Neural Mechanisms Underlying Core Visual Perception of Objects

by

Ha Hong

B.S. Physics
Korea Advanced Institute of Science and Technology (2009)

Submitted to the Harvard-MIT Program in Health Sciences and Technology in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Medical Engineering and Medical Physics

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2015

Author ..... Signature redacted ....
Harvard-MIT Program in Health Sciences and Technology
May 18, 2015

Certified by.. Signature redacted ....
James J. DiCarlo, MD, PhD
Professor of Neuroscience
Head, Department of Brain and Cognitive Sciences at MIT
Thesis Supervisor

Accepted by .... Signature redacted ....
Emery N. Brown, MD, PhD
Director, Harvard-MIT Program in Health Sciences Technology
Professor of Computational Neuroscience and Health Sciences and Technology

Neural Mechanisms Underlying Core Visual Perception of Objects

by

Ha Hong

Submitted to the Harvard-MIT Program in Health Sciences and Technology
on May 18, 2015, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Medical Engineering and Medical Physics

Abstract

Visual perception of objects is a computationally challenging problem and fundamental to human well-being. Extensive previous research has revealed that the inferior temporal cortex (IT), a high-level visual area, is involved in various aspects of visual perception. Yet, little is known about: how IT neural responses to objects support human perception of the objects; and how IT responses are produced from retinal images of objects.

The goal of this research is to tackle these two related questions and find out explicit, quantitative mechanisms that describe human *core visual perception of objects*, a remarkable ability achieved with brief (<200ms) image viewing duration. We first operationally define the core visual perception by measuring behavioral reports of human subjects in hundreds of visual tasks. These tasks are designed to systematically assess subjects' ability to estimate key visual parameters of an object in an image, such as the object's category, identity, position, size, and viewpoint angles. Combined with a rich dataset of monkey visual neural responses to the same task images, we systematically explore a large number of explicit hypotheses that might explain the human behavioral reports. Here, we demonstrate that weighted linear sums of IT responses robustly predict the human pattern of behavior. Moreover, we show that performance-optimized hierarchical neural networks explain a large portion of neural responses of high-level visual areas including IT. These results establish a working mechanistic model of core visual perception by providing an end-to-end understanding of the human visual system from images to neural responses to behavior.

Thesis Supervisor: James J. DiCarlo, MD, PhD
Title: Professor of Neuroscience
Head, Department of Brain and Cognitive Sciences at MIT

Acknowledgments

I am surprised at how quickly my past six years have passed here in DiCarlo Lab at MIT. These years have been a fun learning experience and at the same time a critical period of my professional and personal growth. This would have been impossible without the help of many people that I have met during my stay here, and I want to take this opportunity to thank them.

I have been very fortunate to know and work with my advisor Jim DiCarlo, and I am immensely grateful to him. He has preened me as a new scientist with his endless support and teaching that has fueled my desire to learn more about the visual system and its dual implemented in machines. He always helped me when I needed, encouraging both my professional and personal development. In the academic environment he created, I have received enough care and support, and as an international student, this environment has became my newfound home. For me, he is the greatest teacher that I have ever met, and he has made my long journey to PhD much more pleasurable.

I am grateful to my closest academic friends — especially, Dan Yamins, Najib Majaj, and Ethan Solomon. Their support and teaching have been an indispensable component of this thesis work, and I would not have been able to handle many difficulties that I have encountered without them. I would like to thank all the members of DiCarlo Lab. They not only have helped me scientifically, but also made DiCarlo Lab a welcoming home that always has embraced me as a family member despite many of my shortcomings. I am especially grateful to my thesis committee members — Antonio Torralba and John Maunsell — for their continued guidance and interest in my thesis research, which allowed me to navigate the scientific problems much more effectively with confidence. I would like to thank all of my friends and family

members for their affection, kindness, and support both personally and scientifically. I also thank the following funding agencies, which made this work possible: Samsung Scholarship, NIH NEI, NIH NIMH, DARPA, and NSF.

Finally, I would like to express my gratitude to my wife, Kyung Min Dho, for her endless love and always being with me during the entire journey to PhD. She believed me and trusted me, even when I did not trust myself, and her trust eventually enabled me to overcome the obstacles during the journey. She has been strong and wise, caring of our two little boys and family. I genuinely feel that I am a very lucky man to marry her, and to her, I dedicate this thesis.

Contents

Chapter 1

Introduction

Humans parse complex visual scenes rapidly and accurately. Such rapid visual perception feels effortless to us, yet it serves as a fundamental neural substrate that supports a wide range of human functions that are critical to survival, such as obstacle avoidance, threat detection, resource detection, navigation, to name a few. Its seemingly uncomplicated processing of visual information is deceiving however. In fact, it is far from being trivial, because low-level pixel data can undergo drastic changes from variation in position, size, pose, lighting, occlusion, while still representing the same high-level content. Understanding computational mechanisms that are tolerant to these changes but sensitive to distinctions (e.g., identifying one particular object among similar ones, noticing change in the object position, etc.) is a significant computational feat. Knowing how the brain accomplishes this feat presents a unique scientific challenge and could have tremendous implications for a deep understanding of cortical information processing, artificial intervention and repair of the visual system, building computer vision systems, and a broader understanding of human cognition.

1.1 Visual perception and the ventral visual stream

Visual perception of objects is the ability to interpret, analyze, and understand various aspects of objects in scenes by processing images that carry the information of the scenes. Humans rapidly and accurately interpret visual scenes from their surrounding, an ability that is critical to normal functioning. One facet of visual perception of objects is view-invariant object recognition, which involves detection and identification of objects while discounting changes in low-level image statistics [Gross, 1994; Miyashita, 1993; Rolls, 2000; Orban, 2008; DiCarlo and Cox, 2007; DiCarlo et al., 2012]. Visual perception of objects involves estimating a variety of other properties besides an object's category or identity, such as position, size, pose, lighting, occlusion, clutter, non-rigid deformation, and many other factors of the object that are normally discarded during invariant object recognition but are essential for defining scenes [Edelman, 1999; Koenderink and van Doorn, 1979].

The ventral visual stream in the brain houses the computational machinery that enables visual perception of objects. It consists of a set of brain areas in the occipital and temporal lobes of humans [Grill-Spector et al., 2001; Kourtzi and Kanwisher, 2001; Malach et al., 2002], with highly homologous areas in the non-human primate [Kriegeskorte et al., 2008b]. Anatomically, the ventral stream is composed of a set of cortical areas, each thought to convey a distinct representation of the visual image [Felleman and Van Essen, 1991; DiCarlo and Cox, 2007]. The connectivity patterns and latency of responses in each ventral stream area reveal a hierarchical organization (Fig. 1.1). Visual information travels from the retina to the lateral geniculate nucleus of the thalamus (LGN), and then through successive visual cortical areas: V1, V2, V4, and the inferior temporal cortex (IT) [Felleman and Van Essen, 1991; DiCarlo et al., 2012]. Just about 100 ms after photons strike the retina, IT

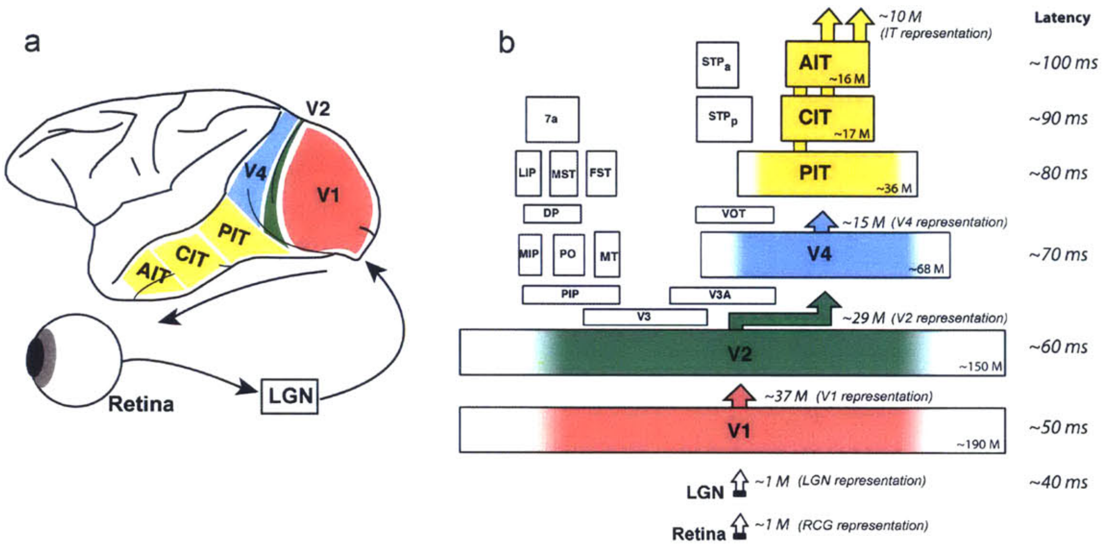


Figure 1.1: (*a*) **The anatomical organization** of the ventral visual stream in macaque, and the flow of visual information from retina to IT. AIT, CIT, and PIT stand for anterior, central, and posterior IT, respectively. (*b*) **Hierarchical structure of the ventral stream.** The area of each box is proportional to the surface area of the corresponding visual area [Felleman and Van Essen, 1991]. The number in the right bottom corner of each box shows the approximate number of neurons in both hemispheres. The number above each area is the approximate representation dimensionality of the area based on the number of layer 2/3 projection neurons [Collins et al., 2010; O'Kusky and Colonnier, 1982]. The colored portion corresponds to the central 10° of the visual field [Brewer et al., 2002]. The right column lists the approximate median response latency [Nowak and Bullier, 1997; Schmolesky et al., 1998]. Figure adopted from [DiCarlo et al., 2012].

represents image-evoked neural responses [Hung et al., 2005c; DiCarlo and Maunsell, 2000a; Desimone et al., 1984a; Kobatake and Tanaka, 1994; Tanaka, 1996; Logothetis and Sheinberg, 1996], most likely produced by a combination of intra-area processing and feed-forward inter-area processing of the visual image [DiCarlo et al., 2012].

Lesions or inactivation of the highest area of the ventral stream — IT — produce selective deficits in object recognition [Yaginuma et al., 1982; Horel, 1996; Holmes and Gross, 1984; Weiskrantz and Saunders, 1984; Schiller, 1995]. Disruption of specific ventral stream sub-regions disrupts specific object discrimination tasks [Pitcher et al., 2009], and artificial activation of sub-regions predictably shifts percepts of complex objects [Afraz et al., 2006; Verhoef et al., 2012].

It has also been experimentally observed that IT cortex normally associated with object recognition appears to retain some sensitivity to object position [Li et al., 2009; DiCarlo and Maunsell, 2003; MacEvoy and Yang, 2012; Sayres and Grill-Spector, 2008; Sereno et al., 2014] and other properties [Nishio et al., 2014]. However, it is not clear how much and exactly what kinds of non-categorical information is present in higher ventral cortex, nor how these properties are integrated with the categorical representation.

A framing hypothesis (Fig. 1.3**a**) for how the brain achieves visual perception is that the ventral visual stream successively transforms and *encodes* low-level pixel-like patterns of neural responses into completely novel patterns of IT population neural responses that more explicitly represent high-level image content (e.g., object identity or category; for reviews, see [Gross, 1994; Miyashita, 1993; Rolls, 2000; Orban, 2008; DiCarlo and Cox, 2007; DiCarlo et al., 2012]) and that the neural responses are *decoded* by the downstream areas to solve various tasks of visual perception [Miyashita, 1993; Freedman et al., 2003]. For example, the initial image-evoked IT neural population responses (100 ms latency) can directly support robust invariant visual object categorization and identification [Hung et al., 2005c; Li et al., 2006; Rust and DiCarlo, 2010], and IT population responses are far more useful for such tasks than are earlier ventral stream representations [Rust and DiCarlo, 2010; Freiwald and Tsao, 2010] or non-ventral stream representations [Lehky and Sereno, 2007]. However, it is *unknown* whether the initial IT neural population responses are sufficient to quantitatively explain human pattern of performance over a wide range of visual perception tasks.

1.2 Models of the ventral visual stream

Model building is a fundamental part of the scientific process, where observations are reconciled and understood through the creation of explanatory frameworks and models. The merit of such models is ultimately measured by their ability to explain existing data and predict new observations. In neuroscience, where the subject of study is effectively a biological computer, the creation of accurate models has not only the power to explain observed data, but also to potentially recreate the abilities of the brain, many of which cannot be rivaled by current artificial systems. As the ventral visual stream enables various aspects of visual perception, the development of robust ventral stream models has been a central problem in the field of computational neuroscience and machine perception. Information processing up to the first stage of the ventral stream (V1) is reasonably well captured by image-based computational models [Lennie and Movshon, 2005; Carandini et al., 2005b; Keat et al., 2001; Jagadeesh et al., 1993; Reid and Alonso, 1996]. However, processing in higher stages (V2, V4, IT) remains poorly understood and difficult to model (but see [David et al., 2006; Connor et al., 2007; Brincat and Connor, 2004; Yamane et al., 2008]).

Over the past decades, neuroscientists have revealed that each ventral stream area shares a common canonical anatomical architecture, including shared motifs for incoming inputs from upstream areas and outputs to downstream areas. This argues for the widely-held hypothesis that each ventral stream area may implement a common information processing strategy, and that the increasingly sophisticated representations found in the ventral stream result from the "stacking" of these areas [DiCarlo et al., 2012]. This "stacked cortex" hypothesis is adopted in many computational models of the ventral stream [Fukushima, 1980; Riesenhuber and Poggio, 1999a; Serre et al., 2007a; Mel, 1997; Lecun et al., 2004; Wallis and Rolls, 1997] and bio-inspired computer vision models [Bengio, 2009; Edelman, 1999; Zhu and

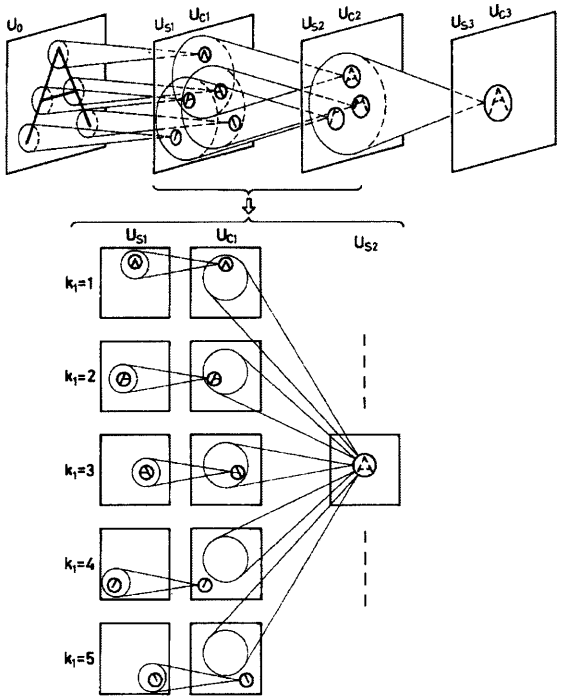


Figure 1.2: An example of "stacked cortex" hypothesis employed in a seminal work on Neocognitron by Fukushima [1980]. This hypothesis generalizes Hubel and Wiesel's idea of simple and complex cells in V1. Figure adopted from [Fukushima, 1980].

Mumford, 2006; Riesenhuber and Poggio, 2000; Pinto et al., 2008b,b]. These models typically include a stack of several hierarchically arranged layers, each implementing AND-like operations to build selectivity and OR-like operations (e.g., MAX, in the HMAX class of models [Riesenhuber and Poggio, 1999a; Serre et al., 2007a]) to build tolerance to identity preserving transformations.

It is thought that these models have a capacity to reliably explain the information processing in the ventral visual stream. These models produce model neurons that signal object identity with invariance to identity-preserving transformations [Wallis and Rolls, 1997; Riesenhuber and Poggio, 1999a; Serre et al., 2007a]. More recent work from our group and others has used high-throughput techniques to search the large parameter spaces of these models families, leading to increases in performance [Pinto et al., 2009; Krizhevsky et al., 2012; Zeiler and Fergus, 2013]. While this is exciting progress, it still remains *unclear* whether these new models are capable of reliably predicting high-level neural responses to a wide range of complex, naturalistic images. We also argue that enormous potential remains untapped in leveraging cutting-edge neurophysiology and psychophysical data to inform modeling efforts.

1.3 Statement of problem and organization of this thesis

While much progress has been made, the field of neuroscience does not yet know a quantitative mechanism of the ventral visual stream that operates on visual images to successfully explain high level neuronal data and human object perception behavior (Fig. 1.3**a**). Specifically, little is known about: how IT neural responses to objects support human perception of the objects (decoding problem); and how IT responses are produced from retinal images of objects (encoding problem).

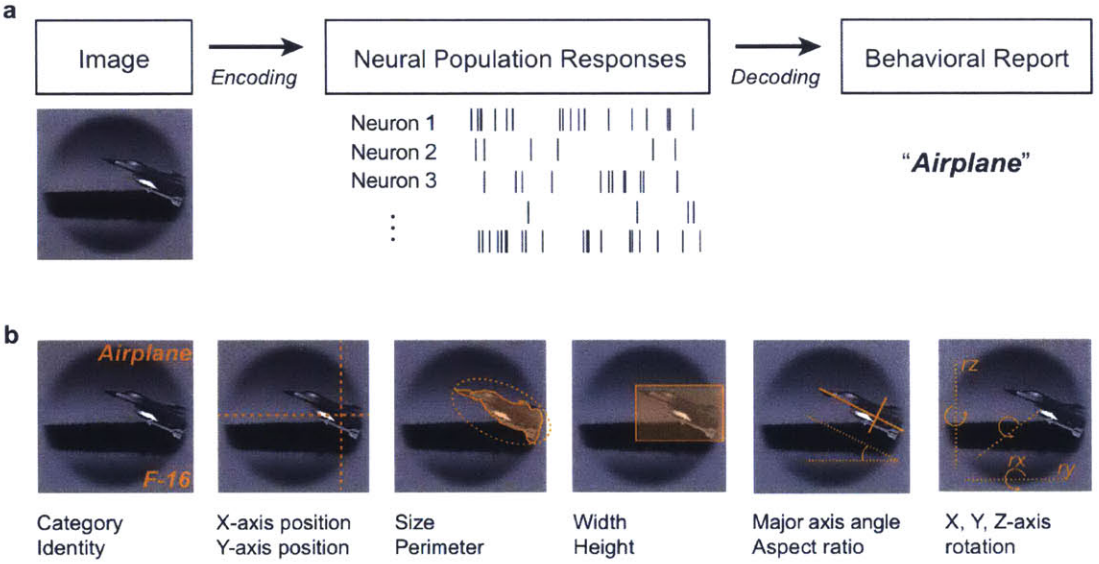


Figure 1.3: (*a*) **Information processing in visual perception.** The ventral visual stream is thought to process information about the form and structure of objects in the environment and *encode* the information as neural population responses. Then, downstream brain areas *decode* the neural responses to extract useful information to perform visual perception. For example, when a subject is asked to solve a visual task (e.g., "what is in this image?"), relevant visual properties are decoded to produce the behavioral report (e.g., "airplane") for the task. (*b*) **Example visual tasks** deployed in this work to operationally define human core visual perception of objects.

The goal of this research is to tackle these two related problems and find out explicit, quantitative mechanisms that describe human *core visual perception of objects*, a remarkable ability achieved with brief (<200ms) image viewing duration within the central 10° of visual field. This viewing time well approximates typical fixation during natural visual exploration [DiCarlo and Maunsell, 2000a; Lehky, 2000; Sheinberg and Logothetis, 2001; Yarbus, 1967], and the spatial range best approximates the retinal region most strongly covered by the ventral visual stream [Ungerleider et al., 2008; Op de Beeck and Vogels, 2000].

1.3.1 Careful measurement of human visual perception behavior

Since the ultimate function of the visual system is to guide behavior, and because our goal is to understand the function and performance of the visual system, careful measuring of the pattern of behavioral report is essential to produce mechanistic models of the ventral stream. It is important to note that human subjects are imperfect on various visual perception tasks. In fact, we observe that there is a rich structure in the pattern of behavioral reports, and we argue that this "finger print" of the ventral stream function should guide the development of mechanistic models of perception.

In Chapters 2 and 3, we first operationally define the core visual perception by measuring behavioral reports of human subjects in hundreds of visual tasks. These tasks are designed to systematically assess subjects' ability to estimate key visual parameters of an object in an image, such as the object's category, identity, position, size, and viewpoint angles (Fig. **1.3b**). Our main experimental paradigm is a brief

(100ms) presentation of an image, followed by a set of possible answer choices to choose from.

1.3.2 High-quality neural data collection

Our goal to uncover mechanisms underlying visual perception critically depends on large quantities of high-quality neurophysiology data in higher-level visual areas (IT and V4). To enable this, our team has developed a powerful preparation with large-scale multi-electrode array recordings in the visual cortex of macaques. This technique allows robust recording from hundreds of sites each day, for up to several months. In Chapter 2, we train non-human primates to view images, and we record population neural responses to the same task images we used to collect human behavioral data. All procedures using animals are done in accordance with the MIT Institutional Animal Care and Use Committee.

Here, we deliberately decide to collect *monkey* neural data to understand *human* visual perception. Key neural data cannot currently be obtained in human subjects at sufficient spatial and temporal resolution, so such data must be obtained in an animal model. Done correctly (i.e., exact same sensory stimuli, as we do here), the only assumption is that the relevant neuronal substrates and mechanisms are shared between the animal model and the human. This monkey-to-human assumption is widely held and is a key justification for the use of non-human primates in neuroscience research.

In vision, it is supported by behavioral data showing very close matches between humans and non-human primates in a range of visual tasks [Shadlen et al., 1996; Newsome et al., 1989; Britten et al., 1992, 1996; DeValois, 1965, 1978], and by consistent organization of visual areas in cerebral cortex [Orban et al., 2004] and consistency between monkey neuronal data and human fMRI in IT [Kriegeskorte et al., 2008a,b]. Nevertheless, we acknowledge that, for complex tasks that, for example, heavily rely on human-specific knowledge, the monkey-to-human assumption might be ungrounded, and thus, we carefully analyze our data to detect any failure in this assumption.

1.3.3 IT-to-behavior linking mechanism (decoding)

With the large-scale human behavioral and monkey neural response data on the task using the same images, we systematically explore a large number of explicit hypotheses about the neural basis of human behavioral reports. Neurons produce spikes, as opposed to the behavioral report of human subjects. In order to predict behavioral reports from neural responses, we first convert each neuronal unit's spikes into per-image scalar value (i.e., rate code spike counting) and concatenate the values across all units to make the response vector for the image. Then, the per-image response vector is "decoded" into (predicted) behavioral report by a decoder (e.g., a simple weighted linear sum), which is an explicit hypothesis of linking mechanism. In Chapters 2 and 3, we employ this approach to compute many explicit, quantitative hypotheses that span ideas in the literature, and we ask which ones robustly explain human patterns of behavioral reports over all visual tasks, and which are ruled out by the data.

1.3.4 Image-to-IT linking mechanism (encoding)

We aim to characterize the cortical transform that produces the output of high-level visual areas (IT and V4) as a function of image input in Chapter 4. We start with combining[1] simple classes of models for a single cortical layer; use high-throughput methods to identify high-performing models on, for example, categorization tasks[2]; and test candidate models against empirical metrics that evaluate the models' power in predicting neural responses; identify places where models succeed or fail to capture IT and V4 structures; and then use that information to constrain a new round of model development, building model sophistication as necessary.

As a starting point, and inspired by previous neuronal modeling work David et al. [2006]; Connor et al. [2007]; Brincat and Connor [2004]; Yamane et al. [2008]; Rust et al. [2006], we limit the constituent operations of cortical processing in a single layer to a set of simple computational elements, including: (1) a filtering operation, implementing template matching; (2) a simple nonlinearity, for example, thresholding; (3) a local pooling/aggregation operation, such as softmax; and (4) a local competitive normalization. Each of these operations is in fact a large family of possible operations, specified by a set of parameters[3]. Each simulated neural unit in the n-th layer is then modeled as a function of population activity of the $(n-1)$-th layer and specific choice of these computation elements, e.g.,

$$\text{Simulated unit in the } n\text{-th layer} =$$

$$Pool_{\theta_p}(Normalize_{\theta_N}(Threshold_{\theta_T}(Filter_{\theta_F}((n-1)\text{-th layer input}))))$$

where the parameters θ_p, θ_N, θ_T, and θ_F describe each of the constituent operations,

[1] That is, "stacked cortex" hypothesis in Section 1.2

[2] We do not optimize directly for e.g. neural responses, as the model is poorly constrained by the neurophysiology data. This approach is not only computationally more tractable, but also: (1) a stronger form of generalization; and (2) capable of providing a key insight into how the ventral steam is "developed." See 1.2 for details.

[3] For example, fan-in and fan-out, threshold values, pooling exponents, the spatial extent over which the operations operate, and the size/shape/content of the templates that are matched

and the starting layer is simply the input image.

Simulated neural units are compared against recorded monkey neural responses. This comparison can be characterized at many levels of abstraction, from low-level neural output prediction to high-level behavioral concordance. A reasonable low level metric could be the per-unit explain variance across spike counts for images during a behaviorally-relevant interval (i.e., rate code). At higher level measure, we attempt to evaluate the population level match by using neural representation similarities, as in [Kriegeskorte et al., 2008a,b].

Chapter 2

Inferior Temporal Neuronal Firing Rates Accurately Predict Human Core Object Recognition Performance*

To go beyond qualitative models of the biological substrate of object recognition, we ask: can a single ventral stream neuronal linking hypothesis quantitatively account for core object recognition performance over a broad range of tasks? We measured human performance in 64 object recognition tasks using thousands of challenging images that explore shape similarity and identity preserving object variation. We then used multi-electrode arrays to measure neuronal population responses to those same images in visual areas V4 and IT (Inferior Temporal Cortex) of monkeys. We

*This chapter is modified from a study that has been submitted to *Journal of Neuroscience* as of May 2015. This work has been done in collaboration with Najib Majaj, Ethan Solomon, and James DiCarlo.

tested leading candidate linking hypotheses and control hypotheses — each postulating how ventral stream neuronal responses underlie object recognition behavior. Specifically, for each hypothesis we computed the predicted performance on the 64 tasks and compared it to the measured pattern of human performance. All tested hypotheses based on low and mid-level visually-evoked activity (pixels, V1, and V4) were very poor predictors of the human behavioral pattern. However, simple learned, weighted sums of distributed average IT firing rates exactly predicted the behavioral pattern. More elaborate linking hypotheses relying on IT trial-by-trial correlational structure, finer IT temporal codes, or ones that strictly respect the known spatial sub-structures of IT ("face patches") did not improve predictive power. While these results do not reject those more elaborate hypotheses, they suggest a simple, sufficient quantitative model: each object recognition task is learned from the spatially distributed mean firing rates (100 ms) of ~60,000 IT neurons, and is executed as a simple weighted sum of those firing rates.

2.1 Introduction

The detailed mechanisms of how the brain accomplishes viewpoint invariant object recognition remain largely unknown, but lesion studies point to the ventral stream (V1-V2-V4-IT) as critical to this behavior [Holmes and Gross, 1984; Biederman et al., 1997]. Previous ventral stream studies have focused on understanding the non-linear transformations between the retina and neural responses [Gallant et al., 1996; Hegdé and Van Essen, 2000; Pasupathy and Connor, 2002; Connor et al., 2007; Freeman et al., 2013], including evidence that IT is better at recognition than early representations [Hung et al., 2005c; Rust and DiCarlo, 2010], and that IT responses are partially correlated with perceptual report [Sheinberg and Logothetis, 1997; Op de

Beeck et al., 2001; Kriegeskorte et al., 2008b]. While such studies tell us much about visual processing and support the belief that the ventral stream is critical to object representation, they do not present a single linking hypothesis that is quantitatively *sufficient* to explain how ventral stream neural activity accounts for object recognition performance over all recognition tasks. This study aimed to provide that link for a subdomain of object recognition — core object recognition [DiCarlo et al., 2012] — in which images are presented for 100 ms in the central 10 degrees of the visual field.

To do so, our strategy was to: (1) develop a stringent behavioral assay, (2) obtain sufficient neuronal sampling, (3) implement specific hypotheses that each predict perceptual report from neural activity, and (4) compare those predictions with actual perceptual reports. We addressed each challenge as follows:

1. We characterized human core object recognition performance using large image sets that explore shape similarity and identity preserving image variation, and assumed that monkey and human patterns of performance are equivalent (see Discussion). The 64 recognition tasks we used set a high bar, because performance on them varies widely and is not explained by low-level visual representations (see below).

2. We measured neuronal responses in visual area V4 and along IT cortex [Felleman and Van Essen, 1991] using the same pool of images used in the behavioral testing. We relied on multi-electrode arrays to monitor hundreds of sites each tested with multiple repeats of 5760 images (See Methods). Our measured neuronal population was adequate for quantifying uncertainty with respect to neuronal sampling and allowed us to extrapolate to larger numbers of neurons.

3. We tested specific quantitative versions of previously proposed hypothetical

links between neuronal activity and recognition behavior, as well as control hypotheses. Each *neuronal linking hypothesis* is a postulated mechanism of how downstream neurons integrate ventral stream activity to make a decision about which object label the observer will report in each image [Connor et al., 1990; Parker and Newsome, 1998; Johnson et al., 2002].

4. Ideally, a sufficient linking hypothesis should predict the perceptual report for each and every image. Here we focused on predicting the mean human recognition accuracy d' for all object recognition tasks, where each task contains many images. Specifically, we compared the predicted pattern of d' with that measured in humans on the same 64 tasks (which could lead to a range of outcomes; see Figure 2.1b).

We report here that simple, Learned Weighted Sums of Randomly-selected Average neuronal responses spatially Distributed over monkey IT (referred to here as "LaWS of RAD IT") are able to meet that high bar (Figure 2.1b, upper right panel). In contrast, other linking hypotheses based on neuronal responses from IT or other visual areas fall short. While this is compatible with previous ideas about IT's role in object recognition [Tanaka, 1993; Kobatake and Tanaka, 1994; Tanaka, 1997], it is, to our knowledge, the first demonstration that a *single, specific* neural linking hypothesis is quantitatively sufficient to explain behavior over a wide range of core object recognition tasks.

2.2 Results

2.2.1 Quantitative characterization of human core object recognition

To characterize human core object recognition abilities [DiCarlo et al., 2012], we designed 64 core object recognition tasks, and obtained an unbiased measure of human ability (d', see Methods) on each task. Figure 2.3**a** uses a color scale to show human d' values for each of the 64 tasks. The wide range of values is not due to subject variability, as the pattern of values over the 64 tasks from any one subject is highly correlated with the pattern of values from the pooled results of all other subjects (median correlation = 0.93, see Figure 2.3**b**). Instead, it shows that all humans find some object recognition tasks to be easy ($d' \simeq 5$; corresponds to an unbiased accuracy of 99.4% percent correct in a "one-vs-rest" classification, where 50% is chance), others to be more difficult ($d' \simeq 2$; 84.1% correct), and others to be very challenging ($d' \simeq 0.5$; 59.9% correct). Two unsurprising qualitative trends are noted. First, human object recognition ability is dependent on shape similarity: we found high d's for basic-level categorization tasks ("cars" vs. "non-cars", "faces" vs. "non-faces", "animals" vs. "non-animals", etc.; mean d' across all levels of object view variation = 3.46), lower d's for car identification tasks (easy subordinate, e.g., "car 1" vs. "not car 1"; mean d' = 1.49), and even lower d's for face identification tasks (challenging subordinate, e.g., "face 1" vs. "not face 1"; mean d' = 0.50). Second, human object recognition ability drops as variation in object view (position, scale, pose) increases: mean d' = 2.39 for low variation, 1.89 for medium, and 1.50 for high variation. While these results show that humans are not completely invariant, they confirm that humans tolerate significant amounts of object view variation. Figure 2.3**a** also shows that tolerance interacts with shape similarity, with humans being least tolerant for the most difficult subordinate tasks [Biederman and Gerhardstein, 1993; Tarr and Bülthoff, 1998; Tjan and Legge, 1998]. We note that these measurements of human object recognition ability were designed and carried out independently of any neuronal data collection.

2.2.2 Candidate linking hypotheses that might predict human object recognition behavior

A candidate linking hypothesis that aims to predict the observed pattern of human recognition accuracy must have at least two components: (1) a specification of the "features" of neural activity that are relevant to behavior (a.k.a. neuronal code), and (2) a specification of a biologically plausible mechanism that translates that neural code to a behavioral choice on each trial.

Based on the existing literature, the "features" of neural activity of high interest include: the tissue region where the neuronal responses are found (V4, IT, IT inside "face patches", IT outside "face patches"), the size of the neuronal population (number of neural sites or units), the temporal window over which the responses are considered, the temporal grain of those measurements (e.g., spike timing codes vs. rate codes), and consideration of the so-called "trial-by-trial" population-wide correlation of activity. One can imagine many variants and combinations of these ideas, not all of which can be fully explored in a single study, but we aimed to specify and then test some of the most widely believed ideas. That is, we used our data to measure the code specified by each hypothesis, and then we asked how well that code predicted the measured object recognition performance.

To compute the predictions of each code, a linking hypothesis must also specify a biologically plausible mechanism (decoder) that translates the measured (neural response) features into an object label on each trial. For example a "car" decoder translates the measured neuronal features into the binary decision: is a car present in the image or not? In this study, we tested simple linear classifiers (a.k.a. linear discriminants) each such decoder computes a simple weighted sum of the features in the proposed population code. While this study is agnostic with respect to how this type of decoder might be implemented in downstream brain areas (e.g., PFC [Freedman et al., 2001], perirhinal cortex [Pagan et al., 2013]), it is known to be biologically plausible: each synapse on a hypothetical downstream neuron corresponds to a "weighting" on part of the neuronal code, and the neuron's output is determined by the weighted sum of all its inputs—corresponds to a decision by the decoder (e.g., [Shadlen et al., 1996]). Different types of decoders correspond to different assumptions about how those downstream neurons learn the synaptic weights (e.g., in humans, this might correspond to learning to map visual inputs to specific words in the lexicon). In this study, because we were primarily interested in the neuronal features that best predict adult object recognition performance, our approach was to start with a simple, well-known decoder, hold that idea, and then later explore different types of decoders to determine their impact on our conclusions (see Figure 2.10). All performance measures reported in this paper are based on neural responses to images that were never previously seen by the decoder (a.k.a. cross validation, see Methods).

In sum, to test each conceptual neuronal linking hypothesis we: (1) instantiated the idea by measuring the proposed neural code in the population data, (2) learned a hypothesized downstream decoder (e.g., one for each of the 24 noun labels) that takes that neuronal code as input and finds the simple weighed sum that gives the highest performance on that task (see Methods for details), and (3) compute the behavioral predictions of that hypothesis for each of the 64 tasks using previously unseen images.

To the extent that each neuronal linking hypothesis predicts a different pattern of behavioral performance across the 64 tasks, not all linking hypotheses can accurately predict the observed pattern of human behavioral performance. A priori, it was also possible that none of the linking hypotheses would accurately predict the human pattern of behavior—for example, we may not have sampled enough neurons to reveal that a hypothesis is sufficient, or perhaps monkey and human performance patterns are different and thus no linking hypothesis tested on monkey neuronal codes can predict human patterns of performance. Nonetheless, we reasoned that we could use the strategy of comparing the predicted vs. actual object recognition performance of each neuronal linking hypothesis to infer which hypothesis corresponds most closely to the mechanisms at work in the brain.

The 64 human population d's were the benchmark to which we compared our candidate linking-hypotheses. To capture the four possibilities for such a comparison (Figure 2.1**b**), we defined two metrics, *consistency* and relative *performance*. To quantify the match between the pattern of predicted and human d's, we computed *consistency*, the rank correlation between predicted d' and actual human d' across all 64 object recognition tasks. To quantify the match on average between predicted and actual human d', we computed relative *performance*, the median ratio of predicted d' and actual human d' across all 64 object recognition tasks. To estimate the human subject-to-subject variability for *consistency* and *performance* we selected one subject from each task set and combined the task performance of the three task sets to produce 64 "individual" human d's (Figure 2.3**b**). We repeated this procedure multiple times to construct an ensemble of individual human subjects. We used this ensemble to compute the median *consistency* (Spearman rank correlation coefficient) and *performance* (1 by definition) between individual human d's and pooled population human d's and the 68% confidence intervals around that median.

In total, we tested 944 linking hypotheses, in each case varying the number of neural sites included, thereby translating that conceptual linking hypothesis into an exact specification which makes falsifiable predictions. For example, one specific linking hypothesis we tested was: learn simple weighted sums of the mean firing rates across 128 IT neural sites, distributed across IT measured in a 70–170 ms time window, ignoring trial-by-trial correlations (Figure 2.4). To facilitate visual inspection, the behavioral predictions of this linking hypothesis are strung out in a single color coded vector (Figure 2.4**c**). Most candidate linking hypotheses produced different predicted patterns of behavioral performance; sometimes these differences were small, but often they were dramatic (see Figure 2.5**a** for examples). For visual comparison, Figure 2.5**a** also shows human performance on the same 64 object recognition tasks (data from Figure 2.3**a**, replotted), and it illustrates that some candidate linking hypothesis lead to very poor predictions of the pattern of human performance, while others lead to surprisingly good predictions. Next we ask: which, if any, candidate neuronal linking hypotheses predict the human pattern of behavior over all 64 tasks (Figure 2.3**a**)?

2.2.3 A specific monkey IT based linking hypothesis predicts human core object recognition behavior

Before delving into the large space of linking hypotheses that we explored, we start by summarizing our main result. Our analyses revealed that Learned Weighted Sums of Randomly-selected Average neuronal responses spatially Distributed over IT (here termed the LaWS of RAD IT linking hypothesis) produced a pattern of behavioral performance that perfectly predicted the observed pattern of human behavioral performance. Figure 2.4 shows the predictions of a specific instance of the LaWS of RAD linking hypothesis based on the following specification: (1) randomly chosen 128 sites of (2) IT, with the response (3) averaged over a (4) 70–170 ms time window after image onset. The pattern of predictions for the 64 recognition tasks is statistically indistinguishable from the pattern of human behavior (Figure 2.3**a**) and is clearly superior to an identical LaWS of RAD linking hypothesis based on 128 V4 neuronal sites (comparisons quantified in Figures 2.5**a**,**b** and 7).

In total, we collected the responses of 168 spatially separate neuronal sites in IT (M1: 58, M2: 110) and 128 sites in V4 (M1: 70, M2: 58). We pooled neuronal sites across IT because we did not see any strong differences between its subdivision (PIT, CIT, AIT; see Figure 2.9). We measured each site's spiking response pattern to each of 5760 images, drawn from the same pool used in the human psychophysical testing. Each image was presented at least 28 (typically > 47) times per site (i.e. a total of $\sim$250,000 visual stimulus tests at each site). We could not collect this large volume of data in a single day—it required $\sim$30 days of recording in each animal. The initial linking-hypotheses we explored were based on multi-unit activity (MUA) and assumed stability of that measure at each recording site over the 30 recording days [Chestek et al., 2011]. We then examined how our main finding might change if we used single unit response data instead, and examined our assumption of the stability of each recording site over days (see Methods and Figure 2.8). We also considered the fact that we only sampled a small number of IT neuronal sites (relative to the millions of neurons in IT cortex). While these factors are important for estimating the *number* of neurons needed to predict behavior, they turn out to have little impact on our main finding.

2.2.4 Quantifying the goodness of a linking hypothesis: *consistency*

Following previous work [Hyvarinen et al., 1968; Newsome et al., 1989; Connor et al., 1990], we reasoned that the neuronal linking hypotheses that are most likely to correspond to the mechanisms at work in the brain are those that produce the most quantitatively consistent relationship with the human behavior (i.e. the linking hypothesis's pattern of colors in Figure 2.5**a** should best match the pattern of colors in Figure 2.3**a**). To quantify that *consistency*, we computed Spearman's rank correlation coefficient over the 64 d's [Yoshioka et al., 2001].

The most stringent application of this method is that, for a linking hypothesis to remain viable, it must produce behavior that is indistinguishable from the behavior of individual subjects. Based on this stringent criterion, all V4-based linking hypotheses we tested failed to accurately predict the observed human behavioral pattern (See Figures 2.5**b** and 2.6), as did all V1-based linking hypotheses. For comparison, Figure 2.5**a,b** also shows the predictions of linking hypotheses based on populations of baseline computer vision features, all of which failed to predict the pattern of behavior (see Methods for details). Despite our best efforts, we found that the V4 and V1 based linking hypotheses could not be "rescued" by increasing the number of neurons in the linking hypothesis, or by changing the type of decoder (i.e. learner; Figures 2.5 and 2.6). We also considered the possibility that V4-based linking hypotheses might have been handicapped by receptive field limitations of our neural sampling. In particular, we narrowed our images to only those with objects presented in the contralateral field or at the center of gaze. While V4 populations showed the expected pattern of higher d's for contralaterally presented objects, neither test substantially improved the ability of V4-based linking hypotheses to predict the pattern of human behavior (see Methods), even though these same V4 populations often *outperformed* humans in some of the behavioral tasks (see below). These results do not argue that V1 and V4 play no role in object recognition behavior. Instead, they suggest that neural representations (a.k.a. codes) conveyed by those areas do not *directly* underlie object recognition behavior (compatible with previous suggestions [Sheinberg and Logothetis, 1997; Brincat and Connor, 2004; Rust and DiCarlo, 2010; DiCarlo et al., 2012]. These results also show that the approach we employed, the combination of images and tasks, is a powerful test of neuronal linking hypotheses that cannot easily be "passed" by lower level (e.g., V1) or even mid-level (V4) representations.

In contrast to the results in V4 and (simulated) V1, wc found that some IT-based linking hypotheses accurately predicted the behavioral pattern of human observers. For example, based on previous work [Hung et al., 2005c], one of the first specific linking hypotheses we tested was: the mean firing rate of each IT neurons in a 70–170 ms time window, where IT neurons are sampled in a distributed manner over IT cortex (i.e. ignoring IT spatial sub-structure such as "face-patches"), and ignoring correlations across the population (Figure 2.5**d**). We tested this linking hypothesis using different numbers of IT neural sites, and we were surprised to find that, once we included $\sim$100 sites, this IT-based linking hypothesis was not only a more accurate predictor of human behavior than other hypotheses (e.g., V4-based linking hypotheses), but that its predictions were statistically *indistinguishable* from the measured pattern of human d's (linking hypotheses that pass into the gray region in Figure 2.5**b**). Following up on this result, we also found this simple IT-based linking hypothesis continued to accurately predict the pattern of human object recognition ability, even when we varied: the number of neuronal sites participating in the linking hypotheses ($>$ 64 sites), the type of decoder used, and the training provided to the decoder (Figures 2.7, 2.8, and 2.10).

We explored several other IT-based linking hypotheses that have been suggested in the literature. First, we considered the idea that trial-by-trial correlations in neuronal firing across the IT population might be important to consider when asking if a neuronal linking hypothesis is consistent with behavior [Zohary et al., 1994; Averbeck et al., 2006; Liu et al., 2013]. Because we had collected responses at many of our neuronal sites simultaneously, we were able to compare neuronal codes produced across the population on actual single trials, with codes produced on artificial single trials in which any population correlation structure is removed by shuffling the trials (e.g., so the responses of IT unit 1 on presentation p of image i are considered along with the responses of IT unit 2 on presentation q of image i). We found that a LaWS of RAD IT linking hypothesis that maintained the trial-by-trial population correlation structure had no increased (or decreased) ability to explain the pattern of human behavior, even when lowering the number of neurons so that we might be able to see that increase (Figures 2.7–8).

Second, we considered the idea of finer grained temporal codes. To do this, we took the simple 70–170 ms post-stimulus time window (above) where the LaWS of RAD IT linking hypothesis was highly predictive, and we broke it into successively smaller and smaller time windows, giving each learned decoder full access to the neural response in all such time windows. Because all of the same spiking information in these finer-grained temporal codes is still available to each decoder, this approach can only maintain or increase performance on each of the 64 behavioral tasks (until data limits are reached). However, because accuracy on some tasks might improve relative to others, it could either increase, decrease, or have no effect on, the *consistency* of the *pattern* of performance over all 64 human behavioral tasks. The results showed that, relative to the simple 100-ms window mean firing rates in the LaWS of RAD linking hypothesis, these more complex, finer-grained IT temporal codes led to no measurable change in the linking hypothesis's *consistency* with the pattern of human behavior.

Third, we considered modular IT linking hypotheses, where different sub-regions of IT are devoted exclusively to certain kinds of tasks. The strongest experimentally-motivated example of a modular linking hypothesis is that certain spatial regions of IT (face "patches" in monkeys; FFA, OFA in humans) are devoted to certain types of "face-related" tasks, such as face discrimination (one face vs. others) and face detection (faces vs. other categories). Our data allowed us to examine such hypotheses because 19 of our 64 tasks are face-related tasks, and we could label ~19% of our IT neural sites as likely belonging to one or more of the 6–10 IT face patches (based on the high purity of these regions for units that have high face vs. non-face object selectivity [Tsao et al., 2006]). We first note that our findings are consistent with weaker forms of modularity of face processing, such as spatial clustering of neural sites that are most important for face detection. Indeed, it was not surprising (as it is nearly by definition) that, of the IT sites that were weighted most strongly by the decoders (i.e. the top 5% most heavily weighted) in our three face detection tasks, 87.5% of those were face-patch-likely sites. More interestingly though, we also found that only 12.5% of the most highly-weighted sites in our 16 face discrimination tasks were face-patch-likely sites, arguing that face discrimination might not rely exclusively on IT face-patch tissue. To test a stronger form of the face modularity hypothesis using the *consistency* approach of this study, we asked if neuronal linking hypotheses based only on the face-patch-likely population of sites were more consistent with the pattern of human performance on face-related behavioral tasks (compared with linking hypotheses based on all of IT or based only on face-patch-unlikely populations within IT; Figure 2.5**e**). We found that this did not significantly change the accuracy of the behavioral fits—if anything, the trend suggested a decreased accuracy. In sum, while our results are consistent with weaker forms of the face-modularity linking hypothesis (i.e. face detection tasks are best performed by "face (detection) neurons" which are spatially clustered [Tsao et al., 2006; Issa and DiCarlo, 2012], our data find no support for the stronger form of the face-modularity hypothesis (i.e. all face-related tasks exclusively depend on the responses of neurons in face patches). However, our data do not falsify that strong form either.

2.2.5 Goodness of a candidate linking hypothesis: *performance*

While the *consistency* metric evaluates the similarity between the *pattern* of d's predicted by each candidate linking hypothesis and the measured human d's, we next asked: what number of neurons is required for a LaWS of RAD IT linking hypothesis to account for the actual d's across all our 64 tasks? In particular, one can imagine neuronal linking hypotheses that are highly predictive of the pattern of d's over the 64 tasks (as in Figure 2.5), but with absolute levels that are far below the measured human d's (see Figure 2.1**b** for a schematic demonstration of correlated but unequal d's). Indeed, we found examples of such linking hypotheses (see blue points in Figure 2.6 that are within the top gray band, but outside of the red dashed circle). We found that, for both V4 and IT-based codes, once the number of neural sites was greater than ~100, measures of *consistency* were largely insensitive to further increases in the number of neural sites in the code. However, *performance*, the median of the ratio between predicted and actual (human) d' across all 64 tasks, of any specific neural code *was* strongly dependent on the number of neuronal sites. For example, while we found it effectively impossible to vary the number of neural sites to make, for example, a V4-based linking hypothesis match the human *pattern* of performance (Figure 2.7**b**), for many V4-based linking hypothesis, we could, by extrapolation, estimate the number of neurons that could, in principle, match the median human d' over the 64 tasks.

2.2.6 Effect of number of units on *consistency* and *performance*

We systematically explored the effect of changing the number of neural sites on *consistency* and *performance.* This is illustrated in Figure 2.7 for two families of linking hypotheses—the simple LaWS of RAD IT linking hypothesis family reviewed above, and the simple LaWS of RAD V4 linking hypothesis family. For both families, median predicted performance increased as the number of sites increased, however, only the LaWS of RAD IT linking hypothesis became fully consistent with human performance. That is, with 128 neuronal sites (or more), the LaWS of RAD IT linking hypothesis in Figure 2.7 perfectly predicted the entire pattern of performance over all 64 tasks in that the Spearman correlation (Figure 2.7**a,b**) was indistinguishable from the human-to-human *consistency* (the horizontal dotted line in Figure 2.7**b**; the gray region indicate the variability of individual human subjects).

Figure 2.7**a** also illustrates why the non-IT-based linking hypotheses we tested failed to explain the pattern of human performance. In particular, it shows that the LaWS of RAD V4 linking hypothesis fails both because it cannot achieve high d's on some tasks (e.g high variation tasks, green filled circles in Figure 2.7**a**), and because it achieves d's that are *better* than humans in other tasks (e.g., some low variation tasks, green open circles in Figure 2.7**a**). Increasing the number of neurons participating in the LaWS of RAD V4 linking hypothesis cannot fix this obvious discrepancy with behavior, and the result argues against the idea that we did not collect sufficient information from V4 neurons. In sum, distributed, learned V4 population rate codes do *better* than humans on some particular behavioral tasks, but they fail to produce the human *pattern* of *d'*s over all 64 tasks.

2.2.7 Sufficient single-trial, single-unit population linking hypotheses

As shown in Figure 2.7, both the *consistency* and *performance* of IT-based linking hypotheses are dependent on the number of neural sites assumed to be participating in the behavior. The plot shows that only a small number (hundreds) of neural sites are needed before the *consistency* of LAWs of RAD IT hypothesis plateaus. That is, that linking hypothesis achieves a pattern of performance that is indistinguishable from humans after it includes ~100 sites (y-axis in Figure 2.7**b**), and the inclusion of more neural sites does not improve *consistency.* However, another constraint on the number of neuronal sites comes from how the mean performance of a specific linking hypothesis compares with the mean performance of human subjects (x-axis in Figure 2.7**b**; see the caption for the definition of *performance*). Unlike *consistency*, *performance* is an unbounded metric that depends on signal-to-noise ratio. Too few neuronal sites lead to median predicted performance that is below observed performance, and too many lead to performance that is superior to behavior. This offers the opportunity to find the number of neural sites where the linking hypothesis matched human performance (Fig. 2.1b, upper right). However, to estimate that number of neurons, it becomes very important to consider exactly how the hypothesis is implemented and its relationship to brain circuitry. In particular, we and others assume that neurons in downstream brain areas can listen to the spikes of some number of single neurons (e.g., neurons in IT) and produce, on each behavioral trial, a guess as to the object label (the task we asked the humans to perform, Figure 2.3**a**).

In that regard, it is critical to note that the neural data and analyses used to generate Figures 2.4-7 differed from that assumption in two ways: (1) we did not distinguish single units from multi-units, and (2) we averaged the responses of each neural site over many repetitions (typically 50, minimum 28). Neither of these details substantially altered our conclusions about the behavioral *consistency* of LaWS of RAD IT linking hypotheses. However, they are important for estimating how many single neurons would be needed to match human level accuracy on single image presentations.

Figure 2.8**a** examines the difference between multi-unit activity (MUA) and the activity of sorted single units (SUA). Linking hypotheses based on single-unit and multi-unit IT data shared highly comparable *consistency-performance* relationship, except that single-unit linking hypotheses required approximately two times as many neural sites to reach a similar level of *consistency* or *performance.* This similarity is perhaps surprising (see Discussion), but is compatible with previous work that examined the same issue [Hung et al., 2005c].

Figure 2.8**b** explores the issue of averaging and compares the results of a simple model of single trial decoding to the results of decoding while averaging across all available trials. While we did not expect this analysis choice to change our conclusions about the behavioral *consistency* of LaWS of RAD IT linking hypotheses, we expected that it would affect the estimated number of IT neurons that must participate in that linking hypothesis to achieve human-level performance on single trials (because averaging improves the single-to-noise of each neuronal site). The single-trial analysis in Figure 2.8**b** gives a *consistency-performance* relationship similar to that of average-trial analysis if ~60 times as many IT units are provided. That is, we estimate that ~60 independent IT neural sites (operating in parallel) are sufficient to stand in for a single, "repetition-averaged" neural site, and this estimate accounts for how neuronal variability ("noise") affects both the decoding (e.g., as in [Shadlen et al., 1996]) and the learning of the decoder (see more below).

Taken together, the analyses presented in Figure 2.8 converge to suggest that spike counts from ~60,000 (529 rep-averaged IT multi-unit (see Fig. 2.7**b**) × ~2 × ~60) distributed single units in IT cortex can, when read with simple, biologically-plausible downstream neural decoders, perfectly predict both the behavioral pattern of performance and the median level of performance over all 64 tasked object recognition tasks. This number is an extrapolation, as our methods are not yet capable of recording that many IT neurons, and other factors such as "noise correlation" might alter that estimate (see Discussion). Furthermore, because *performance* depends on parameters of how the code is learned to be read (decoded), this estimate could be somewhat higher or lower, as analyzed in detail in Figure 2.10. However, we note that this number is far less than the total number of neurons estimated to project out of IT to downstream targets (~10 million [DiCarlo et al., 2012]).

2.2.8 Effect of time window on *consistency*

To further explore the precise parameters of the LaWS of RAD IT family of linking hypotheses, we varied the starting time and duration of the time window over which the mean rate was read from the IT population (Figure 2.7**c**,**d**). We found that the LaWS of RAD IT linking hypothesis begins to be highly consistent with behavior at a center latency of 100 ms (time window of [50, 150] ms after image onset), and that *consistency* remains at a high plateau for nearly 100 ms before dropping off. During this entire plateau, the predicted pattern of performance of this linking hypothesis is statistically indistinguishable from the human pattern of performance. For comparison, all LaWS of RAD V4 linking hypotheses we tested failed to pass this *consistency* test for all temporal windows.

2.3 Discussion

We propose a framework for comparing neural responses to behavior. Instead of *qualitatively* comparing performance on a selected set of conceptual tasks, we devised a "Turing" test—a battery of behavioral tasks that explore the range of human subjects' capabilities in core object recognition. This operational definition of object recognition provided a strong *consistency test* by which we could quantitatively evaluate different neuronal linking hypotheses that might explain behavior. As expected, many neural (and non-neural) linking hypotheses failed to predict object recognition behavior, including: Pixel-based codes, V1-like-based codes, multiple Computer Vision codes, V4-based codes, and several IT-based codes. However, we were surprised to find that simple, learned weighted sums of randomly-selected average responses of distributed IT neurons (LaWS of RAD IT linking hypothesis family) perfectly predicted the human pattern of behavioral performance across *all* 64 recognition tasks. More precisely, the data argue that a simple rate code (100 ms time scale, [70, 170] ms onset latency) read out on single-trials, learned from a distributed population of $\sim$60 thousand single IT units can fully explain both the pattern and the magnitude of human performance over a large battery of recognition tasks.

Initially, we were surprised that this simple linking hypothesis is virtually perfect at predicting the pattern of performance. Nevertheless, we explored other ideas motivated from theoretical considerations [Averbeck et al., 2006] and neuronal response findings [Sugase et al., 1999; Tsao et al., 2006]. First, we found that the LaWS of RAD linking hypothesis was not strongly affected by trial-by-trial correlational structure in the population responses (Figure 2.5**d**). We suspect that this is due to the dimensionality of our neuronal populations (>100) combined with the fact that correlational "noise" structure can either increase and decrease performance depending on the layout of the task-relevant "signal" structure in the population representation [Averbeck et al., 2006]. Second, we explored finer-grained temporal codes (Figure 2.5**c**), which revealed no change in the accuracy of the behavioral predictions. We are careful to note that our results do not imply that trial-by-trial correlational structure is not a limiting factor for some tasks (e.g., [Mitchell et al., 2009; Cohen and Maunsell, 2010]), or that finer-grained temporal neuronal codes in IT are falsified by our data. Instead, our results argue that such ideas do not yet add any measurable value for the real-world motivated set of object tasks explored here.

Our study was not aimed to improve upon the previously documented spatial-clustering of "face neurons" in IT [Desimone et al., 1984b; Tsao et al., 2006; Issa and DiCarlo, 2012]. However, we did explore the idea that IT is not best considered as a distributed neural representation, but that it consists of at least two spatially segregated parts—"face patches" that are *a priori* devoted to "face" tasks (part A) and other parts that are devoted to non-face tasks (part B). Our results are entirely consistent with the hypothesis that "Part A" neurons are heavily weighted in adult face detection tasks. That is, prior to learning face detection, downstream neurons accept inputs that are distributed over all of IT, but in the adult, learned state, those downstream readers will most heavily weight neurons that are best at supporting face detection. This hypothesis is consistent with the idea (e.g., [Tsao et al., 2006]) that "face neurons" (and "face patches") are heavily causal in adult face detection behavior [Afraz et al., 2006]. We also considered a stronger form of domain-specific face processing: that all face related tasks causally depend only on neurons in part A, while all other tasks causally depend on neurons in part B [Tsao et al., 2006]. We tested this idea by restricting the parts of IT the downstream decoders are allowed to read from—decoders for face-related tasks can read only from neurons in part A and decoders for all other tasks can read only from part B. Our results showed that such parcelling did not improve the accuracy of the behavioral predictions. Instead, the (non-significant) trend was in the wrong direction (Figure 2.5**e**). As such, our results do not support the strong face modularity hypothesis, but they do not falsify that idea either.

We are not the first to compare neural responses to object recognition behavior. Using shape similarity judgements some studies have shown agreement between neural representation in monkey IT and perceptual "distances" between parameterized shapes in both monkeys and human [Op de Beeck et al., 2001; Kriegeskorte et al., 2008b]. While pioneering, there is a limit to such qualitative comparisons. Primarily, there is a question as to whether shape similarity is a good surrogate for recognition behavior. But even if that assumption was granted, previous work did not attempt to rule out V4 or even V1 as viable candidates, nor did it attempt to distinguish among the large space of IT-based linking hypotheses.

Other studies focused on documenting IT's computational prowess at invariant object recognition [Hung et al., 2005c; Rust and DiCarlo, 2012]. Absolute accuracy was used as the metric, with IT neural populations having a clear advantage over pixels [Hung et al., 2005c] and over V4 [Rust and DiCarlo, 2010] in discriminating between objects across limited changes in view. Here we show that V4-based rate codes are unlikely to directly underlie all object recognition tasks because they outperform humans on some tasks and underperform in others. This points out the fragility of using performance on a single task as a metric for determining which neuronal linking hypothesis underlies behavior. Absolute performance strongly depends on parameters that control the noisiness of a neuronal population (e.g., number of neurons), making it very difficult to expose the key factors of interest (i.e. which neurons and which features of the neuronal response). For example, we here replicate previous work [Zohary et al., 1994; Hung et al., 2005c; Cohen and Maunsell, 2009; Rust and DiCarlo, 2010] showing that increasing the number of neurons improves performance on our recognition tasks, but we now show that it keeps the relationships between easy and difficult tasks the same. Thus, the *pattern* of performance *across many tasks* emerges as a more robust measuring stick by which we can evaluate different neuronal codes [Johnson et al., 2002].

Our comparison of non-human to human primates deviates from approaches that combine neural recording with behavioral testing in the same subjects [Britten et al., 1996; Luna et al., 2005; Cohen and Maunsell, 2011]. It was motivated by our desire to get both high fidelity behavioral and neuronal population data, a fruitful first-line strategy when a perceptual domain is poorly understood (e.g., [Mountcastle et al., 1969; Johnson et al., 2002]). Such comprehensive characterization of object recognition ability is difficult and time consuming in non-human primates, and current human fMRI lacks the appropriate spatial and temporal resolution necessary for characterizing neuronal population at the level we accomplished here (But see [Kay et al., 2008; Naselaris et al., 2009]).

The fact that monkey neuronal population responses can accurately predict human performance patterns adds evidence to the assumption of highly conserved visual capabilities across the two species [Merigan, 1976; Sigala et al., 2002]. Furthermore, our results show that simple, learned weighted sums of randomly-selected distributed average responses (LaWS of RAD) in non-human primate IT are sufficient to account for human performance, even on object categories outside the realm of typical monkey experience (e.g., planes, cars, boats, etc.). We interpret this to mean that primates share a generic neural representation of "shape" [Kriegeskorte et al., 2008b; Zoccolan et al., 2009], suitable for dealing with the difficulties of identity preserving image transformations without being restricted to object categories and a lexicon that is shaped by each subject's real world experience [Freedman et al., 2001]. Specifically, our results argue that primates share a non-semantic IT visual "feature" representation upon which semantic understanding can be learned, and constitutes a performance bottleneck in primate object recognition. This inference is agnostic as to how much of this feature representation is innate, versus learned during the statistically shared postnatal experience of primates [Li and DiCarlo, 2008].

Our results set the stage for new directions in linking neurons to object behavior. One natural extension is to increase the scope of our images and tasks and explore non-categorical visual properties, such as position, size, and viewpoint variation. This avenue of research is studied in Chapter 3.Another obvious direction is to obtain more precise behavioral data for the images we already tested neurally to look closely at the ability of the LaWS of RAD IT linking hypothesis to predict the image-by-image confusion patterns in humans. Both directions will facilitate more stringent neuronal-to-behavioral comparisons, and increase the resolution at which neuronal linking hypotheses can be distinguished. Eventually, more comprehensive behavioral tests might force us to turn to more complex underlying neural codes that were not necessary here (e.g., fine-timing or synchrony based codes [Engel et al., 2001]), and might open the door for investigating a role for feedback in tasks that require inference (e.g., [Kersten et al., 2004; Oliva and Torralba, 2006]).

More comprehensive behavioral assays will necessitate conducting them in both humans and non-human primates to determine when the cross-species assumption breaks down. As in other sensory areas [Connor et al., 1990; Shadlen et al., 1996;

Cohen and Maunsell, 2010], simultaneous recording from behaving animals will reveal a better estimate of the neuronal population size needed for object recognition, and produce accurate trial-by-trial performance predictions. The LaWS of RAD IT linking hypothesis reported here brings us a step closer to predicting the impact of direct manipulation of IT neurons on object recognition behavior. In such a framework future investigations of the behavioral changes in recognition induced by artificial neuronal manipulation (e.g., [Afraz et al., 2006; Verhoef et al., 2012] can be used to further refine IT based linking hypotheses.

This study sidesteps the important question of how IT neuronal responses are produced. Ongoing work is systematically characterizing the non-linear transformations from retina, through V1, V2, and V4 [Pasupathy and Connor, 1999; Hegdé and Van Essen, 2000; Rust and DiCarlo, 2012; Freeman et al., 2013; Yamins*, Hong*, Cadieu, Solomon, Seibert, and DiCarlo, 2014]. Those approaches need to be combined with the framework presented here to achieve an end-to-end understanding of the neuronal mechanisms that support core object recognition behavior [DiCarlo et al., 2012]. We explore this in Chapter 4.

2.4 Methods

2.4.1 64 object recognition tasks

To characterize human object recognition abilities (which we assume are similar to those of monkeys, see Discussion), we designed a behavioral assessment tool, images and tasks that span the range of human performance in core object recognition. To explore shape similarity, we used objects that can be parsed into basic-level categories with multiple exemplars per category, allowing us to test human performance on coarse and fine discriminations. To explore identity preserving object transformations — the "invariance problem," a hallmark of object recognition [DiCarlo and Cox, 2007; DiCarlo et al., 2012] — we used ray tracing software to photo-realistically render each object while parametrically varying, its position, size and pose. Finally, to insure that the tasks were challenging for current computer vision algorithms, we placed each object on a randomly chosen natural background that was uncorrelated with its identity [Pinto et al., 2011]. To focus on the so-called "core object recognition" —recognition during a single, natural viewing fixation [DiCarlo et al., 2012] — each task image was presented as an 8 degree patch directly at the center of gaze for 100 ms. The culmination of our effort was a set of 64 core object recognition tasks (24 noun labels, each at 2 or 3 levels of variation; see Figure 2.3) that constitutes a reasoned attempt at exploring the power of human object recognition. We do not claim this to be an exhaustive characterization of human object recognition, but as an initial operational definition that can be sharpened and extended to explore other aspects of object recognition and shape discrimination (see Discussion).

2.4.2 Image generation

High-quality images of single objects were generated using free ray tracing software (`http://www.povray.org`; [Plachetka, 1998]). Image consisted 2D projections of 3D models (purchased from Dosch Design and TurboSquid) added to random backgrounds. No two images had the same background, in a few cases the background was, by chance, correlated with the object (plane on a sky background) but more often they were uncorrelated, with the background on average giving no information about the identity of the object.

This general approach allowed us to create a database of 5760 images, based on

64 objects. The objects were chosen based on eight "basic-level" categories (animals, boats, cars, chairs, faces, fruits, planes, tables), with eight exemplars per category (BMW, Z3, Ford, ...). By varying six viewing parameters, we explored three types of identity preserving object variation, position (x and y), rotation (x, y, and z), and size. The parameters were varied concomitantly, each was picked randomly from a uniform range that corresponded to one of three levels of variation (low, medium, and high). For the low variation image set the parameters were fixed and picked to correspond to a fixed view and size of each object centered on the background. For example, cars were presented in their side view, while faces were presented with a frontal view. We did vary the backgrounds however, so that each object was presented on 10 randomly picked backgrounds resulting in a total 640 images. For medium and high variation, we generated 40 images per object resulting in 2560 images per variation (total $5760 = 640 + 2 \times 2560$). Each image was rendered with a pooled random sample of the 6 parameters and presented on a randomly picked background. In sum, object view parameter ranges for the three variation levels were:

- **Low variation:** All objects placed at image center ($x = 0, y = 0$), with a constant scale factor ($s = 1$) translating to objects occluding 40% of image on longest axis, and held at a fixed reference pose ($rxy = rxz = ryz = 0$).
- **Medium variation:** Object position varies within one-half multiple of total object size ($|x|, |y| \leq 0.3$), varying in scale between $s = 1/1.3 \sim .77$ and $s = 1.3$, and between -45 and 45 degrees of in-plane and out-of-plane rotation ($\leq 45°$).
- **High variation:** Object position varies within one whole multiple of object size ($|x|, |y| \leq 0.6$), varying in scale between $s = 1/1.6 \sim .625$ and $s = 1.6$, and between -90 and 90 degrees of in-plane and out-of-plane rotation ($\leq 90°$).

2.4.3 Human psychophysics and analysis

A total of 104 observers participated in one of three visual task sets: an 8-way classification of images of eight different cars, an 8-way classification of images of eight different faces, or an 8-way categorization of images of objects from eight different basic-level categories. Observers completed these 30 to 45 minute tasks through Amazon Mechanical Turk, an online platform where subjects can complete experiments for a small payment. All the results were confirmed in the lab setting with controlled viewing conditions, and virtually identical results were obtained in the lab and web populations (Pearson correlation $= 0.94 \pm 0.01$).

Each trial started with a central fixation point that lasted for 500 ms after which an image appeared at the center of the screen for 100 ms, following a 300 ms delay, the observer was prompted to click one of 8 "response" images that matched the identity or category of the stimulus image. The image presentation time (100 ms) was chosen based on results showing that core object recognition performance increase quickly such that accuracy for that presentation time is within 92% of performance at 2 seconds (see Figure S2.2 in [Cadieu et al., 2014]). Results were very similar with slightly shorter (50 ms) or longer (200 ms) viewing duration. To enforce the need for view tolerant "object" recognition (rather than image matching) each response image displayed a single object from a canonical view, without background. After clicking a response image, the subject was given another fixation point before the next stimulus appeared. No feedback was given. The "response" images remained constant throughout a block of trials that corresponded to one set of tasks (i.e., an 8-way categorization block contained eight embedded binary tasks).

Human object recognition performance was determined by computing a d' for each binary task. Specifically, for a given 8-way task set and variation level (e.g., basic-level categorization at hard variation or car subordinate identification at medium variation), we constructed the raw 8×8 confusion matrix for each individual observer. Then, we computed the population confusion matrix by taking the sum of these raw confusion matrices across individuals. From the population confusion matrix, we computed the d' for each task of recognizing one target class against seven distractor classes (a.k.a. "binary" task). We obtained 72 d' measurements by performing this procedure over all combinations of three task sets and three variation levels (3 task sets × 8 targets per task set × 3 levels of variation). We excluded face-identification at high variation because none of the 8 d's were statistically distinguishable from random guessing, leaving a total of 64 behavioral tasks for the main results presented. Inclusion of these 8 d's had no significant effect on the results. All human studies were done in accordance with the MIT Committee on the Use of Humans as Experimental Subjects.

Typically each human observer only participated in one of the three task sets (basic-level categorization and car and face subordinate-level identification task set; 4 out of 104 subjects participated in both the car and face task sets). For the 8-way basic-level categorization task set, each observer ($n = 29$) judged a subset of 400 randomly sampled images at each variation level (400 out of 640 for low variation and 400 out of 2560 for medium and high variation levels). For the 8-way car ($n = 39$) and 8-way face ($n = 40$) identification task sets, each observer saw all 80 images at the low variation level and all 320 images at both medium and high variation levels. The presentation of images were randomized and counterbalanced so that the number of presentations of each class was approximately the same in the given variation level. Variation levels were presented in successively harder blocks, so observers would see a full set of low variation ("easy") images before moving to medium and then high variation ("difficult") images. On a few additional observers

(n = 10) we interleaved the different task sets (basic categorization, car and face identification at low variation) and we saw no significant effect of interleaving on the pooled population d's (the Pearson correlation coefficient between blocked and interleaved was 0.903 ± 0.057; see the next paragraph for the procedures to compute the pooled population d's).

While no single observer judged all the images in our image database, our pool of human observers did. To compute the pooled population human d's, we started with each observer's data, and computed a $8{\times}8$ confusion matrix for each variation level. We then constructed the population-confusion matrix for each task set and variation level (e.g., 8-way low variation car identification) by summing across individual subject's confusion matrices. We used standard signal detection theory to compute population d's from the pooled population confusion matrix ($d' = Z(TPR) - Z(FPR)$, where Z is the inverse of the cumulative Gaussian distribution function, and TPR and FPR are true positive and false positive rates respectively).

The 64 human population d's were the benchmark to which we compared our candidate linking-hypotheses. To capture the four possibilities for such a comparison (Figure 2.1**b**), we defined two metrics, *consistency* and relative *performance*. To quantify the match between the pattern of predicted and human d's, we computed *consistency*, the rank correlation between predicted d' and actual human d' across all 64 object recognition tasks. To quantify the match on average between predicted and actual human d', we computed relative *performance*, the median ratio of predicted d' and actual human d' across all 64 object recognition tasks. To estimate the human subject-to-subject variability for *consistency* and *performance* we selected one subject from each task set and combined the task performance of the three task sets to produce 64 "individual" human d's (Figure 2.3**b**). We repeated this procedure multiple times to construct an ensemble of individual human subjects. We used this ensemble to compute the median *consistency* (Spearman rank correlation coefficient) and *performance* (1 by definition) between individual human d's and pooled population human d's and the 68% confidence intervals around that median.

To investigate the effect of image subsampling on our results, we computed the sampling induced standard error of the pooled population d's on the basic-level categorization task set. The standard error was minimal (median = 2.1% of corresponding d') since the entire image set was presented multiple times to our large pool of observers ($n = 29$). Assuming the effect of this error to be additive and independent, the predicted consistencies of a linking hypothesis would be increased by ~0.15% if each of our observers judged the entire 5760 image in the basic-level categorization task set.

We also generated two predictions on how *consistency* might improve if we had collected human data on all images. If we assume that the human-to-human *consistency* will eventually be 1 as the number of presented images increases to infinity, the Spearman-Brown prediction formula allows us to estimate the human-to-human *consistency* and its confidence interval (CI) as if we had collected human data on all images in our image set. This assumption resulted in an increase of only ~1.9% to the human-to-human *consistency* and the CI results presented in the main text. If we assumed a more reasonable asymptote of 0.95, the increase was only ~0.59%. In combination the above two analyses suggest that image subsampling in the human basic-level categorization task set had no significant effect on our main results.

2.4.4 Animals, surgeries, and training

The non-human subjects in this experiment were two adult male rhesus monkeys (*Macaca mulatta,* 7 and 9 kg). Before training we surgically implanted each monkey with a head post under aseptic conditions. We monitored eye movement using video eye tracking (SR Research EyeLink II). Using operant conditioning and juice reward, our two subjects were trained to fixate a central red square (0.25°) within a square fixation window that ranged from ±1° to ±2.5° for up to 4s. Outside of maintaining fixation, no additional attempt was made at controlling spatial or feature attention.

We recorded neural activity using 10×10 micro-electrode arrays (Blackrock Microsystems). (96 electrodes were connected, the corners were not connected). Each electrode was 1.5 mm long, and the distance between adjacent electrodes was 400µm. Before recording, we implanted each monkey with three arrays in the left cerebral hemisphere, 1 array in V4 and 2 arrays in IT, as shown in Figure 2.2**b**. Array placement was guided by the sulcus pattern which was visible during surgery. The electrodes were accessed through a percutaneous connector that allowed simultaneous recording from all 96 electrodes from each array (three connectors on each animal). All behavioral training and testing was performed using standard operant conditioning (juice reward), head stabilization, and real-time video eye tracking. All surgical and animal procedures were performed in accordance with the National Institute of Health guidelines and the Massachusetts Institute of Technology Committee on Animal Care.

2.4.5 Monkey behavior, image presentation, and recording procedures

Our goal was to assess neuronal activity patterns that are automatically evoked by presenting a visual image to an awake, alert visual system. Thus, the monkey's only task was to maintain gaze on a fixation dot in the middle of a screen for 2-4 seconds as images were serially presented at the center of gaze. The monkeys initiated each trial by fixating a central red square. After initiating fixation (160ms), a sequence of 5-10 images were presented each for 100ms with 100ms of blank screen in between. Each image was presented at the center of gaze, and subtended 8° of the visual field with a resolution of 32 pixels/deg and a pixel luminance range of 0.3–300cd/m^2. The images were presented in a randomized order, and each image was presented for at least 28 repetitions (typically ~50). We recorded neural responses for 5760 images drawn from the same pool that we used in our human psychophysical testing, with nearly identical visual presentation parameters.

During each recording session, bandpass filtered (250Hz to 7.5kHz) neural activity was monitored continuously, sampled at 30kHz using commercially available hardware (Blackrock Microsystems). The majority of the data presented in this paper were based on multi-unit activity (see Figure 2.8**a**,**b**) for single unit analysis). A multi-unit spike event was defined as the threshold crossing when voltage (falling edge) deviated by less than three times root mean squared error (RMSE) of baseline voltage. Threshold was typically set once during the beginning of a recording session, while the animal was viewing a blank gray screen. Out of 576 implanted electrodes (3 arrays × 96 electrode x 2 monkeys), we focused on the 296 (128 V4 and 168 across PIT, CIT, and, AIT) most visually driven neural sites. To pick these sites we estimated evoked visual response using an independent set of images (typically 795 images with a minimum of 350 images). Visual drive was then defined as the cross-validated average of the top 10% evoked image responses (d' between neural response to image versus blank). Receptive fields were mapped with briefly flashed bars, and expected contralateral receptive field biases were observed in V4.

We recorded all spike time events at all recorded neural sites. As described in the text, we defined different neuronal codes by considering spike counts in different time windows relative to image presentation. Our array placements allowed us to sample neural sites from V4 and different parts of IT. For most analyses, we grouped all sites into either a V4 or an IT population. The response of all neural sites in a population (V4 or IT) to an image formed a vector, the image vectors in turn formed a matrix (described further below) summarizing the population response to all 5760 tested images (Figure 2.2**a**). To fill a response matrix and its multiple repetitions neural responses were collected over multiple days (68 days for M1 and 65 days for M2; stability and its impact on the results is discussed later).

2.4.6 The construction of specific, candidate linking hypotheses and their predicted behavioral performance

A neuronal linking hypothesis is a formal rule for converting neural activity to overt behavior (e.g., a choice of object class). Here, each candidate linking hypothesis learns a neural code that converts neural responses into a prediction of the type of object that is present in the world (as conveyed by the visual image). Defining each linking hypothesis requires the specification of two components: (1) A "response matrix" of neural (or, in some case, computer generated) responses to each and every image. This specification includes which neurons are included (e.g., responses of 100 spatially distributed IT neurons) as well as a specification of the relevant aspects of that neural activity (e.g., time window, mean rate). (2) A specific type of presumed downstream neural decoder along with a training procedure for the decoder that specified how to estimate its final learned state. After specifying these two components for each linking hypothesis, we computed its predicted behavioral performance for each of the 64 object recognition tasks using independent test images.

Response matrix

Neural response matrix: For neuronal linking hypotheses, the response matrix is a $N \times I$ matrix where N is the number of neuronal sites considered to be part of the linking hypothesis and I is the total number of images tested. Since our image set ($n = 5760$) was very large, we collected neural responses piecemeal over multiple days. Each entry of the matrix is the "response" of a particular neural site to a particular image. We considered V4 and IT separately. For each visual area, the "response" was computed as follows. First, we counted the number of spike events elicited by each image, in each neural site, over a given time window. For example, one possibility is the time window 70–170 ms after image onset, but many other possibilities exist, and we explored some of those. From this response, we subtracted the neural site's background response for that day (mean response to "blank" images). Finally, the evoked response of each neural site, was normalized by the site's sample standard deviation (over all tested images that day). This normalization was done to compensate for day-to-day variation and had no effect on pattern of performance and a small effect on absolute performance (~5% increment). The full matrix was collected multiple times (typically ~50 repetitions, minimum 28) and averaged across all repetitions.

Feature response matrix: We also constructed linking hypotheses where the "responses" were simulated, rather than directly measured from neural activity. These included pixels, V1-like simulated neurons, and several popular algorithms in the computer vision community. These algorithms each take an image, and produce the values of a fixed number of "features" (operators on the image). For each algorithm we computed the response of all its feature outputs for each of our images. We treat these feature outputs as analogous to neuronal populations, and thus construct a responses matrix for each algorithm. We explored pixel ($n \simeq 16000$ features that have comparable visual-drivenness as neuronal features), V1-like ($n \simeq 76000$ features, again, visually-driven), PHOG ($n \simeq 3000$ visually-driven features), SIFT ($n \simeq 59000$ visually-driven features), an HMAX variant called SLF ($n \simeq 4000$ visually-driven features), and an L3 algorithm ($n \simeq 4000$ visually-driven features) [Pinto et al., 2011].

Downstream neuronal decoders and training procedures

To estimate what information downstream neurons could easily "read" from a given neural population, we used simple, biologically-plausible linear decoders (a.k.a. linear classifiers, linear discriminants). Such decoders are simple in that they can perform binary classifications by computing weighted sums (each weight is analogous to the strength of synapse) of input features and separate the outputs based on a decision boundary (analogous to a neuron's spiking threshold). The decoders differ in how the optimal weights and decision boundary are learned. We mainly explored two types of linear decoders, support vector machines (SVMs) and correlation-based classifiers (CC). The SVM learning model generates a decoder with a decision boundary that is optimized to best separate images of the target object from images of the distractor objects. The optimization is done under a regularization constraint that limits the complexity of the boundary. We used LibSVM software package [Chang and Lin, 2011] with the linear C-SVC algorithm and L_2 regularization (the regularization constant C was set to 5x10^4 except for the linking hypotheses in Figures 2.5**c**–**e** where the C was optimized by a 3-fold cross validation on training data). The CC (correlation classifier) learning model [Meyers et al., 2008] produces a decoder with the use of the target class center estimated by computing the mean across the target images in the training data. The resulting decoder determines the test image's membership by computing the Pearson correlation coefficient between the class center and the image. Correlation-based decoders are simpler than SVMs in two regards: (1) they are determined by class centers in the training data without mathematical optimization, and (2) they do not have free parameters that are unrelated to the data that impact the optimization procedure [Meyers et al., 2008]. For completeness we also explored simpler single feature decoders (max, 95th quantile, 90th quantile, median). These decoders were built by searching for features based on certain criteria. For example, a "max" decoder is built by finding the feature or neural site that has the best d' for each behavioral task. All of these decoders could potentially be implemented by downstream neurons, as they involve two basic operations: weighted sums of inputs followed by a threshold.

For a given task set (e.g., 8-way basic-level classification), and variation level (low, medium, or high) (see "Human Psychophysics" above for details), the corresponding portion of the response matrix was split into "training" and "testing" sets. The mean and variance of each unit or feature was normalized so that its responses to the training set have zero mean and unit variance. The training set was then used to optimize eight "one-vs-rest" linear decoders by finding weights that would maximize classification performance of each. To construct an 8-way decoder, analogous to what the human observers were asked to do, we applied all eight decoders and scored the decoder with the largest output margin as the behavioral "choice" of the linking hypothesis.

Generating the predicted behavioral performance of each candidate linking hypothesis

After constructing each candidate linking hypothesis (i.e. after learning how to read the "code" for each task), we used the "testing" image set (never seen by the decoder) to generate the linking hypothesis's predicted behavioral performance in each of the three task sets. Each such 8-way classification scheme resulted in an 8×8 confusion matrix summarizing the predicted performance (hits and false arms) of a particular linking hypothesis on a particular task set and variation level. This was done multiple times with at least 50 training/testing splits. The average confusion matrix across all splits was then used to compute linking hypothesis d's exactly analogous to the human d's. We also tested a binary two-way classification scheme more common in the computer vision community. The two alternative schemas resulted in similar absolute performance (~5% difference in average performance level) and the practically identical pattern of performance (~2% difference in *consistency* with humans).

2.4.7 Face Selectivity Index

We defined face selective IT sites as the ones that have absolute face selectivity index (FSI) larger than 1/3. The FSI of a site was computed as the following, where F and NF denote the site's mean response to face and non-face stimuli respectively [Tsao et al., 2006; Issa and DiCarlo, 2012]:

$$FSI = \frac{F - NF}{|F| + |NF|}$$

2.4.8 Stability and assumption of combining neural activity across recording days

To collect a large number of repetitions from the thousands of tested images, we had to collect data from the recording arrays over ~45 days (M1, 43 days and M2, 47 days). While the recording arrays are fixed in tissue and are thus sampling the same cortical location across days, these methods cannot guarantee that the exact same neurons are recorded over all days. Such absolute stability, while desirable, is not strictly required to test the neuronal linking hypotheses that we consider here (which assume randomly selected samples of IT neurons). Nevertheless, we sought to understand if our presented results might be different if the exact same neurons had been recorded over all days. To do this, we compared performance obtained by averaging the neural responses to six presentations of all images collected on the same day (assuming stable set of neurons during the day) to performance obtained by averaging the responses to the same number of image presentations (6 presentations), but sampled randomly from multiple days without replacement (always sampled from the same electrode). Each of the two methods produced a pattern of 64 predicted d' values (as in the main text), and we found that those patterns were very similar —the mean Pearson correlation coefficients between the two sets of performances was 0.908 (± standard deviation of 0.016 across different samples of trials; $n = 64$ d's) for IT and 0.923 (± standard deviation of 0.016) for V4. Thus while it is possible that there is some day-to-day variation of recorded activity on each electrode, that variation is small in that it does not substantially change the pattern of results (e.g., some IT linking hypotheses predict human performance and V4 linking hypotheses do not) and thus is unlikely to change our main result.

2.4.9 *Consistency* and *performance* of neuronal linking hypotheses when objects were presented in the ipsilateral versus the contralateral visual field

Since all arrays were placed in the left hemisphere, we wondered whether performance of our neuronal linking hypotheses was affected by object position in the visual field. To address this we divided the response matrix of each visual area (V4, IT) into two groups based on whether the object centers were in the ipsilateral, or contralateral visual field. We then compared performance on the two groups of images using analogous training and testing procedures to what we used for our main results. Consistent with the known contralateral visual field bias in V4, our results showed a ~20% reduction in *performance* of V4 for ipsilaterally-presented objects, while IT showed only a ~3% reduction. However, even when only considering objects in the contralateral visual field, the pattern of behavioral performance predicted by V4 was still very different than the actual human performance ($Consistency = 0.470 \pm 0.111$, error is computed by sampling over of behavioral tasks and assuming that human pattern of performance does not depend on visual hemifield).

2.4.10 *Consistency* and *performance* of neuronal linking hypotheses when objects were only presented foveally

Since V4 units typically have smaller receptive fields than eccentricity-matched IT units, and the array placements focused on foveal V4, we also wondered whether V4-based linking hypotheses could be improved by restricting our image set to objects positioned close to the fovea. To test this hypothesis we re-measured human behavior and neuronal responses (V4 and IT) for a new set of images that did not contain any variation in object position. We used a total of 32 behavioral tasks —24 low variation tasks (8 basic-level, 8 car identification, and 8 face identification tasks) and 8 new basic-level tasks based on images rendered specifically for this analysis (i.e., objects were rendered with randomly picked pose (rotation around x, y and z) and size parameters, but position (x, y) was fixed at the center of the image). Each linking hypothesis consisted of 58 units, and we used correlation decoders for this analysis. All other details were optimized to obtain best performance. For this set of 32 tasks, the median human-to-human *consistency* was 0.887 (with the 68% CI = [0.740, 0.947] due to the sampling of individuals and object recognition tasks). The *consistency* between the LaWS of RAD IT linking hypothesis and human performance was 0.868 (with a 68% CI = [0.791, 0.909] due to the sampling of behavioral tasks). And the *consistency* between the LaWS of RAD V4 linking hypothesis and human performance was -0.196 (CI = [-0.358, 0.001]). While the performance of the LaWS of RAD IT linking hypothesis was indistinguishable from human subjects in terms of *consistency* ($p = 0.411$, bootstrap test), the LaWS of RAD V4 linking hypothesis had significantly lower *consistency* ($p < 0.001$, bootstrap test). This low *consistency* was not caused by the low performance of the V4-based linking hypotheses (similar to Figure 2.7**a,** open green circles); in 12 behavioral tasks (usually low variation identification tasks), V4-based linking hypotheses *outperformed* the pooled human population. This analysis confirms that the performance of these V4 linking hypotheses is not limited by receptive field size and argues instead for an inferior and potentially more tangled V4 representation [DiCarlo and Cox, 2007; DiCarlo et al., 2012].

2.4.11 Computer vision algorithm based linking hypotheses

We compared our biological results on *consistency* and *performance* to a variety of computational models, including: *The trivial pixel control*, in which the original 256x256 square images were down-sampled into 150x150 pixels and flattened into a 22500-dimensional "feature" representation. The pixel features provided a control against the most basic types of low-level image confounds. All following computer vision features were computed based on this downsized 150x150 pixel features. *An optimized V1-like model*, built on grid of Gabor edges at a variety of frequencies, phases, and orientations [Pinto et al., 2011], each image was represented by 86400 features. *PHOG* (Pyramid Histogram Of Gradients) is a spatial pyramid representation of shape based on orientations gradients of edges extracted with a Canny detector [Lazebnik et al., 2009]. We fixed the angular range to 360 degrees and the number of quantization bins to 40 to produce 3400 dimensional features. *The baseline SIFT computer vision model* provided another control against low-level image confounds [Lowe, 2004]. The SIFT descriptors were computed on a uniform dense grid with a spacing of 10 pixels and a single patch size of 32 by 32 pixels. Each image was represented by 67712 features. *The bio-inspired Sparse Localized Features (SLF)* are extensions of the C2 features from the Serre et al. HMAX model [Riesenhuber and Poggio, 1999a; Serre et al., 2007a; Mutch and Lowe, 2008]. HMAX is a multi-layer convolutional neural network model targeted at modeling higher ventral cortex. Because it is a deep network, HMAX has large IT-like receptive fields. HMAX is one of many existing "first-principles"-based models that attempt to build up invariance through hierarchical alternation of simple and complex cell-like layers. There were 4096 features per image. *L3* is a recent three-layer convolutional neural network, which also has large IT-like receptive fields and which was discovered via a high-throughput screening procedure [Pinto et al., 2011]. We used the top-5 models identified in [Pinto et al., 2011], and the dimensionality of each was 15488, 6400, 2048, 4608, 10368, respectively.

Acknowledgements

This work was supported by DARPA (Neovision2), the NIH (NEI-R01 EY014970 to James DiCarlo) and the NSF (IIS-0964269 to James DiCarlo). Ha Hong was supported by a fellowship from the Samsung Scholarship. We are grateful to Nicolas Pinto for help in creating the framework to generate the images and help with the design and testing of the computer vision models.

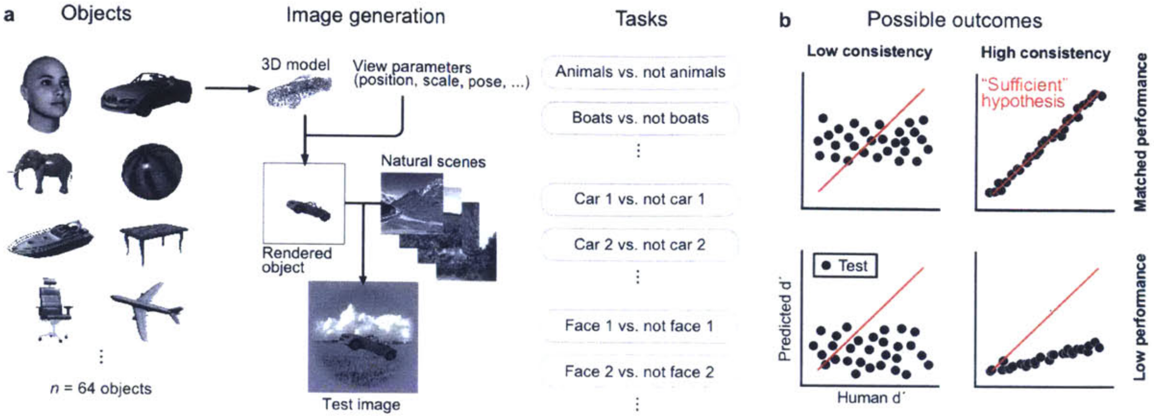


Figure 2.1: (*a*) **Object recognition tasks.** To explore a natural distribution of shape similarity, we started with eight basic-level object categories and picked eight exemplars per category resulting in a database of 64 3D object models. To explore identity preserving image variation, we used ray tracing algorithms to render 2D images of the 3D models while varying position, size and pose concomitantly. In each image, six parameters (horizontal and vertical position, size, rotation around the three cardinal axes) were randomly picked from pre-determined ranges (see Methods). The object was then added to a randomly chosen background. All task images were achromatic. Human observers performed all tasks using an 8-way approach (i.e. see one image, choose among eight; see Methods). Two kinds of object recognition tasks were tested: (1) Basic-level categorization (e.g., "Car" vs. "not car") or (2) Subordinate identification (e.g., "Car 1" vs "not car 1"). We characterized performance for each of 8 binary tasks (e.g., "Animals" vs "not animals", "Boats" vs. "not boats", etc.) in each 8-way recognition block at 2-3 levels of variation, resulting in 64 behavioral tasks (64 d' values). (*b*) **Possible outcomes for each tested linking hypothesis.** We defined multiple candidate neuronal and computational linking hypotheses (Figure 2.5), and determined the predicted (i.e. cross-validated) object recognition accuracy (d') of each linking hypothesis on the same 64 tasks (y-axis in each scatter plot), and compared those results with the measured human d' (x-axis in each scatter plot). A priori, each tested linking hypothesis could produce at least four possible types of outcomes. The pattern of predicted d' might be unrelated to or strongly related to human d' (left vs. right scatter plots). We quantified that by computing *consistency*—the *correlation* between predicted d' and actual human d' across all 64 object recognition tasks. Average predicted d' might be low or matched to human d' (lower vs. upper scatter plots). We quantified that by computing *performance*—the median ratio of predicted d' and actual human d' across all 64 object recognition tasks. For brevity, we will refer to these two metrics as *consistency* and *performance* from here on. Our goal was to find at least one "sufficient" code: a linking hypothesis that perfectly predicted the human d' results on all object recognition tasks (upper right scatter plot).

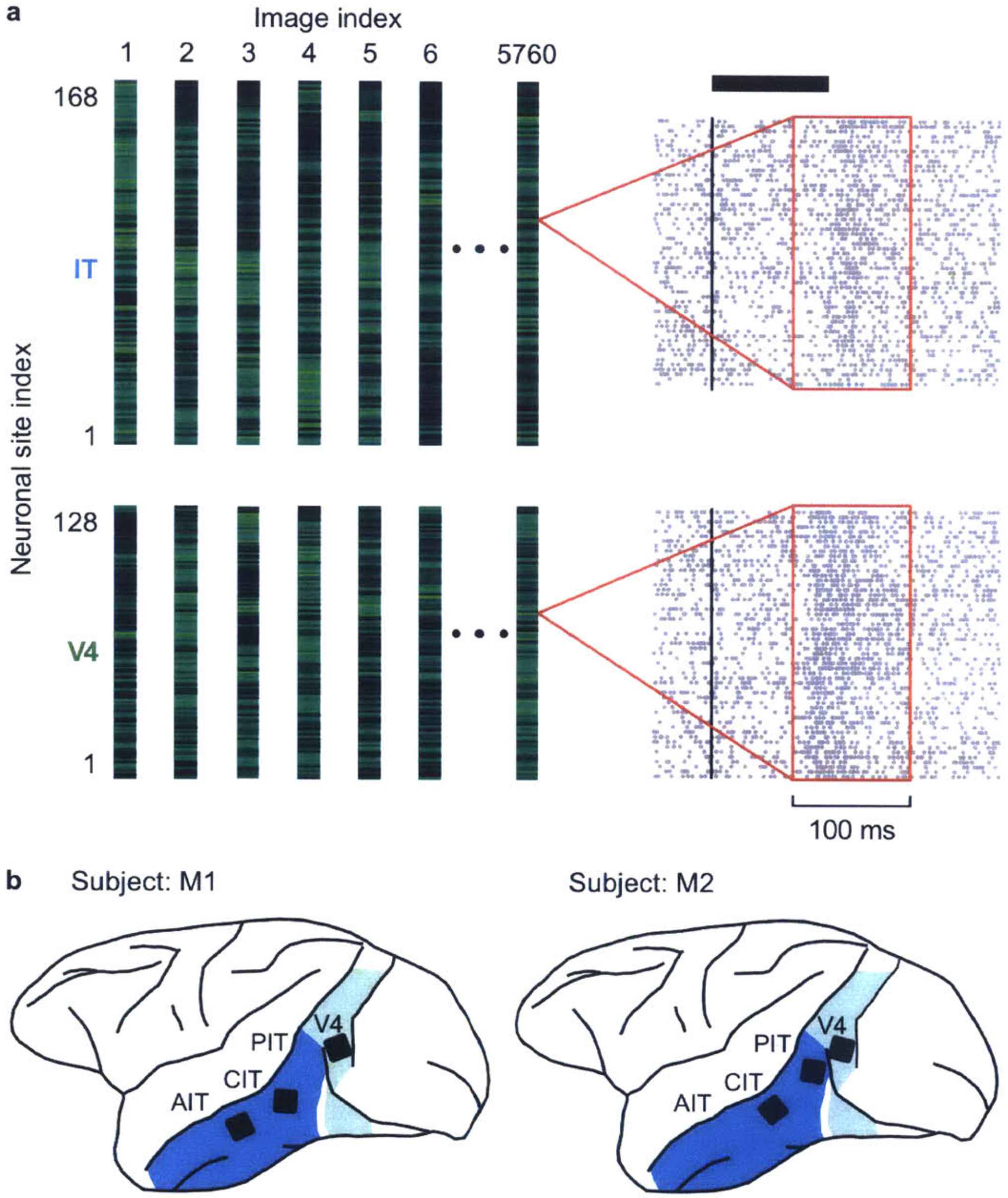

a
Image index
1
2
3
4
5
6
5760
168
IT
1
128
V4
1
Neuronal site index
100 ms
b
Subject: M1
Subject: M2
PIT
V4
CIT
AIT
PIT
V4
CIT
AIT

Figure 2.2 *(preceding page)*: **Neural Responses.** (***a***) We used multi-electrode arrays to record neural activity from two stages of the ventral visual stream (V4 and IT (= PIT + CIT + AIT)) of alert rhesus macaque monkeys. We recorded neural responses to the same images used in our human psychophysical testing. Each image was presented multiple times (typically ~50 repeats, minimum 28) using standard rapid visual stimulus presentation (RSVP). Each stimulus was presented for 100ms (Black horizontal bar) with 100ms of neutral gray background interleaved between images. While some of our neural sites represented single neurons, the majority of our responses were multi-unit (see Figure 2.8**a**). The rasters for repeated image presentations were then tallied within a defined time window (e.g., 70–170ms after image onset, red rectangle, black vertical line indicated stimulus onset) to compute an average firing rate in impulses per second. The mean evoked firing rate is an entry in a response vector (green vertical vector, green saturation is proportional to response magnitude) that summarizes the population response to a single image. The concatenation of the response vectors produces a response matrix representing the population neural response of a particular visual area to our database of 5760 images. We parsed our neural population into V4 and IT, treating the various parts of IT as one population. We recorded from 168 neural sites in IT and 128 neural sites in V4. (***b***) The approximate placement of the arrays in V4 (Green shaded areas) and IT (Blue shaded area) is illustrated by the black squares on two line drawings representing the brains of our two subjects.

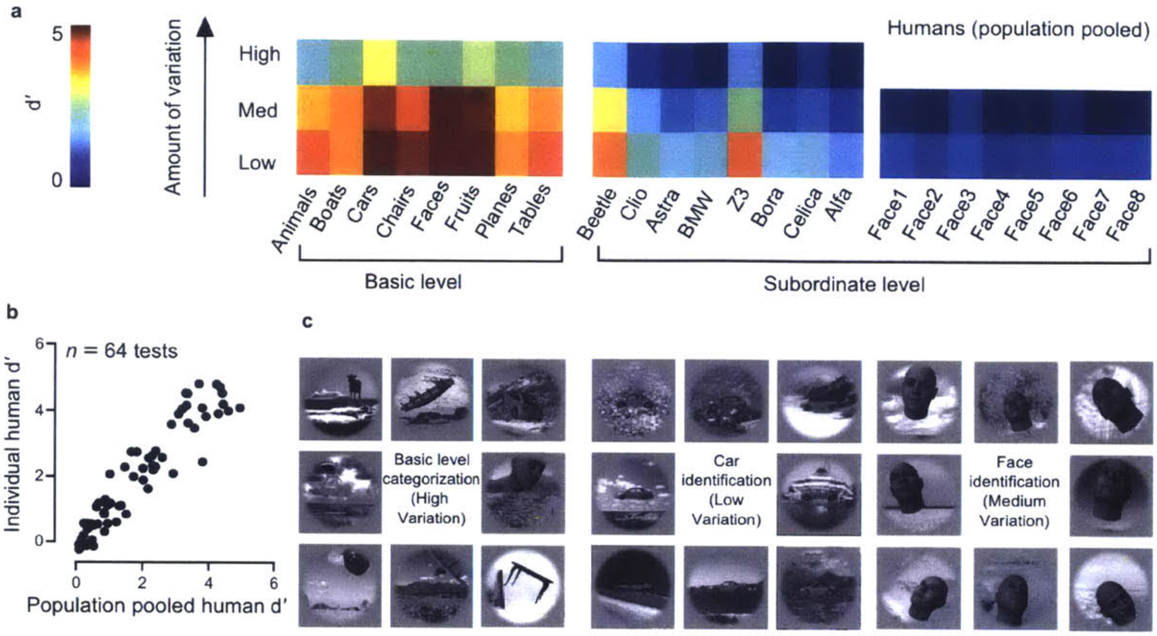


Figure 2.3: **Human core object recognition results.** (*a*) Each color matrix from left to right summarizes the pooled human d' for each of the three task sets ranging from basic level categorization, to subordinate level face identification. In each matrix, the amount of identity preserving image variation was increased from low (bottom) to high (top), resulting in a total of 64 behavioral tasks. Red, represents high performance ($d' = 5$), and blue, low performance ($d' = 0$). For each 8-way task set and each level of variation the computed eight d's were based on the average confusion matrix of multiple observers (Basic level categorization, n=29, car identification, $n = 39$, face identification, n=40; see Methods for more information). (*b*) Human to human consistency. The scattergram shows the performance (d') of one human observer plotted against the performance (d') of the pooled population of human observers across all 64 tasks. The individual human observer was created by randomly combining the performance of three subjects on the three tasks sets (basic-level categorization and car and face subordinate-level identification). The population performance was computed based on a confusion matrix that combined the judgement of our entire pool ($n = 104$) of human observers. The Spearman correlation coefficient in this example was 0.941 (with a 68%-CI = $[0.921, 0.946]$ over the choice of behavioral tasks). Median relative *performance* was 0.999 (with a task-induced 68%-CI = $[0.965, 1.073]$ over the choice of behavioral tasks). (*c*) Example Images. Each octet of images are image samples representing all eight objects used for each of the three tasked task sets at three example variation levels (Basic level categorization (high variation), car identification (low variation), face identification (medium variation)).

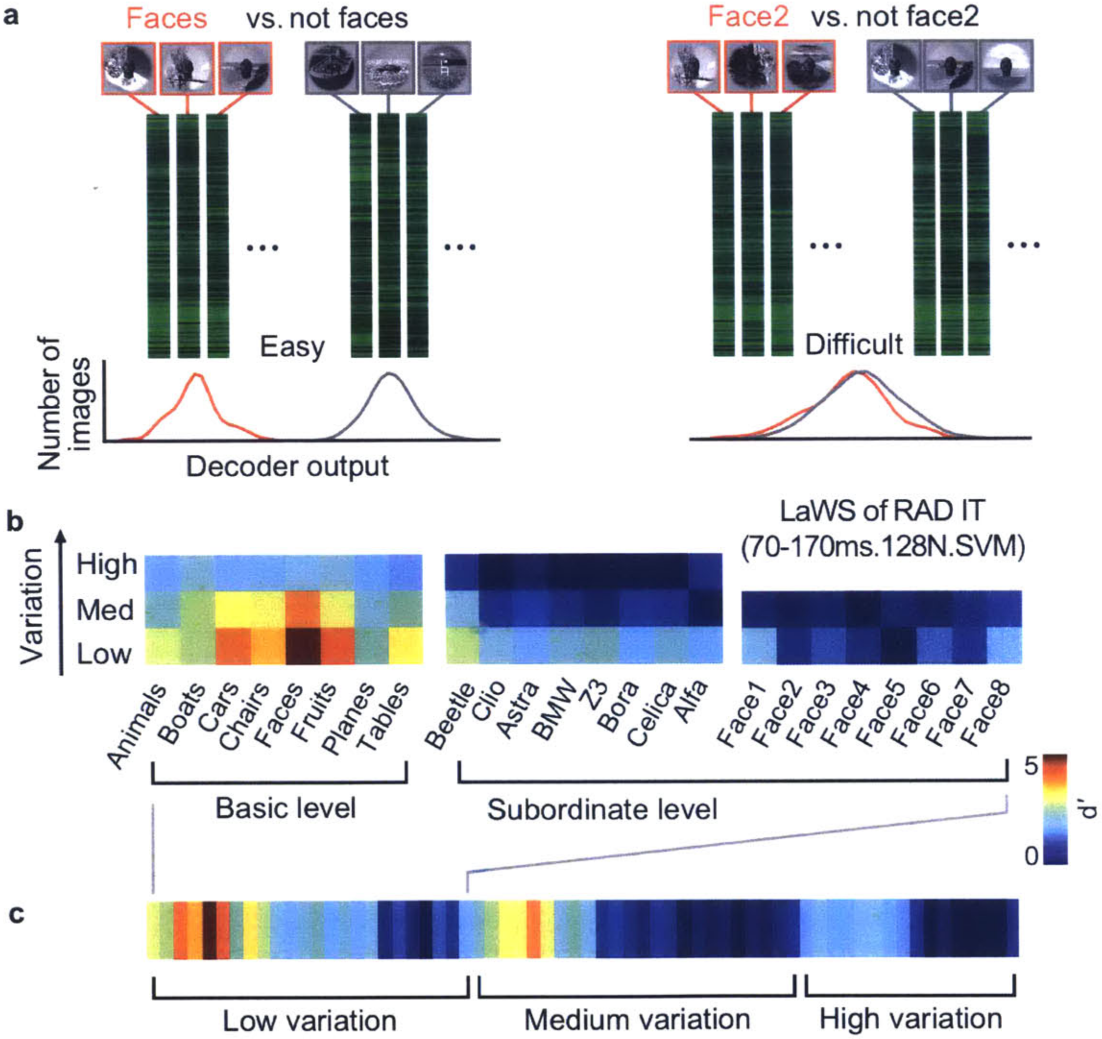

a
Faces vs. not faces
Face2 vs. not face2
...
Easy
Difficult
Number of images
Decoder output
b
LaWS of RAD IT
(70-170ms.128N.SVM)
Variation
High
Med
Low
Animals
Boats
Cars
Chairs
Faces
Fruits
Planes
Tables
Beetle
Clio
Astra
BMW
Z3
Bora
Celica
Alfa
Face1
Face2
Face3
Face4
Face5
Face6
Face7
Face8
Basic level
Subordinate level
5
0
d'
c
Low variation
Medium variation
High variation

Figure 2.4 *(preceding page)*: **Predicted performance pattern of an example LaWS of RAD IT neuronal linking hypothesis.** In this example, the hypothesized neural activity that underlies behavior is: in IT, from 128 units, mean firing rate, in a time window of 70–170 ms; and the decoder is an SVM decoder. (***a***) Based on the aforementioned features of neural activity, a depiction of the outputs of two example decoders for two tasks from two different task sets. For each task set (basic categorization, subordinate identification), and each variation level (low, med, high), we randomly divided our image responses into "training" and "testing" samples. We used the "training" samples, depicted by the green response vectors, to optimize eight "one-vs-rest" linear decoders. The performance of each decoder was then evaluated on the "testing" images. The red and black distributions summarize the response output of two such decoders to a sample of "testing" images. (***b***) Predicted pattern of behavioral performance for all 64 behavioral tasks. To generate these predictions, we constructed an 8-way decoder for each of the three task sets. Analogous to what the human observers were asked to do, for each task set, we applied all eight decoders and scored the decoder with the largest output margin as the behavioral "choice" of the linking hypothesis. Our final *d*'s are the average of at least 50 iterations of randomly picked train/test splits. Similar to Figure 2.3, the color matrices depict predicted performance (*d*′) for this example linking hypotheses for all task sets and variation levels (64 predicted *d*′ values). (***c***) To facilitate comparison amongst different linking hypotheses and with human behavior (see Figure 2.5) we strung out the color matrices into a color vector grouping task sets at each variation level.

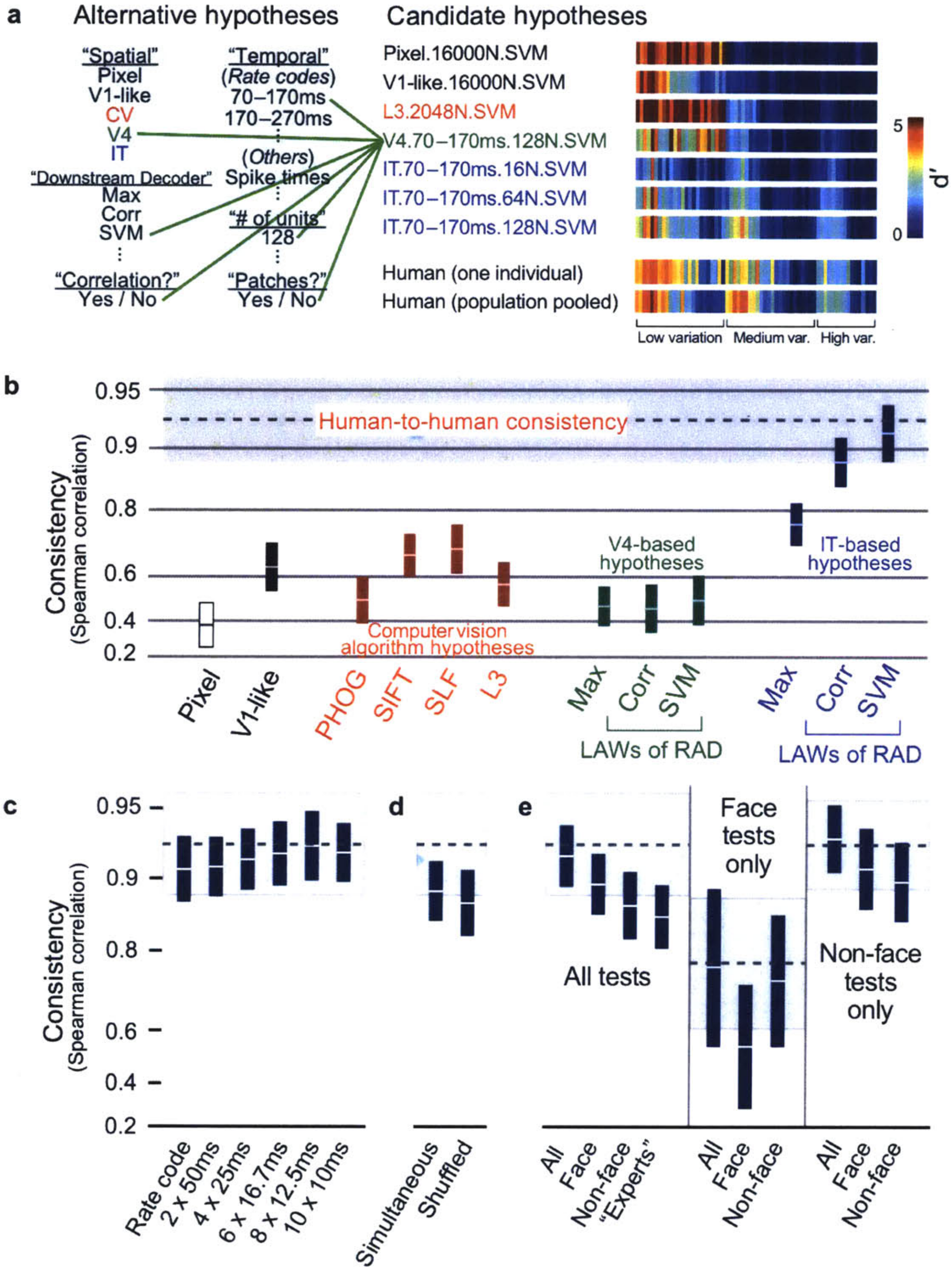

a
Alternative hypotheses
"Spatial"
Pixel
V1-like
CV
V4
IT
"Downstream Decoder"
Max
Corr
SVM
"Correlation?"
Yes / No
"Temporal"
(Rate codes)
70–170ms
170–270ms
(Others)
Spike times
"# of units"
128
"Patches?"
Yes / No
Candidate hypotheses
Pixel.16000N.SVM
V1-like.16000N.SVM
L3.2048N.SVM
V4.70–170ms.128N.SVM
IT.70–170ms.16N.SVM
IT.70–170ms.64N.SVM
IT.70–170ms.128N.SVM
Human (one individual)
Human (population pooled)
Low variation
Medium var.
High var.
5
0
d'
b
0.95
0.9
0.8
0.6
0.4
0.2
Consistency
(Spearman correlation)
Human-to-human consistency
Computer vision algorithm hypotheses
V4-based hypotheses
IT-based hypotheses
Pixel
V1-like
PHOG
SIFT
SLF
L3
Max
Corr
SVM
LAWs of RAD
Max
Corr
SVM
LAWs of RAD
c
0.95
0.9
0.8
0.6
0.4
0.2
Consistency
(Spearman correlation)
Rate code
2 x 50ms
4 x 25ms
6 x 16.7ms
8 x 12.5ms
10 x 10ms
d
Simultaneous
Shuffled
e
All tests
All
Face
Non-face
"Experts"
Face tests only
All
Face
Non-face
Non-face tests only
All
Face
Non-face

Figure 2.5 *(preceding page)*: **Candidate linking hypotheses.** (***a***) The candidate linking hypotheses we explored were drawn from a space defined by four key parameters: spatial location of sampled neural activity, the temporal window over which the response of our units was computed (mean rate in this window), the number of units, and the type of hypothesized downstream decoder. Each candidate linking hypothesis is a specific combination of these parameters. For example, in green is a V4 based linking hypothesis, with a temporal window of 70–170ms, that includes 128 neural sites and uses a support vector machine (SVM) decoder. The predicted performance of each linking hypothesis for each behavioral task is depicted as a color vector where blue signifies low predicted performance ($d'= 0$) and red signifies high predicted performance ($d'= 5$). The goodness of each linking hypothesis can be visually evaluated by comparing its color pattern to that of the human population. (***b***) ***Consistency.*** To quantify the ability of each linking hypothesis to predict the pattern of human performance (i.e. the similarity between color vectors in panel (a)), we computed the Spearman rank correlation coefficient between predicted performance and actual (pooled human, 104 subjects) across all task d's. Median human-to-human correlation is indicated by the dashed line (median Spearman correlation coefficient of 0.929). The gray region signifies the range of human-to-human *consistency* (68%-CI $= [0.882, 0.957]$). Each bar represents a different candidate linking hypothesis (bar length is proportional to task-induced variability). For pixel features (open symbol), V1-like features (filled black symbol), and computer vision features (red filled symbols) we picked the linking hypothesis that performed best. For neural features (V4 (green) and IT (blue)) we matched the number of units at 128. Only bars that enter the gray region correspond to linking hypotheses that successfully predict human behavior. Within the context of IT-based linking hypothesis, we explored finer grain temporal codes (***c***) We also took advantage of our simultaneous multi-electrode array recording to assess the impact of trial-by-trial firing rate correlation on the pattern of performance predicted by our most successful linking hypothesis (***d***) We considered the idea of a modular IT linking hypothesis, with different sub-regions of IT being devoted exclusively to certain kinds of tasks (***e***) First we compared the performance of "face patch likely" sites to "non-face patch" sites on all tasks. We then stitched together an "expert linking hypotheses where each task is performed by neuronal sites that are tailored to that task (e.g., "face" detection is only done by "face neurons" while "car" detection is done by non-face neurons). To be complete, we compared the performance of our different modular IT linking hypotheses on both face tasks only ($n = 17$ of the 64 tasks) and non-face tasks only. Like in panel (b) pattern of performance was always compared to human-to-human *consistency* indicated by the gray region.

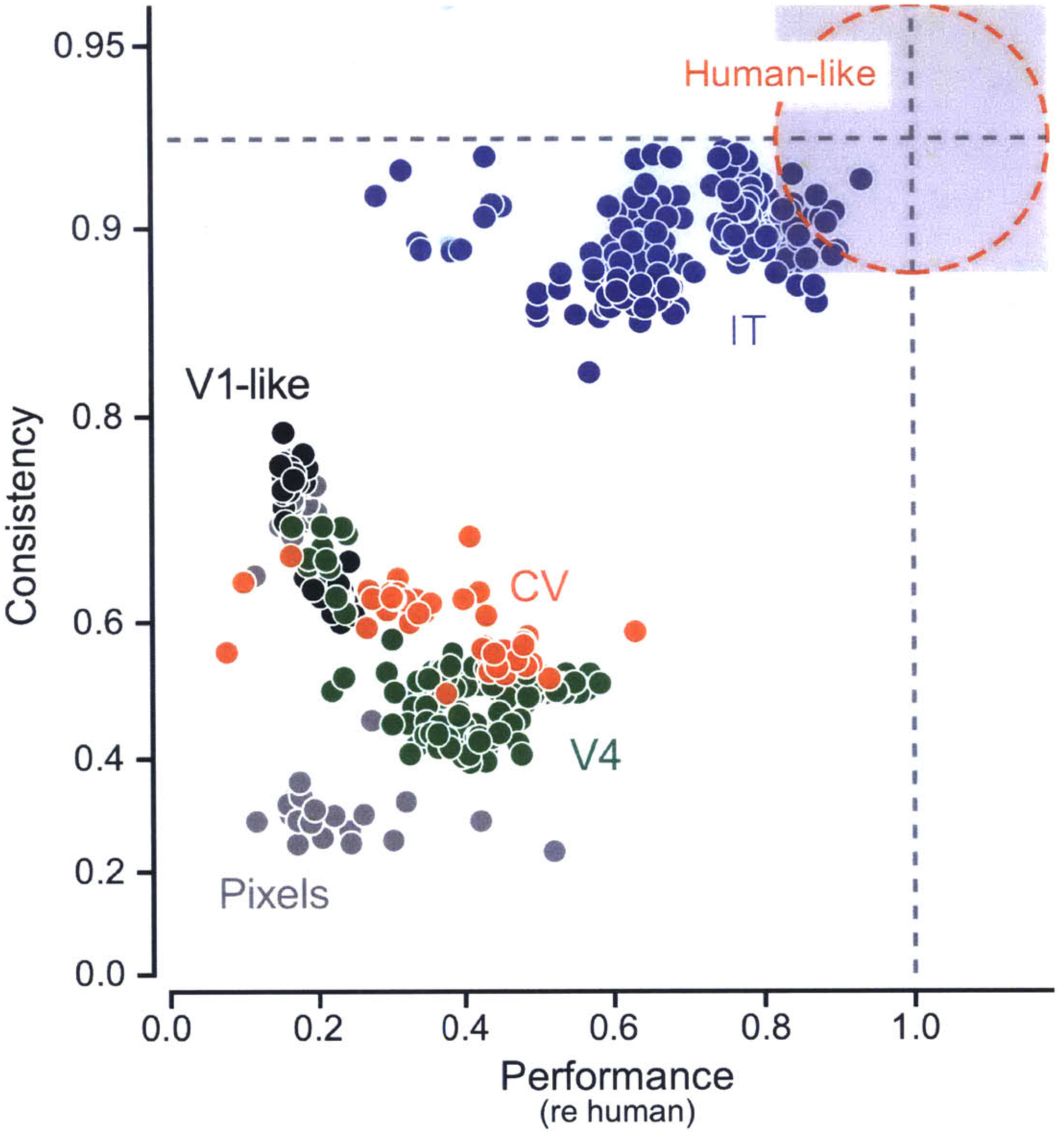

Human-like
IT
V1-like
CV
V4
Pixels
Consistency
0.95
0.9
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Performance
(re human)

Figure 2.6 *(preceding page)*: **Exploring a large set of linking hypotheses.** The y-axis shows *consistency* (defined in Figure 2.5**b**), and the x-axis shows *performance*—the median of the ratio between predicted and actual (human) d' across all 64 tasks. In total, we tested 944 types of linking hypotheses, varying the number of neurons/features in each case, for a grand total of 50,685 instantiations considered. Here we show the results of 755 of those hypotheses. The result of each specific instantiation is shown as a point in the plot with color used to indicate the "spatial" location of the features (IT, V4, V1, or computer vision). We show these examples to illustrate the parameters that we varied which included, spatial location, temporal window, number of units, type of decoder, as well as a variety of training procedures and train/test splits (see Figure 2.10**a**). The horizontal dashed line indicates the average human-to-human *consistency*, and the horizontal grey band represents variability in human-to-human *consistency.* The vertical dashed line indicates the average relative human-to-human performance and by definition is at 1, and the vertical grey band shows the human-to-human variability in relative *performance.* Any linking hypothesis that falls in the red dashed circle is perfectly predicting human performance on these 64 tasks. Note that much of the scatter in the IT-based linking hypotheses (blue) is due to varying the number of neural sites, as illustrated in Figure 2.7**b**.

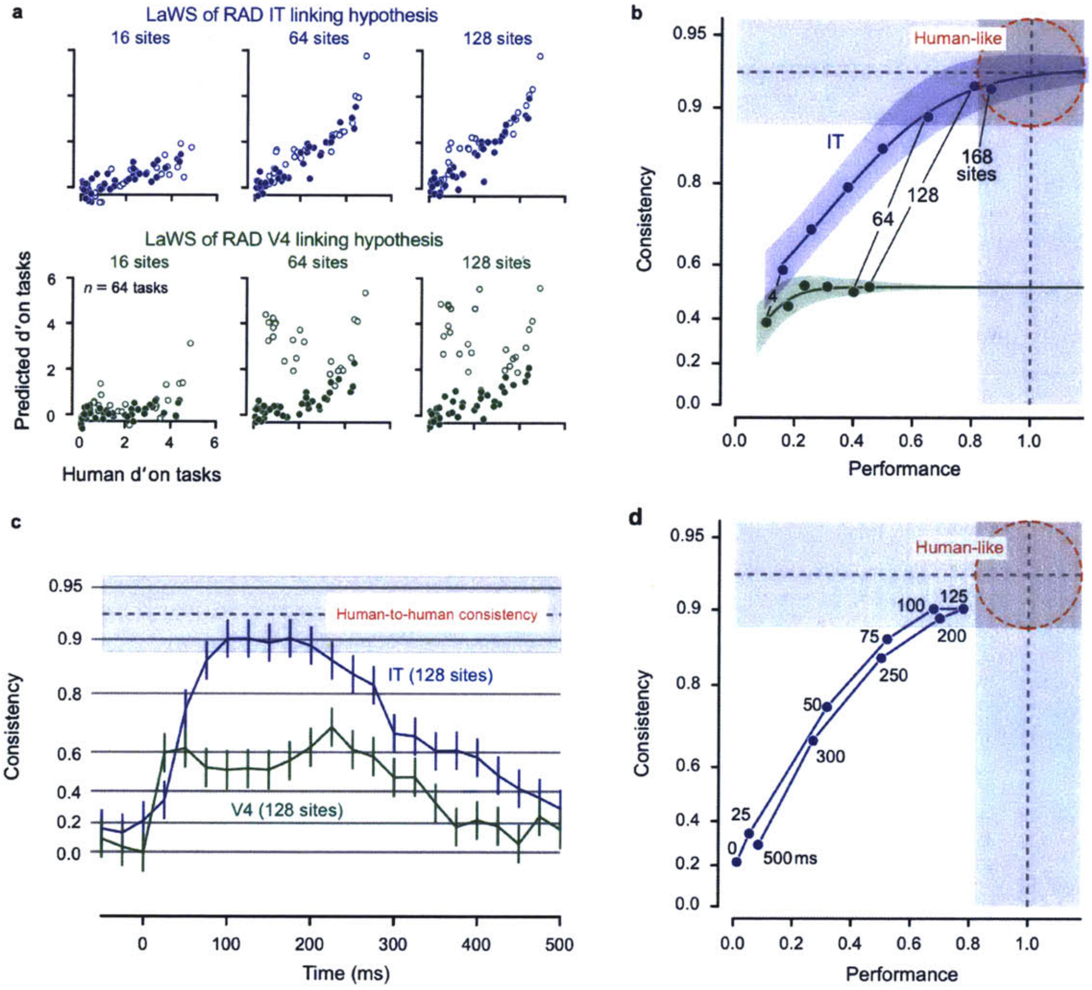

a
LaWS of RAD IT linking hypothesis
16 sites
64 sites
128 sites
LaWS of RAD V4 linking hypothesis
16 sites
64 sites
128 sites
n = 64 tasks
Predicted d' on tasks
0
2
4
6
Human d' on tasks
b
0.95
0.9
0.8
0.6
0.4
0.2
0.0
Consistency
Human-like
IT
4
64
128
168
sites
0.0
0.2
0.4
0.6
0.8
1.0
Performance
c
0.95
0.9
0.8
0.6
0.4
0.2
0.0
Consistency
Human-to-human consistency
IT (128 sites)
V4 (128 sites)
0
100
200
300
400
500
Time (ms)
d
0.95
0.9
0.8
0.6
0.4
0.2
0.0
Consistency
Human-like
0
25
50
75
100
125
200
250
300
500 ms
0.0
0.2
0.4
0.6
0.8
1.0
Performance

Figure 2.7 *(preceding page)*: **Effect of number of units and temporal window on *consistency* and *performance*.** Here we show the results for the LaWS of RAD linking hypotheses (see text), but results are qualitatively similar for other hypotheses. (***a***) The scattergrams show predicted performance (d') of these two neuronal linking hypotheses (IT (blue), V4 (green)) plotted against the actual human performance on all 64 tasks (low variation (open circles), medium and high variation (filled circles). The number of units increases from 16 neural sites (left) to 128 neural sites (right). For each linking hypothesis we also computed its *performance*: the median of the ratio between predicted and actual human performance across all d's for all 64 tasks. (***b***) *Performance* (defined in Fig. 2.6) versus *consistency* for the V4- and IT-based linking hypotheses as a function of the number of (trial-averaged) units. The curve fits are: r^2 of 0.996 for IT; and r^2 of 0.91 for V4. They predict that ~529 IT trial-averaged neural sites and ~22,096 V4 trial-averaged neural sites would match human performance under the LaWS of RAD linking hypothesis. (***c***) *Consistency* for different temporal windows of reading the neural activity. Each point is computed with a 100 ms-wide window, and the x-axis shows the center of that window. The number of trial-averaged neural sites was fixed at 128. (***d***) *Consistency* versus *performance* for the LaWS of RAD IT linking hypothesis at several progressive temporal windows with the center location starting at the time of image onset (0 ms) and up to 500 ms after image onset. The width of the temporal window was fixed at 100 ms (code details are same as (***b***) except the number of trial-averaged neural sites was fixed at 128).

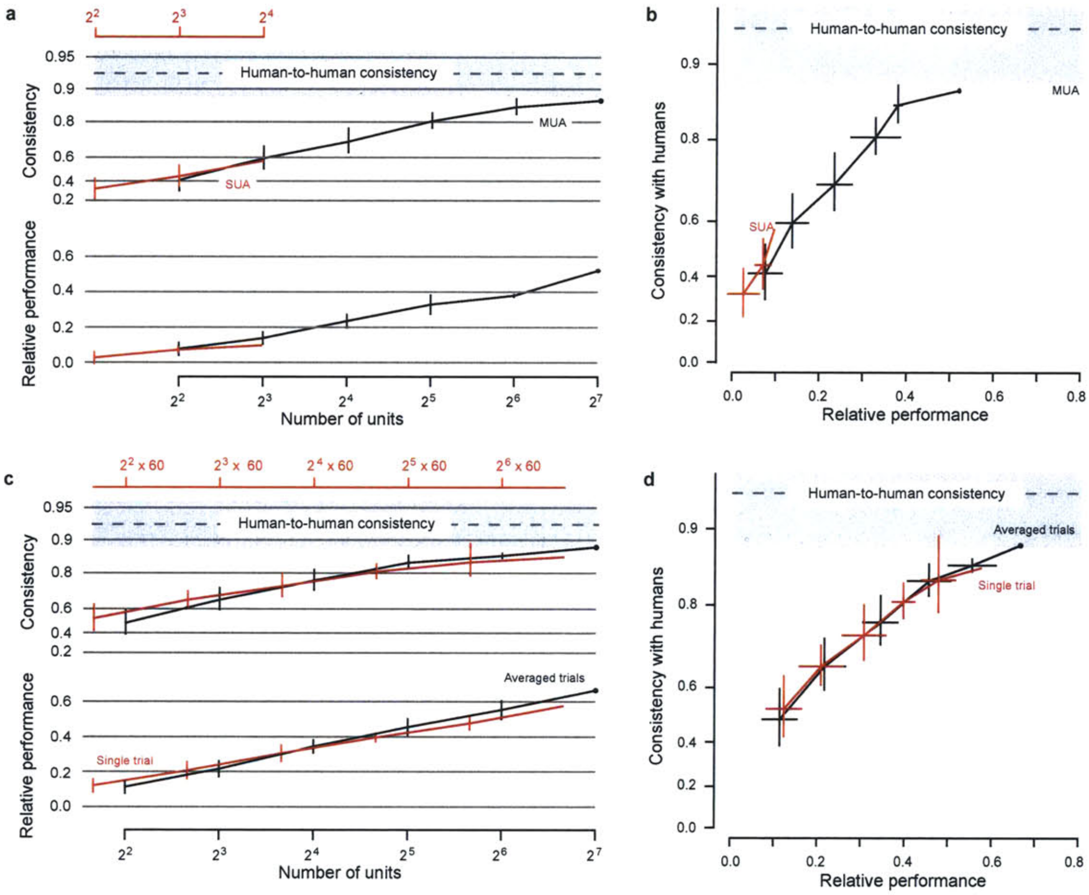

a
2^2
2^3
2^4
0.95
0.9
0.8
0.6
0.4
0.2
Human-to-human consistency
MUA
SUA
Consistency
Relative performance
0.6
0.4
0.2
0.0
2^2
2^3
2^4
2^5
2^6
2^7
Number of units
b
Human-to-human consistency
0.9
0.8
0.6
0.4
0.2
0.0
MUA
SUA
Consistency with humans
0.0
0.2
0.4
0.6
0.8
Relative performance
c
2^2 x 60
2^3 x 60
2^4 x 60
2^5 x 60
2^6 x 60
0.95
0.9
0.8
0.6
0.4
0.2
Human-to-human consistency
Consistency
Relative performance
Averaged trials
Single trial
0.6
0.4
0.2
0.0
2^2
2^3
2^4
2^5
2^6
2^7
Number of units
d
Human-to-human consistency
Averaged trials
Single trial
0.9
0.8
0.6
0.4
0.2
0.0
Consistency with humans
0.0
0.2
0.4
0.6
0.8
Relative performance

Figure 2.8 *(preceding page)*: (***a***) **Single unit activity (SUA) versus Multi-unit activity (MUA) linking hypotheses.** We employed a profile based spike sorting procedure [Quiroga et al., 2004] and an affinity propagation clustering algorithm [Frey and Dueck, 2007b] to isolate the responses of 16 single units from our sample of 168 IT neuronal sites (minimum signal-to-noise ratio (SNR) for each single unit cluster was set to 3.5, with SNR defined as the amplitude of the mean spike profile divided by root mean square error (RMSE) across time points). (***b***) ***Consistency*** **with the human pattern of performance versus *performance* for SUA (red) and MUA (black).** We estimate that twice as many neurons are needed so that the *consistency–performance* relationship of our SUA linking hypothesis matches that of our MUA linking hypothesis. All parameters and training procedures of SUA and MUA based linking hypotheses were identical (performance was based on the average of 5 repetitions using a correlation based decoder (CC) where the units were randomly divided into non-overlapping groups to estimate error from independent sampling of units). (***c***) **Single trial versus averaged trials linking hypotheses.** Because human subjects were asked to make judgements on single image presentations, we also explored a "single trial" training and testing analysis where we treated the responses of the neural units to each images presentation as a new and independent set of neural units (i.e., "unrolled" the trial dimension into the unit dimension). (***d***) ***Consistency*** **versus *performance* for of the single-trial LaWS of RAD linking hypothesis (red) and the averaged-trial LaWS of RAD linking hypothesis (black).** We estimate that ~60 as many neurons are needed so that the *consistency–performance* relationship of our single-trial linking hypothesis matches that of our averaged-trials linking hypothesis. Error bars are standard deviations induced by independent sampling of units as in (a).

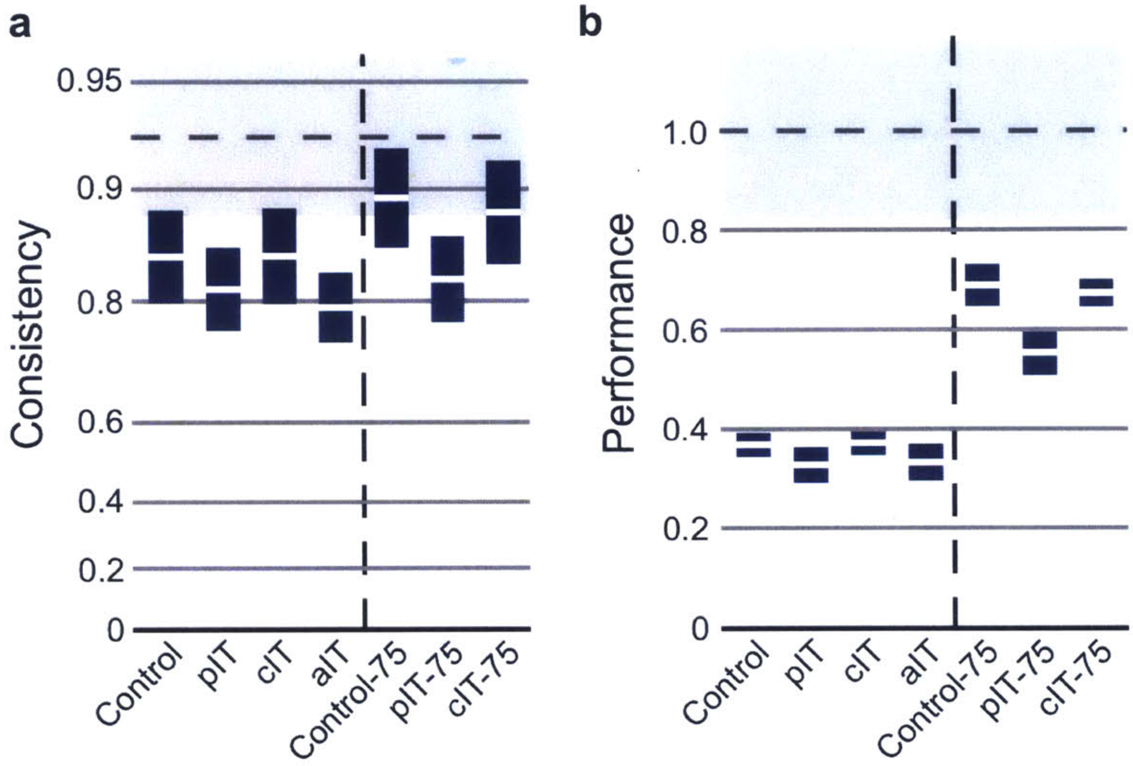


Figure 2.9: **No significant difference in *consistency* and *performance* of IT subpopulations when parsed based on anatomical subdivision: PIT versus CIT versus AIT**. Based on anatomical landmarks we could conservatively divide our population of 168 IT neural sites into: 76 in PIT, 75 in CIT, and 17 in AIT. (***a***) A comparison of the *Consistency* values for IT populations when neural sites respected anatomical boundaries (PIT versus CIT versus AIT) in contrast with a "Control" populations where the sites were randomly picked from all three anatomical subdivisions. There was no significant difference between the IT populations independent of whether we restricted our population to 17 neural sites (limiting our analysis to the number of neural sites in AIT our least sampled anatomical subdivision), or we expanded to 75 neural sites and compared PIT and CIT). Similarly, *performance* (***b***) showed no significant differences between the different IT populations. It is important to note that the decrease in *consistency* and neural *performance* is expected based on the smaller population sizes (see Figure 2.7**b**). *Consistency* and *performance* were computed based on our typical 70–170 temporal window using an SVM decoder.

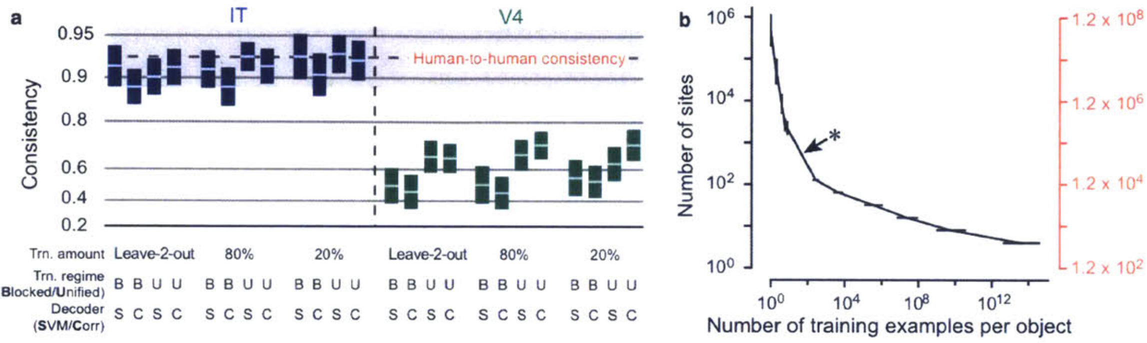


Figure 2.10: (*a*) **The effect of training procedure**. *Consistency* values for LaWS of RAD V4 and IT linking hypotheses under different training procedures. The number of units was fixed to 128 units and the temporal window was 70–170ms after the onset of the image presentation. Two types of decoders were tested (Support vector machines and correlation decoders). We also varied the number of images used to train the decoder (Leave-2-out: for each class, all images but two were used as the training set, and the remaining two were used for testing; 80%: 80% of images were used for training, and the held-out 20% were used for testing; 20%: similar to 80%, but 20% were used for training, and 80% for testing). In the blocked training regime, the training and testing of a decoder was done for each variation level separately. For the unified training regime, the decoders were trained across all variations and tested on each variation level separately. (*b*) **Trade-off** between the sufficient number of units and the number of training images per object for the LaWS of RAD IT linking hypothesis (where temporal window was fixed at 70–170ms and SVM decoders were used). In each data point, the performance of the linking hypothesis was projected to reach the human-to-human *consistency* (within the subject-to-subject variability) and the human absolute performance (relative performance of one). On the y-axis, the numbers shown in black indicate the projected number of repetition-averaged, multi-unit neural sites that are sufficient, while the numbers in red indicate the number of single-trial, single-unit sites that are sufficient (120× larger). For example, the asterisk indicates a LaWS of RAD IT linking hypothesis of $\sim$60,000 single units discussed above, and the plot shows that it would require $\sim$40 training examples per object to learn de novo (with a 68%-CI $\simeq [30, 60]$, not shown in the plot).

Chapter 3

Representation of Non-Categorical Visual Properties in Inferior Temporal Cortex*

Extensive previous research, including our work in Chapter 2, has examined the role of inferior temporal (IT) cortex in viewpoint-invariant object recognition, revealing robustness of the IT neural population's category encoding to identity-preserving transformations. Here we systematically explore IT encodings for object position, size, pose, and a variety of other "identity-orthogonal" visual properties. We find that IT outperforms lower visual areas such as V1 and V4 in estimating all these visual properties, including those (e.g., position) that are normally considered low-level visual features. We also find high IT–human consistency in both cross-task performance patterns and a plausible number of neural sites to match human performance. Information is distributed broadly in the neural population, rather than

*This work has been done in collaboration with Dan Yamins, Najib Majaj, and James DiCarlo.

factored into property-specific units. Our results suggest that IT jointly encodes a spectrum of object-based visual features relevant for scene understanding.

3.1 Introduction

Humans rapidly and accurately process visual scenes from their environment, an ability that is critical to normal functioning. One facet of scene understanding is view-invariant object recognition [DiCarlo et al., 2012], a challenging computational problem because two images of objects in the same high-level category can have vastly different low-level statistics due to variation in object geometry, position, size, pose, lighting, occlusion, clutter, non-rigid deformation, and many other factors [DiCarlo and Cox, 2007]. Extensive research in visual systems neuroscience has uncovered the role of the ventral visual stream, a series of connected cortical areas present in humans and non-human primates, in solving this challenge. The ventral stream is thought to function as a sequence of hierarchical processing stages [Tanaka, 1996; Logothetis and Sheinberg, 1996; Gross, 1994] that encode image content (e.g., object identity and category) increasingly explicitly in successive cortical areas [Vogels and Orban, 1994; DiCarlo and Cox, 2007; DiCarlo et al., 2012]. For example, neurons in the lowest area, V1, are well-described by Gabor-like edge detectors [Carandini et al., 2005a], though the V1 population does not show robust tolerance to complex image transformations [DiCarlo et al., 2012]. In contrast, rapidly-evoked population activity in inferior temporal (IT) cortex, the cortical area at the top of the ventral hierarchy, can directly support real-time, invariant object categorization [Hung et al., 2005a; Rust and DiCarlo, 2010; Yamins*, Hong*, Cadieu, Solomon, Seibert, and DiCarlo, 2014].

Scene understanding involves estimating a variety of other properties besides an object's category or identity [Edelman, 1999; Koenderink and van Doorn, 1979]. However, many of these properties — where is the object? how big is it? what orientation and heading is it at? — are precisely the "nuisance" variables that must be discounted to achieve invariant recognition. Since humans do in fact perceive all these visual object properties in images, this begs the question: what overall neural architecture underlies both the ability to discount identity-preserving variable transformations for object recognition tasks while being sensitive to these same variables for other scene-understanding tasks? One major class of hypotheses [Mishkin et al., 1983; Goodale and Milner, 1992; Ungerleider and Haxby, 1994] is that identity-specific properties (e.g., category membership) are represented in higher ventral cortical areas such as IT, while identity-orthogonal variables (e.g., position) are either represented in lower cortical areas (e.g., V1, [Bosking et al., 2002]) or outside the ventral stream (e.g., the dorsal stream [Mishkin et al., 1983]). These ideas are attractive, because they are consistent with the fact that higher ventral areas have larger receptive fields and are less retinotopic than lower areas, and suggest an intuitively understandable mechanism for how invariance is built in the ventral stream — namely, by aggregating view-tuned units at each physical scale to produce partially view-invariant units that can themselves be aggregated at a larger scale.

However, this separation induces the so-called binding problem, in which multiple separate streams of information would then have to be brought together somewhere in the brain, potentially using feedback connections [Deco and Rolls, 2004; Chikkerur et al., 2010]. A line of theoretical work has suggested that *factored* representation schemes that retain the "nuisance" variable information while still building category selectivity could avoid this binding problem to begin with [Edelman, 1999; DiCarlo and Cox, 2007]. It has also been experimentally observed that IT cortex normally associated with invariant recognition appears to retain some sensitivity to object position [Li et al., 2009; DiCarlo and Maunsell, 2003; MacEvoy and Yang, 2012; Sayres and Grill-Spector, 2008; Sereno et al., 2014] and other properties [Nishio et al., 2014]. However, it is not clear how much and exactly what kinds of non-categorical information is present in higher ventral cortex, nor how these properties are integrated with the categorical representation.

Here, we investigated this issue systematically by recording neural responses in IT and V4 cortex to a large set of visual stimuli containing a range of real-world objects with significant simultaneous variation along object position, size, and pose variables (see Chapter 2 and [DiCarlo et al., 2012; Yamins*, Hong*, Cadieu, Solomon, Seibert, and DiCarlo, 2014]). This image set allows us to characterize neural encodings for standard categorical tasks as well as a variety of identity-orthogonal estimation tasks. We quantify the amount and distribution of information with respect to biologically plausible downstream decoders for each of these tasks at both the single-site and population levels, comparing between cortical areas as well as to psychophysical measurements of human behavior. We find that for all tasks we investigated, including those normally considered low-level (e.g., position), more information is accessible in IT than in V4, which in turn has more accessible information than a V1-like model. We also find that the IT population performance pattern is more consistent with human behavioral measurements that those from lower layers. Moreover, information for all these tasks in IT appears generally to be well-distributed, as opposed to highly concentrated in task-specialist subpopulations. In addition to these experimental findings, in Section 4.2.7, we also describe a computational model of the ventral stream that explains the results from simple assumptions. Taken together, our results strongly favor a joint-encoding hypothesis in which the ventral stream builds explicit representations both for categorical and non-categorical visual object properties simultaneously.

3.2 Results

3.2.1 Battery of visual tasks

We continued to use our main neural test stimulus set of Chapter 2, which consisted of 5760 images of 64 distinct objects chosen from one of eight categories (animals, boats, cars, chairs, faces, fruits, planes, tables), with eight specific exemplars of each category (e.g., BMW, Z3, Ford, etc. within the car category). The set was designed specifically to (1) include a range of everyday objects, (2) support both coarse, "basic-level" category comparisons (e.g., "animals" vs. "cars") and finer subordinate level distinctions (e.g., distinguish among specific cars) [Rosch et al., 1976], and (3) require strong tolerance to object viewpoint variation, e.g., pose, position and size. The objects are shown at high levels of the position, scale and pose variation on cluttered natural scene backgrounds that are randomly selected to ensure that background content is uncorrelated with category identity (see Fig. 3.1a and 3.7a; also, see Section 2.4.2 for additional details). The high levels of variation expose key dimensions that make invariant object recognition challenging for artificial vision systems, but to which humans are robustly tolerant [DiCarlo and Cox, 2007; Pinto et al., 2008a; Yamins* et al., 2014].

The simultaneous variation of object properties in the image set allows a battery of discrete-valued classification tasks and continuous-valued "identity-orthogonal" visual estimation tasks (Fig. 3.1b). We first defined the following discrete-valued visual tasks, as in Chapter 2:

- **Basic-level object categorization.** This is a discrete-valued eight-way object categorization task of Chapter 2, in which the goal is to report the category of the object in the image, from the set of choices: Animals, Boat, Car, Chair, Face, Fruit, Plane, Table.

- **Subordinate-level object identification.** These are discrete-valued eight-way object identification task, in which the goal is to report the specific identify of an object in each image from the list of eight exemplars of that object's category. There are eight such tasks, one for each category in the dataset. For example, in the case of the car category, the eight-way subordinate-level object identification task is identify an image as containing one of: Beetle, Alfa Romeo, Vauxhall Astra, BMW 325, Maserati Bora, Toyota Celica, Renault Clio, or BMW z3. In Chapter 2, we studied car and face subordinate tasks. Here, he considered subordinate tasks from all other categories (e.g., Tables, Boats; but not cross pairs such as Table 1 vs. Boat 2).

In addition to the above, we also introduced the following set of *non-categorical* visual tasks:

- **Position Estimation.** These are a set of related continuous-valued location estimation task, in which the goal is to identify an object's center location. Tasks are to identify the location in pixels from the object center, along the x-axis ("X-Axis Position") and the y-axis ("Y-Axis Position"), and the distance in linear pixels of the object center to any fixed point location ("Center Distance").

- **Bounding-box location and size estimation.** These are a set of related continuous-valued bounding-box related tasks. The bounding box for an object is defined to be the smallest axis-aligned rectangular subset of the image that fully contains the pixels of the object. Location of each corner is measured, as is the size in linear pixels along both axes ("X-Axis Size" and "Y-Axis Size", respectively). The area of the bounding box in square pixels is also measured ("Bounding Size").

- **2-D Retinal Area.** This continuous-valued task measures the area in square pixels that the object takes up in the image. Each image pixel is either covered by the object, in which case the pixel is counted toward this metric, or it is not covered by the object, in which case the pixel is not counted. For example, pixels surrounded by an object but not actually covered by it (e.g., the hole of a donut) do not count toward this measure.

- **Perimeter.** This continuous-valued task measures the area in linear pixels on the boundary of the object. Pixels in the object not completely surrounded by other pixels also in the object do count toward this measure; any other pixels do not count.

- **3-D Object Scale.** This continuous-valued task measures the 3-D scale parameter used to generate the image in the original rendering process, relative to a fixed canonical size — namely, $s = 1$ in the object parameterization discussed above. This relationship of this property to the 2-D retinal area depends in a complex manner on the object's geometry.

- **Major Axis Length, Aspect Ratio and Angle.** The major axis of an object is defined to be the longest line segment such that both ends of the line segment are pixels within the object. The minor axis is the shortest perpendicular line segment so that the rotated bounding box defined by the major and minor axes covers the object. The continuous-valued measure axis length is measured in linear pixels. The aspect ratio is the ratio of the lengths of minor to the major axis. The major-axis angle is the 2-dimensional angle, in degrees, made by the major line with the horizontal ($x = 0$) line.

- **3-D Rotation.** These three rotations are the angles, in degrees, used by the renderer to orient the object in the original image creation process. The angles are described via standard Euler rotations using the XYZ order. The $(0, 0, 0)$ rotation is defined separately for each of the 64 exemplar objects in the dataset. However, the exemplar angles are fairly well-defined "semantically", meaning that they are reasonably consistent across the eight exemplars each for the eight basic object categories. Specifically, for each category the $(0, 0, 0)$ angle is the one in which:
  - Animals: animal is facing forward, with its head upright.
  - Boats: boat is oriented with bow facing forward and keel point downward.
  - Cars: car grille is facing forward, while tires on the bottom.
  - Chairs: chair legs are facing downward, with the seat facing forward.
  - Faces: looking straight the viewer, with top of the head oriented upward.
  - Fruits: stem attachment at the top. [Note that many of the fruits possess a rough rotational symmetry around the vertical axis.]
  - Planes: cockpit facing forward, with plane in upright position.
  - Tables: table legs facing straight downward, with longest side along the x-axis.

For the first two discrete-valued categorization tasks, performance is measured using *balanced accuracy*. Balanced accuracy is defined for a prediction of binary task with positive and negative classes as:

$$\textbf{AccBal} = \frac{TP}{P} + \frac{TN}{N} - 1$$

where TP is the number of correct positive predictions, P is the number of positives examples in the data, TN is the number of correct negative predictions, and N is the number of negative examples in the data. Balanced accuracy for a multi-class prediction problem is the average of one-vs-all (OVA) prediction problems over the classes. For continuous-valued estimation tasks, performance is measured as the Pearson product-moment correlation between the predicted and actual values. Specifically:

$$\textbf{Corr} = \frac{covariance(\vec{p}, \vec{a})}{\sqrt{variance(\vec{p}) \cdot variance(\vec{a})}}$$

where $\vec{p}$ is the vector of predictions for a sequence of images and $\vec{a}$ is vector of corresponding ground-truth values for that property. Both metrics range from -1 to 1, with 0 being chance-level prediction and 1 being perfect prediction.

3.2.2 Large-scale array electrophysiology in macaque higher visual cortex on a high-variation stimulus set

In Chapter 2, we collected a large-scale data of macaque IT and V4 neural response to the stimulus set of 5760 images described above. Here, we collected an additional set of neural data by chronically implanting three more multi-electrode arrays. Combined with the previous data, using nine chronically implanted electrode arrays in total, we collected responses from 309 neural sites in cortical area IT and 211 neural sites in cortical area V4 to each image in the set (see Fig. 3.1c and Methods). In this chapter, we used most visually driven 266 IT and 126 V4 neural sites (see 2.4.5 for the definition of visual drivenness). The cut-off threshold was chosen to precisely include all of Monkey 1 data in Chapter 2, which had higher visual drivenness than Monkey 2 data. We then investigated the ability of these neural populations and simulated low-level visual area populations (V1-like model and pixels, see Methods) to support both the categorical and non-categorical tasks discussed above.

3.2.3 Comparing categorical and non-categorical task representations across cortical areas

For many tasks, including object category, position, size and pose, we found individual sites in our IT sample whose responses contained reliable information for that task, despite simultaneous variation in all these variables (Fig. 3.2a-c). For binary categorical tasks, we defined single-site performance as the absolute value of the site's discriminability for the task on a set of held-out test images (see Methods). For continuous-valued estimation tasks, we defined single-site performance as the absolute value of the Pearson correlation of that site's response with the actual value for task, again on a set of held-out test images. For most tasks, the best sites from our IT cortical sample contained significantly more information than those from our V4 sample (Fig. 3.2d).

Because information about visual properties is likely to be spread over multiple neural sites, we next investigated the extent to which our battery of tasks was encoded at the neural population level (Fig. 3.3a). For discrete-valued tasks, such as basic categorization and subordinate identification tasks, we used linear classifiers to identify thresholded weighted sums of the neural populations that best predicted category membership on a set of training images, and evaluated prediction accuracy on held out test images [Hung et al., 2005b; Pedregosa et al., 2011]. For each continuous-valued estimation task, we used linear regression to identify weighted sums of the neural population that best estimated the continuous target values on a set of training images, and then evaluated the Pearson correlation between predicted and actual property values (0=chance, 1=perfect), again on held-out test images (see Methods and [Pedregosa et al., 2011]). Once trained, both classifiers and regressors can be considered as a specific linking hypothesis that might explain how the downstream brain area reads visual properties to solve visual task, as we discussed in Chapter 2. That is, they form LaWS of RAD (Learned Weighted Sums of Randomly-selected Average neuronal responses spatially Distributed; see Section 2.2.2) linking hypotheses.

Population performance levels were higher than that from individual sites, as would be expected. Moreover, as with the single-site data, the IT population (Fig. 3.3a, blue bars) significantly outperformed the V4 population (green bars) on all tasks. To compare the results from these higher visual areas to lower-level visual response properties, we also evaluated a Gabor-wavelet-based V1 model [Pinto et al., 2008b] on our stimulus set (gray bars, and see Methods). In all cases, the IT sample population outperformed the V1-like model, and in most cases, the V4 population did as well. The trivial pixel control (black bars) performed least well in nearly all cases. Results were evaluated for each task using an equivalent number of sites ($n = 126$). We performed several controls to ensure that the differences between IT and V4 were not due to differences in receptive field size, sampling sparsity, or number of training examples used to train the classifier (see Methods and Fig. 3.8).

We also recorded V4 and IT neural responses on a simpler stimulus set consisting of grating-like patches placed on gray backgrounds at varying positions and orientations (Fig. 3.3b and 3.7b). From this data, we then measured decoding performance for x-position, y-position and orientation estimation tasks in the IT and V4 populations. We found that performances were typically higher than chance. However, in contrast to the results shown in Fig. 3.3a for complex stimuli, for these simpler stimuli the IT population was not better than the V4 population on position tasks, and both IT and V4 populations were significantly less good than the V1-like model. This comparison clarifies our main result in relation to existing data on low-level task performance in early visual areas [Carandini et al., 2005a]: while the larger receptive fields in IT and, to a lesser extent V4, indeed do lose resolution for low-level pixel-level judgements needed for the simplified stimuli, this type of information loss does not strongly interfere with the abilities of neuronal populations to decode apparently similarly-defined properties (e.g., position or orientation) in more complex image domains.

3.2.4 Consistency of the IT neural encoding with human performance patterns

We also collected human performance data on a variety of tasks in the task battery, including categorization, position, size, pose, and bounding-box estimation (see Methods). We then sought to characterize, for each neural population and each task, how many neural sites would be required to reach parity with human performance levels. For the V4 and IT neural populations, as well as the V1 model and the pixel control, we subsampled sites to produce performance curves as a function of population size, for each task (Fig. 3.4a). For the IT and V4 neural populations, we produced curves out to the limit of the neural data (266 and 126 units, respectively), while for the V1 model (and pixel control) we sampled increasing numbers of units until performance saturated. We then fit each task's neural performance curve to a logarithmic functional form to extrapolate performance levels at sample sizes beyond that in our data (see Methods for details). For the IT population, all tasks had roughly similar logarithmic growth rates, with the predicted IT performance curve intersecting the human performance level at less than 2000 multi-unit sites (Fig. 3.4b), with a mean across tasks of 704 sites. This result suggests that each additional IT unit typically contributes approximately the same amount of additional performance benefit for each task in our task battery. In contrast, the V4 representation performance curves were more variable and in many cases would require several orders of magnitude more sites to match human performance. The V1 representation typically requires at least several orders of magnitude more sites in addition, in many cases unrealistically many more (i.e., greater than 10^{10} sites). The pixel representation is not within realistic bounds for any task in our task battery.

To investigate the neural-behavior link at a more detailed level, we compared the performance *patterns* between human subjects and neural populations using a fixed decoding mechanism. That is, we sought to determine whether the relative difficulty of tasks for humans across our range of tasks corresponded to the relative difficulty predicted by the neural populations (Fig. 3.5a). We constructed a vector of performances (with vector length being the number of tasks), for the human subject pool as well as each neural population. We then computed the Spearman rank correlation between the neural performance vectors and the human performance vector (see Methods). We found that the pattern of IT population performances is a significantly better predictor of the human performance pattern than that from the other cortical areas (Fig. 3.5b). Together with the above result on parity-size estimation, this result shows that IT is likely to be more directly responsible for downstream behavior-generating neurons than lower cortical areas, across the spectrum of non-categorical as well as categorical tasks.

3.2.5 Distribution of information across IT sites

We next sought to characterize how the IT neural population simultaneously represents multiple visual properties. Are properties estimated by dedicated subpopulations of neurons that separately solve individual estimation problems, or instead tightly integrated in a joint population representation with the encodings for different tasks highly overlapping with each other?

We first considered the distribution of information across sites for each task separately (Fig. 3.6a). For this analysis, we used the weights assigned to each site by that task's linear estimator as a proxy for the amount of information contributed by that site for that task. If the linear estimator for a given task assigns a given site a high absolute value weight compared to the weights of other sites, that site is taken to be more relevant for the task; high positive values correspond to strong correlation between the site's output at the task, and high negative values correspond to strong anticorrelation. (See Methods for discussion of alternative proxy metrics.) We characterized each task's site-weight distribution using two statistical metrics: *skewness* and *kurtosis*. Skewness is a measure of the balance of the site-weight distribution, with high values indicating a bias towards sites that are anticorrelated with the task, and low values indicating the opposite. Kurtosis is a measure of the sparseness of the site-weight distribution, with high values indicating that only a very few sites are highly informative for the task, and low values indicating little differentiation between sites. Across all tasks, we found a minimum skewness of -0.85 and a maximum skewness of 1.04, with a median of 0.10 (Fig. 3.6b-c, top panels). These skewness values indicate that the number of sites weighted at above average level is no less than 70% and no more than 130% of the fraction of weights below the mean value. For a majority of tasks (68 out of a total of $N = 108$ individual tasks), the skewness values were not statistically distinguishable from that of equally-sized samples from a standard normal distribution. Across all tasks, we found a minimum kurtosis of -0.26 for the face-detection task, and a maximum kurtosis of 4.73 for the x-axis position estimation task; the cross-task median was 0.45 (Fig. 3.6b-c, bottom panels). These kurtosis values correspond to the fraction of highly-weighted sites making up between 15% and 35% of all sites, with a mean of 26.3%; 47 of 108 tasks having sparseness statistically indistinguishable from that of the standard normal distribution. Overall, these results suggest a picture in which the encoding of each task in the IT population is comparatively well-distributed and not especially sparse.

We then quantified *information overlap* between pairs of tasks. We defined overlap as the correlation of the absolute values of the weight vectors for each task pair (Fig. 3.6d; and see Methods for discussion of alternative overlap metrics). A high positive overlap between weight patterns for a task pair (red color Fig. 3.6d) indicates that downstream neurons could use overlapping sets of neurons in similar ways when reading out the two tasks, whereas high negative correlation (blue color) would indicate that downstream neurons would likely need to draw on non-overlapping sets of neurons. Across all pairs of tasks in our dataset, the maximum observed overlap was 0.82, the minimum is -0.13, and the median is 0.07. 56.5% of pairs have positive overlap, 16.6% have negative overlap, and 26.9% have overlap statistically indistinguishable from 0. Unsurprisingly, high overlap tends to occur between groups of highly semantically related tasks (e.g., the various size-related tasks). However, even apparently unrelated tasks typically had more overlap than would be expected from purely random distribution of units (see Methods and Fig. 3.10). An exception is case of the face-detection task, where the true overlap with other categorical tasks is significantly less than random. Taken together, these results suggest that, holding faces aside, the IT neural population jointly encodes both categorical and non-categorical visual tasks using an integrated representation in which many units participate in tasks.

3.2.6 Computational modeling

Recent work has shown that neural responses in higher ventral cortical areas can be modeled effectively by hierarchical convolutional neural networks that are optimized for performance on challenging high-variation categorization tasks [LeCun and Bengio, 1995; Yamins* et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014]. To determine how consistent these models were with our results on non-categorical properties, we implemented one such model containing six hierarchical layers in Chapter 4, specifically in Section 4.2.7. Here, we reproduced key results for the reader's convenience:

- Each layer in the model is composed of extremely simple, biologically-plausible operations include template-matching, non-linear activation thresholding, and local pooling (see Section 4.2.7, Fig. 4.18; for methods, see [Yamins*, Hong*, Cadieu, Solomon, Seibert, and DiCarlo, 2014]). The layers are stacked hierarchically to produce increasingly complex transformations of input image stimuli. The model is convolutional, meaning that it is applied identically at every point in the image stimulus, but becomes less retinotopic at each layer as the effective receptive field size of each unit becomes larger.

- We investigated the model's performance on the same tasks described above. We found that test categorization performance increased throughout the course of training (Fig. 4.19, grayed panels). More surprisingly, we found that performance on all non-categorical tasks also increased during training, and that performance on non-categorical tasks was highly correlated with categorization performance across training timepoints (Fig. 4.6a-b, Fig. 4.19 white panels, and 4.20).

- We also investigated performance on each task for each layer within the model hierarchy. We found that performance increased with each successive layer of the network, both for categorization tasks as well as identity-orthogonal tasks (Fig. 4.6c and 4.21). This result was in direct accord with the neural results seen in Fig. 3.3a, and moreover, the performance pattern across tasks of the fully-trained networks' top layer is highly consistent with the IT neural performance pattern (Fig. 4.6d-e).

3.3 Discussion

Our results suggest that the same neural circuit mechanisms in the ventral stream (and in particular, IT cortex) build explicit representations for both categorical and non-categorical visual object properties. Though this may sound like a contradiction in terms, it can be interpreted in light of existing theoretical and empirical results that discuss the efficiencies of a joint representation of multiple image properties, especially in relation to avoiding unnecessary binding problems [Edelman, 1999; DiCarlo and Cox, 2007; Sayres and Grill-Spector, 2008]. A key contribution of our experimental results is a systematic, large-scale confirmation that these theories are more consistent with the empirical data than existing alternatives [Mishkin et al., 1983; Goodale and Milner, 1992; Ungerleider and Haxby, 1994]. An additional crucial difference between our results and these existing ideas, however, is that our data and modeling suggest that transformation "sensitivity" is not merely *retained* by successive areas in the ventral stream hierarchy, but rather that is *increased* at each layer in concert with transformation tolerance (see Section 4.2.7 for modeling details).

This observation suggests that, at each stage of the ventral stream, mechanisms that provide a partial solution to any one task (e.g., position estimation) help provide a basis on which to build a more complete solution to each other task (e.g., categorization) at the next layer — and vice-versa. Given the distribution of information across units, however, there is no reason to suspect that the specific tasks we identified here form a privileged basis, in that we likely could obtain a similar result had we measured other similar but not identical tasks.

Additionally, our computational model provides an explanation for a neurally-plausible mechanism that can achieve this type of simultaneous building of explicit representation for both categorical and non-categorical properties, without the need for feedback or attentional mechanisms [Deco and Rolls, 2004; Chikkerur et al., 2010]. However, a potentially deeper observation is that producing this joint tolerance does *not* require directly optimizing for it as such. We find that optimizing just for robust category selectivity brings along performance on all the other non-categorical tasks "for free". This suggests a series of interesting follow-up studies investigating whether the converse is true — is just solving for a non-categorical property (e.g., object position estimation) enough to guarantee categorization performance, or is categorization a much stronger constraint driving the development of IT neural responses?

It may be unintuitive that some properties (e.g., position estimation) that are typically thought of as low-level visual features [Bosking et al., 2002] are actually more effectively captured in higher-level cortical neural populations. Our results are nonetheless consistent with the prior data that formed existing intuitions, because while most studies examining putatively low-level properties do so with very simple stimulus sets (e.g., bars and gratings), our results focus on complex stimuli containing realistic objects on cluttered backgrounds. Our results illustrate that visual concepts present in complex naturalistic stimuli may not map cleanly to, and potentially engage neural mechanisms quite distinct from, those exposed with simpler stimuli (Fig. 3.3).

Existing literature suggests that the ventral stream representation achieves reduction in dimensionality of the original image stimulus by strategically throwing out irrelevant information [Edelman, 1999]. A general hypothesis suggests that this throwing-out is implemented via aggregative operations like the pooling stages of the neural network models discussed above [LeCun and Bengio, 1995]. A more specific version of this same hypothesis is that these pooling operations aggregate over object-identity-preserving transformations at each scale, like higher-level analogs of simple-cell/complex-cell relationship observed in V1 [Riesenhuber and Poggio, 1999b; Serre et al., 2007b]. Our results do not contradict the more general hypothesis, but do show that the more specific version cannot be true. Instead of thinking of ventral-stream dimension reduction as averaging over (or otherwise discounting) the transformations to which the system must become tolerant, we hypothesize an alternative in which what is thrown out is: precisely those lower-level details that do not somehow contribute to *behaviorally useful* visual properties that humans can rapidly assess. This hypothesis could directly by falsified by identifying a visual property defined in complex naturalistic images that humans can report quickly and with high accuracy in the parafoveal visual field, but is not supported by the IT population representation. Equally interesting would be a visual property such that, even if it is supported by IT, does not appear to come along for free with categorization-performance optimization in computational models. Such questions would be especially interesting in the domain of face recognition, where previous data indicates the existence of numerous face-specific processing patches in IT that are differentiated according to their performance on tasks like face identification and orientation estimation [Tsao and Livingstone, 2008].

One limitation of our work is that we make a number of implicit assumptions about the way that IT neurons would be read out (e.g., decoded) by the downstream units directly responsible for behavior. Our results largely involve multi-unit sites, though as in previous work [Yamins* et al., 2014], we sorted single units from our multi-unit recordings [Frey and Dueck, 2007a; Quiroga et al., 2004] and repeated many of the analyses shown above, finding little difference from the multi-unit case. We also do not make use of precise spike times, relying instead on a 100ms rate code on which to build linear classifiers and regressors. These classifiers and regressors are a technical tool for quantifying the amount of information populations have for given tasks. However, because they consist only of linear weightings and at most a single threshold value, they also express a mechanistically simple and plausible model for putative rate-code-based downstream units [DiCarlo et al., 2012]. It would be of interest to determine if more sophisticated codes (e.g., temporal decoding schemes) are involved in the processing of the visual properties we investigate here.

Another limitation of our data is that all images were shown within a parafoveal regime encompassing an 8° window around the animal's gaze fixation point. This regime is large enough to present a wide range of object positions, with maximal displacement greater than object diameter. However, it is not large enough to show, nor do we mean even to suggest, that the ventral stream builds up an ability to estimate properties of objects in the visual periphery normally associated with dorsal stream function [Brown et al., 2005; Sereno and Lehky, 2010]. Instead, given our results as well as recent data showing shape and category selectivity in parietal areas [Rishel et al., 2013; Janssen et al., 2008; Swaminathan and Freedman, 2012], we speculate that both the dorsal and ventral stream contain representations for many types of visual properties, categorical and otherwise — but at different levels of spatial resolution and scale, the ventral being fine-scale a parafoveally biased and the dorsal being large-scale and with peripheral coverage. If borne out, this arrangement would naturally support behavior in which dorsal machinery directed foveation around an environmental saliency map, while the ventral machinery parsed details of each salient foveal snapshot, information which then be integrated downstream to produce an overall scene understanding.

3.4 Methods

3.4.1 Simple stimuli

In addition to our photorealistic image set, we also gathered neural data on a simpler set of stimuli (see Fig. 3.7b), consisting of small grating patches placed on gray backgrounds. The grating objects were shown at different positions in a 5-by-5 location grid. At each location, gratings were shown at each of 4 orientations, including 0°, 45°, 90°, and 135°, for a total of 100 images). The overall intensity of the images are all identical.

3.4.2 Array electrophysiology

We used the same methodology used in Chapter 2 to collect and process neural data. We reproduced key details here for the reader's convenience.

Neural data were collected in the visual cortex of two awake behaving rhesus macaques (*Macaca mulatta*, 7 and 9 kg) using parallel multi-electrode array electrophysiology recording systems (BlackRock Microsystems, Cerebus System). All procedures were done in accordance with NIH guidelines and approved by the MIT Committee on Animal Care guidelines. None 96-electrode arrays (three arrays in each hemisphere, with a total of three hemispheres, two left, one right, across two monkeys) were surgically implanted in anatomically-determined V4, posterior IT, central IT and anterior IT regions [Felleman and Van Essen, 1991]. Of these, 392 neural sites (266 in IT and 126 in V4) were selected as being visually driven with a separate image set. Fixating animals were presented with testing images in pseudorandom order with image duration comparable to those in natural primate fixations [DiCarlo and Maunsell, 2000b]. Images were presented one at a time on an LCD screen (Samsung SyncMaster 2233RZ at 120Hz) for 100ms, occupying a central 8° visual angle radius on top of a gray background, followed by a 100ms gray "blank" period with no image shown. Eye movements were monitored by video tracking (SR Research, EyeLink II), and animals were given a juice reward each time fixation was maintained for 6 successive image presentations. Presentations in which eye movement jitter exceeded ±2° from screen center were discarded. In each experimental block, responses were recorded once for each image, resulting in 25 – 50 repeat recordings of the each testing image. For each image repetition and electrode, scalar firing rates were obtained from spike trains by averaging spike counts in the period 70–170ms post-stimulus presentation, a measure of neural response that has recently been shown to match behavioral performance characteristics very closely [Majaj et al., 2012].

In the same way we processed the neural data in Chapter 2, background firing rate, defined as the mean within-block spike count for blank images, was subtracted from the raw response. Additionally, the signal was normalized such that its per-recording block variance is 1. Final neuron output responses were obtained for each image and site by averaging over image repetitions. Recordings took place daily over a period of several weeks, during which time neuronal selectivity patterns at each recording site were typically stable. Based on firing rates and spike-sorting analysis, we estimate that each individual electrode multi-unit site in this study picks up potentials from 1-3 single neural units.

To determine whether results would likely differ for direct single-unit recordings, we sorted single units from the multi-unit IT data by using affinity propagation [Frey and Dueck, 2007a] together with the method described in [Quiroga et al., 2004]. In our IT sample, we obtained 154 well-isolated single units; in our V4 sample, we obtained 191 well-isolated single units. Throughout, we repeated analyses both for our raw multi-unit site data, as well as for these isolated single-unit populations. Moreover, we have supplemented with serially sampled, single-electrode recording [Hung et al., 2005a; Rust and DiCarlo, 2010], and have found that neuronal populations from arrays have very similar patterns of image encoding as assembled single-electrode unit populations.

Receptive field analysis

Using the simple grating-like stimuli, we were able to compare receptive field locations and sizes in our V4 and IT populations. We found that for both populations, receptive fields were concentrated near the center of gaze. In the case of V4 population, these fields covered the approximately central 4° relative to the center of case; in our IT population, the fields covered roughly central 8°. To investigate the effect of receptive field coverage on our results, we performed versions of each of our analyses restricting to images in the central 4 degrees of the field of view, but did not see substantial differences.

3.4.3 Neural performance assessment

We assessed the performance of neural sites and populations on each of the tasks in our task battery.

For discrete-valued tasks, performance was assessed by training linear classifiers on neural output. Linear classifiers are a standard tool for analyzing the performance capacity of a featural representation of stimulus data on discrete classification problems [Hung et al., 2005a; Pedregosa et al., 2011]. For neuronal sites, the output features are defined as the vector of scalar firing rates for each unit, as is typical in neural decoding studies [Hung et al., 2005a; Rust et al., 2006]. For any fixed population of output features (from either a model or neural population), a linear classifier determines a linear weighting of the units, followed by a discrete threshold, which best predicts classification labels on a sample set of training images. Category or identity predictions are then made for stimuli held out from the weight training set, and accuracy is assessed on these held-out images. For continuous-valued estimation tasks, performance was assessed by training linear regressors on neural output [Pedregosa et al., 2011]. A linear regressor determines a linear weighting of the units that best predicts the target property on a set of training images. Predictions for that property are then made for a set of held-out images, and accuracy is assessed using the Pearson correlation measure discussed above.

For both discrete classifiers and continuous regressors, to reduce the noise in estimating accuracy values, results are averaged over a number of independent cross-validation splittings of the data into training and testing portions. In the data shown in figures 3.3-3.6, results show cross-validated test performance averaged over 50 splits in which each training split contained a randomly selected 80% of the data, and the corresponding testing split contained the remaining 20% of the data. While absolute values of performances depend on the size of training split, the results discussed in this paper do not. In all cases, classifiers and regressors were trained using an $l2$ regularization penalty on the weights, and the penalty weight C was chosen separately for each task via cross-validation sub-splits of the training data [Pedregosa et al., 2011].

3.4.4 Human psychophysical experiments

Data on human object recognition judgement abilities shown in Figs. 3.4 and 3.5 were obtained using Amazon's Mechanical Turk crowdsourcing platform, an online task marketplace where subjects can complete short work assignments for a small payment.

We measured human performance for a subset of the tasks on which we decoded neural performance (see below for detailed list). We recruited MTurk subject pools separately for each task $N = 80$, though there ended up being a small amount of overlap between the subject pools for the various tasks.

For each participant and each task, task sessions consisted of a training phase containing 10 trials (except as indicated below) and a testing phase containing 100 trials. On each trial, a sample image was shown, following by a 500ms pause, and then a response screen was shown. The nature of the response screen depended on the task type (see below for details).

For each of the sessions, we measured 20 of the testing images 2 times, to enable calculation of within-subject reliability. For categorical variables, reliability was calculated as 1 minus the average Hamming distance between two length-20 vectors of distinct repeats, taken over 200 splittings of the repeats into two groups. For continuous variables, reliability was calculated as the average correlation of the target variable between two length-vectors of distinct repeats, again taken over 200 splittings of the repeats into the two group.

During the training trials, sample images were shown for an extended period of time and in which correct answers were indicated both via annotation on the original sample image and in the response screen. Subject correctness in the training phase provided us with an estimate of motor noise for each task. During the 100 testing trials, sample images were shown for 100ms, followed by a 500ms pause, and then a response screen was shown. The accuracy values reported in the figures and text were generated from the testing trials only.

During trial sessions, each trial was assessed on-line for accuracy (that is, how close the subject's answer was to the correct ground-truth answer). The accuracy metric depended on the task type (see below). A small bonus was paid to subject based on their average correctness at the end of the session, and subjects were told at the beginning of each session that their bonus would depend on accuracy.

The tasks we measured included:

- Basic categorization tasks. This is an eight-way alternate forced choice (8-AFC) task. The response screen for this task consistent of 8 response images, one for each of the eight basic categories in our image set. Subjects were required to click with their mouse on the image representing the category they thought they saw in the sample image. The accuracy metric for this task was balanced accuracy. During the training phase, correct answers were indicated by a blue box highlighting the correct choice. Average within-subject reliability for this task was 0.97.

- Subordinate identification tasks. This consisted of eight separate 8-AFC tasks, one for each category. These tasks were not intermixed, e.g., sessions involving subordinate car identification were not intermixed with subordinate boat identification. The response screen for each the eight category tasks consisted of 8 response images, one for each specific object identity within that category. The accuracy metric was balanced accuracy. Average within-subject reliability for this task was 0.92.

- Position estimation. Response screens consisted of a blank canvas the same size as the sample image, and subjects were required to click at the location where they estimated the centroid of the object in the sample image was located. x-position and y-position estimates were computed from the indicated centroid. In the training phase, the location of the centroid was shown with a blue dot, both on the sample image while it was being presented, as well as on the blank response canvas. Accuracy was assessed for each trial as the euclidean distance of the subject's indicated estimate to the actual location. Average within-subject reliability for the x-position estimate was 0.91 and for the y-position estimate was 0.94.

- Axis-aligned bounding box estimation. Response screens consisted of a blank canvas the same size as the sample image, and subjects were required to click on the locations where they thought the top-left and bottom-right of the axis aligned bounding box had been for the object in the sample image. x-axis size, y-axis size, and bounding-box area were computed from the indicated bounding box. During training phase, the correct locations of all four corners of the bounding were indicated using blue dots, and the edges of the bounding box were indicated using black outline, both on the original sample image while it was being shown, as well as in the blank response canvas. Accuracy was assessed for each trial using an area overlap criterion for the estimated versus true bounding box (intersection area divided by total area). Average within-subject reliability of x-axis size was 0.96, for y-axis size was 0.92, and for bounding-box size was 0.84.

- Rotated bounding box estimation. Response screens consisted of a blank canvas the same size as the sample image. Subjects were first required to click on two points indicating one side of the rotated-area bounding box, and then on a third point indicating the extent of the rotated bounding-box in the orthogonal direction. Major axis length, major axis angle, and aspect ratio where computed from the subject's rotated bounding box estimate. Training-phase answers and trial accuracy were as in the axis-aligned bounding box case. Average within-subject reliability for major axis length was 0.85, for major axis angle was 0.79, and for aspect ratio was 0.91.

- Object 3-D scale. Response screens consisted of an image of the object in the sample image, but shown from a single fixed canonical angle (chosen on a per-category basis as described above). On each testing phase trial, the size of the response image was randomized by uniformly drawing from the full size range in the dataset. Subjects were given a slider and were required to resize the image so that the object was at the same 3-dimensional size as they perceived it to be in the sample image. Once subjects felt they had correctly resized the object they pressed a "submit" button. During training phase trials, the correct size was indicated via a marker along the slider, and subjects were required to move the slider to the correct location to within 0.5% size tolerance. Accuracy was assessed using absolute difference between correct and estimated size. Average within-subject reliability for object scale estimate was 0.87.

- Object 3-D rotation. Response screens consisted of a 3-D graphical "pointer" indicating defined "top" and "front" orientations. Subjects were required to rotate the pointer into alignment with the top and front orientations that they perceived in the sample image. Once subjects felt that had corrected posed the pointer, the clicked a "submit" button. During the training phase, the original sample image was repeated on the response screen, and two copies of the pointer were also shown simultaneously: one fixed at the correct 3-D orientation of the object; the other was a movable pointer that subjects were required to rotate into within 2°-solid angle of the correct orientation before proceeding to the next training trial. Training was provided on a per-category basis to teach subjects our definition of the canonical (0, 0, 0) angle for each category, and 32 training examples were provided (containing training images for 4 exemplars each for each of 8 categories). Accuracy was assessed using distance between correct and indicated rotation in the quaternion representation [Shoemake, 1985]. Average within-subject reliability for z-axis rotation was 0.76; for x-axis rotation was 0.69; and for y-axis rotation was 0.71.

3.4.5 Weight pattern analysis

Having determined that the IT population is able to sustain behaviorally plausible linear coding for a variety of tasks, our next goal was to understand the distribution of information for each of the tasks amongst the various sites. To formalize the concept of "relevance of a task at a given site", we used the classifier/regressor weights trained in the population analyses described above (see below for a discussion of alternative metrics). In mathematical terms:

$$\text{site } i \text{ relevance for task } T = w_{Ti}$$

where $\vec{w}_T = (w_{T1}, w_{T2}, \ldots, w_{Tn})$ is the vector of weights of a l_2-regularized linear estimator for task T on site i, and n is the number of neural sites. In the case of the continuous regression tasks, the weights are simply the regression coefficients, whereas in the case of the discrete categorization tasks, the weights are classifier coefficients, prior to the final threshold value. The absolute value of the classifier weight, $|w_{Ti}|$, is a proxy for the amount of information contributed by site i for task T. If $|w_{Ti}|$ is large compared to the weights w_{Tj} for other sites j, site i is taken to be more relevant for the task; $w_T i \gg 0$ corresponds to strong correlation between the site's output at the task, while $w_{Ti} \ll 0$ corresponds to strong anticorrelation.

Let D_T by the distribution of weights for task T (Fig. 3.6a shows example distributions for several selected tasks). In this work, we assume that the weights in $\vec{w}_T$ are IID samples from D_T. We consider the distributions for 108 separate binary tasks, including:

- The 8 1-vs-all basic-level categorization tasks (e.g., Animals vs all, Boats vs all, &c).
- 8 1-vs-all subordinate categorization tasks for each of 8 categories, for a total of 64 binary tasks.
- 12 size, position, bounding box, and pose estimation tasks, as described above.
- 24 subordinate 3-d pose estimation tasks, eight each for the three pose axes, as described above.

In several of the panels in Figure 3.6, we only show results for the non-subordinate tasks, for visual clarity.

We had two basic analysis goals with these distributions: (a) what do the individual task distributions of information look like for each task? and (b) how do they overlap between tasks?

Individual task information distribution

In mathematical terms, our first goal was to characterize the shape of D_T for each task T. To do this, we used two statistical properties of the distributions: skewness and kurtosis.

The γ_1 *skewness* of the weight vector is a measure the balance or asymmetry of the distribution of the weights about the mean weight. Positive skewness means that the positive tail of the weight distribution is longer than the negative tail, e.g., the majority of the weight distribution is below the mean. In the context of this work, high skewness for the weight distribution associated with a given task would indicate that the population was biased towards having sites that are *anticorrelated* with the task, while high negative skewness would indicate the opposite. Formally, skewness is a statistical third-moment measure defined as:

$$\gamma_1(\vec{w}_T) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{w_{Ti}-\mu}{\sigma}\right)^3$$

where $\mu = \frac{1}{n}\sum_{i=1}^{n} w_{Ti}$ is the average weight and $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(w_{Ti}-\mu)^2}$ is the standard deviation of the weights.

We measured the sparseness of weight distributions via *excess kurtosis*, γ_2. Excess kurtosis measures how spread out the weights are, relative to a normal distribution.

Positive excess kurtosis means that the distribution is more peaked than a gaussian distribution with the same mean and standard deviation. High kurtosis values indicating that only a very few sites are highly informative for the task, and low values indicating little differentiation between sites. Formally, excess kurtosis is defined as

$$\gamma_2(\vec{w}_T) = \frac{\frac{1}{n}\sum_{i=1}^{n}(w_{Ti} - \mu)^4}{\sigma^4} - 3.$$

To ensure that we accurately took into account the effects of noise and sparse sampling of image space, the skewness and sparseness shown are computed by averaging the skewness and sparseness computed separately for the weights of 50 classifiers/regressors, each trained on a different split containing 50% of the training data. We also resampled sites with replacement, to ensure we were properly accounting for uncertainty due to site sampling. Error bars shown in Fig. 3.6b are standard deviations computed over both site samples and image splits.

To help interpret the meaning of these skewness and sparseness values, we compared them to two types of controls:

1. Gaussian control. With a statistically large enough sample, gaussian distributions have 0 skew and 0 excess kurtosis. However, finite samples of a gaussian distribution will not have 0 skewness or kurtosis. We matched the size of the empirical distribution of IT sites ($N = 266$) and drew 1000 samples of size from a standard gaussian, and computed the skewness and kurtosis for each sample. The gray bars in Fig. 3.6b show the standard deviation spread of these values.

2. Three-point distribution control. The other end of the statistical spectrum from the gaussian control are *three-point distributions*, distributions that have support on three distinct points, $x_- < x_0 < x_+$. For each task T, we ap-

proximated the empirical distribution D_T with a three-point distribution by solving for x_-, x_0, x_+ as well as probabilities $0 < p_-, p_0, p_+ = 1 - p_- - p_0$, such that x_0 is the empirical mean of D_T and the three-point distribution had the same mean, standard deviation, skew and kurtosis as D_T. Conceptually, the interpretation of these approximations are to divide the population of sites for each task T into three subpopulations: the x_--sites that are the "highly-anticorrelated" with the task T, the x_+-sites that are highly correlated with the task, and the x_0-sites that are not highly relevant to the task. The reference values shown in the skewness histogram (Fig. 3.6b, top right) are, by definition, $(p_+ + 0.5 \cdot p_0)/(p_- + 0.5 \cdot p_0)$, measuring the ratio of above-mean to below-mean sites. The reference numbers shown in the sparsity histogram (Fig. 3.6b, bottom right) are, by definition $p_- + p_+$, measuring the proportion of "high-relevance" sites.

As shown in Fig. 3.6b-c and discussed in the text, we found that the distributions of weights are:

- On average, comparatively symmetric, in which most tasks are statistically indistinguishable in their skewness from size-matched gaussian control, and the proportion of above-mean to below-mean sites range from 0.7 to 1.3.
- On average, slightly more sparse than normally distributed, in which the proportion of high-relevance sites (as defined above) ranges from 15% to 35% of the total, with a median of 26.3%. The normal distribution has 32.5% high-relevance sites, and a significant proportion of tasks are not statistically distinguishable in their sparsity from that of the size-matched gaussian control.

Taken together, these results suggest a picture of information distribution across sites that is comparatively well distributed, as opposed to each task being supported by a small number highly-dedicated sites.

Using the $l2$ metric (as shown in the main text), we also estimated the skewness and kurtosis measures for the population of single units that we sorted from our raw multi-unit site data. Though the specific skewness and kurtosis values were different, the overall summary results were quite similar. Specifically, the kurtosis values corresponded to a fraction of highly-weighted sites between 5.2% and 4.6% of all sites, with a mean of 32.1%, with 53 of 108 tasks having sparseness indistinguishable for that of the standard normal distribution. The skewness values indicated that, for all tasks, the number of sites weighted at above the mean is no less than 75% and no more than 152% of the fraction of weights below the mean, with 73 out of 108 tasks having skewness not statistically distinguishable from that of the standard normal distribution.

Task-pair information overlap

Having characterized the per-task distributions, we sought to characterize the *overlap* of weights for each task pair, seeking to understand how the sites that are likely to be useful for any one task are related to those that are relevant for each other task. We defined the overlap between tasks i and j as the pearson correlation between the absolute values of the weight vectors for the two tasks (see below for discussion of alternative metrics). Formally, the overlap matrix (see figure 3.6c) has i, j-th element defined as

$$M_{i,j} = corr(|\vec{w}_{T_i}|, |\vec{w}_{T_j}|) = \frac{cov(|\vec{w}_{T_i}|, |\vec{w}_{T_h}|)}{\sqrt{var(|\vec{w}_{T_i}|) \cdot var(|\vec{w}_{T_h}|)}}$$

where T_i and T_j are the i-th and j-th tasks, respectively. This value ranges between 1 (perfectly correlated, meaning complete overlap) and -1 (perfectly anticorrelated, meaning totally non-overlapping). In practice, given that the number of tasks is comparable to the number of sites in our sample, and that (as seen in the previous section), each task utilizes between 15% and 35% of all sites, the minimum possible average overlap will be significantly larger than -1.

In figure 3.6d, we show the average of the correlations of 1000 random draws of weights of classifiers/regressors over a set of 50 splits containing 50% of the training data. That is, each matrix element is the average of 1000 correlations $corr(w_{T_i}^{s_k}, w_{T_j}^{s_l})$ where $w_{T_i}^{s_k}$ is the weight vector for the i-th task, trained on the k-th (of 50) splits, and where s_k an s_l where chosen randomly for each of the 1000 repeats.

We were particularly interested in quantifying the overlap between category-detection tasks and non-categorical tasks. To provide reference points against which to compare our results, we considered two controls:

1. Random overlap model. Weights are randomly assigned to each task subject to the constraint of matching per-task and per-site marginal weight distributions, but in which task pair overlap is unconstrained.

2. Minimum overlap model. Weight assignments are constrained as in the random overlap model but additional constrained to result in as little overlap as possible.

In both cases, we used gradient-based optimization methods to solve for weights $\eta_{Ti}, 0 \leq i < n$ for each task T, such that

- $\sum_{T=0}^{N} \eta_{Ti}^2 = \sum_{T=0}^{N} w_{Ti}^2$ for each unit i, where N is the number of tasks.

- $mean(\vec{\eta}_T) = mean(w\eta_T)$ for all tasks T

- $variance(\vec{\eta}_T) = variance(w\eta_T)$ for all tasks T

- $skewness(\vec{\eta}_T) = skewness(w\eta_T)$ for all tasks T

- $kurtosis(\vec{\eta}_T) = kurtosis(w\eta_T)$ for all tasks T.

Using the l-bfgs algorithm [Pedregosa et al., 2011], we minimized the square difference objective function summed over the above 5 terms. In the case of the minimum overlap model also simultaneously minimized $\sum_{T_i<T_j} corr(|\vec{\eta}_{T_i}|, |\vec{\eta}_{T_j}|)$. For both the random and minimum overlap models, we ran the optimization over 1000 random initializations of the η values.

In summary, and shown in Figure 3.10 we found that:

- Overlap is generally positive.

- The average overlap of (non-face) categorical tasks with each other is higher than would be predicted by the random overlap model, except for the case of faces.

- The average overlap of the face-detection task with other categorical tasks is lower than would be predicted by random overlap, but higher than would be predicted by the minimal overlap model.

- The average overlap of (non-face) categorical tasks with non-categorical tasks is lower but not statistically different from the prediction of the random model.

- The average overlap of face detection with non-categorical tasks is not statistically distinguishable from that predicted by the minimal overlap model.

Taken together, these results suggest that, holding faces aside, the IT neural population jointly encodes both categorical and non-categorical visual tasks using an integrated representation in which many units participate in tasks. However, our this observations are consistent with well-established observations of segregated face-specific sites [Kanwisher et al., 1997; Tsao and Livingstone, 2008], and provides a positive control that the overlap-measurement methodology used here is able resolve module-like structure when it exists.

Alternative measures of per-site relevance and overlap

To ensure that our results were not biased by the type of regularization we used, we computed sparsity, skewness, and overlap for measures of site-relevance as defined by two additional types of weight-generation procedures:

1. l_1 **estimator weights:** The same as , but using a different classifier regularization scheme in which l_2 regularization is replaced with l_1 sparse regularization.

2. **Single-site performances:** The weight-signed value of the single site's task balanced accuracy or correlation value (depending on whether the task is categorical or continuous). That is,

$$\text{site } i \text{ relevance for task } T = \begin{cases} sign(w(i,T)) \cdot AccBal(i,T) & \text{if } T \text{ is categorical} \\ Corr(i,T) & \text{if } T \text{ is continuous} \end{cases}$$

   where $w(i,T)$ is the single weight associated with the one-feature classifier for site i with task T, and $AccBal(i,T)$ (resp. $Corr(i,T)$) is the balanced accuracy (resp. correlation) of site i for task T.

These metrics express different physiological hypotheses about the rules that create weightings of downstream neurons that read off the IT population to actually carry out specific tasks. Each metric is based on an algorithmic procedure for learning linear weight patterns that would be useful for reading off a given task. In the future, we believe that experiments that could distinguish between these hypotheses would be of significant utility, because they could shed light on the learning rules used in cortical areas. For the purposes of this work, however, we ended up carrying out analyses for all metrics to check that our results were robust to the choice of metric. We mostly focus on the analysis for the case of the l_2 estimator weight metric, but it is largely identical in all cases.

We also considered several alternative measures of overlap, including:

1. **Percentile matrix:** Defining the overlap between task i and j as the average percentile in the distribution D_{T_i} of all sites at or above some percentile threshold in D_{T_j}.

2. **Knock-down analysis:** Using the same classifiers/regressors trained in the population analyses described above, we could knock down each site by setting that site's input to baseline values for all stimuli, and then assess the change in performance for each task. In this analysis, the amount of change in performance with the site removed, relative to the expected performance drop from the removal of a randomly chosen single site, would serve as the proxy for the amount of information carried by that site for the task.

Again, results for these alternative overlap measures were not qualitatively different than for the correlation metric we show in figures 3.6c-e.

Acknowledgments

This work was partially supported by National Science Foundation grant IS 0964269 and US National Eye Institute R01-EY014970. Ha Hong was partially supported by the Samsung Scholarship.

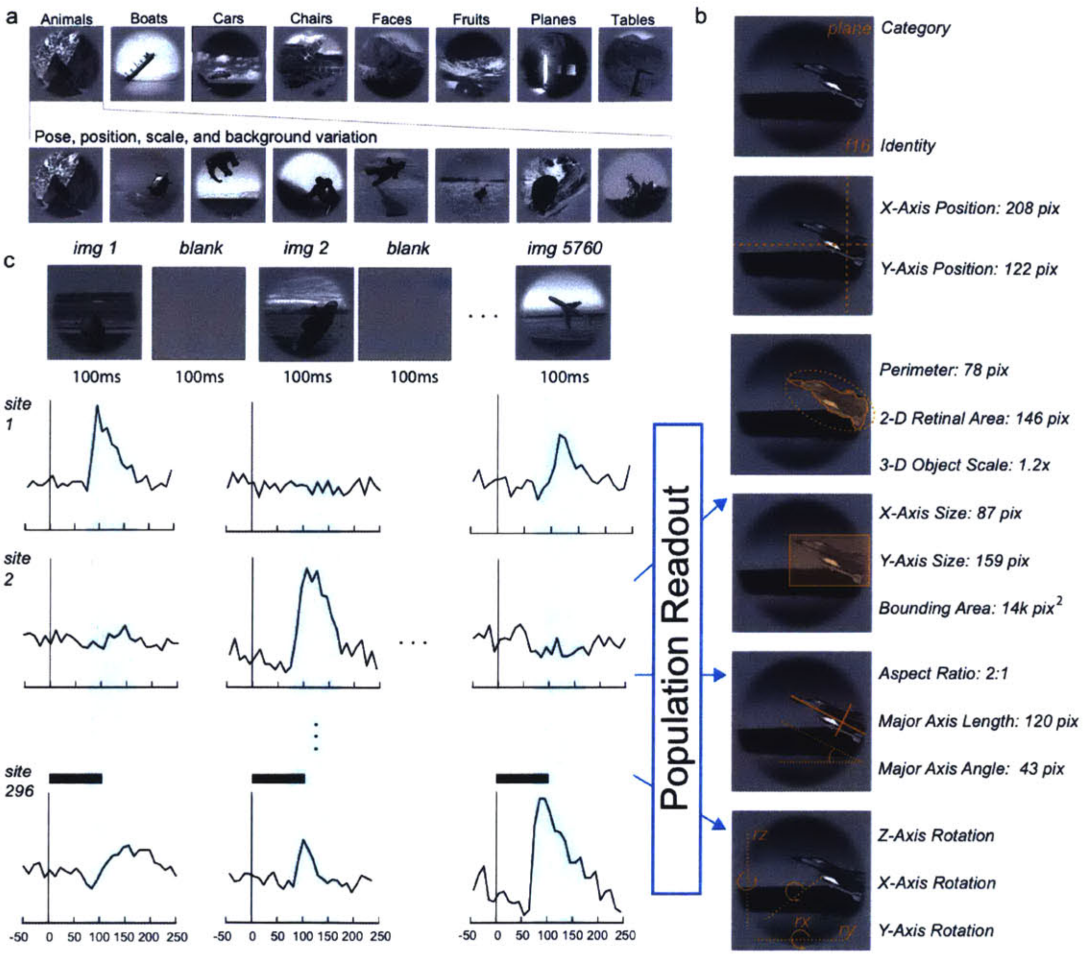


Figure 3.1: **Large-scale electrophysiological measurement of neural responses in macaque IT and V4 cortex to visual object stimuli containing high levels of object viewpoint variation.** (***a***) We recorded neural responses to 5,760 high-variation naturalistic images consisting of eight exemplar objects each in eight categories (Animals, Boats, Cars, Chairs, Faces, Fruits, Planes, Tables), placed on natural scene backgrounds, at a wide range of positions, sizes, and poses. Stimuli were presented to awake fixating animals for 100ms in a Rapid Serial Visual Presentation (RSVP) paradigm. Object centers varied within 8° of fixation center. (***b***) Recordings were made using chronically-implanted electrode arrays, collecting a total of 392 neuronal sites in inferior temporal (IT, $n = 266$) and V4 ($n = 126$) visual cortex. Each stimulus was repeated between 25 and 50 times. Spike counts were binned in the time window 70ms-170ms post stimulus presentation and averaged across repetitions, to produce a 5760 x 392 neural response pattern array. (***c***) We then used linear readouts to decode a variety of types of image information from the neural responses, including categorical data such as object category and exemplar identity, as well as continuous data such as, for example, object position, retinal and 3-D object size, 2-D and 3-D pose angles, object perimeter and aspect ratio.

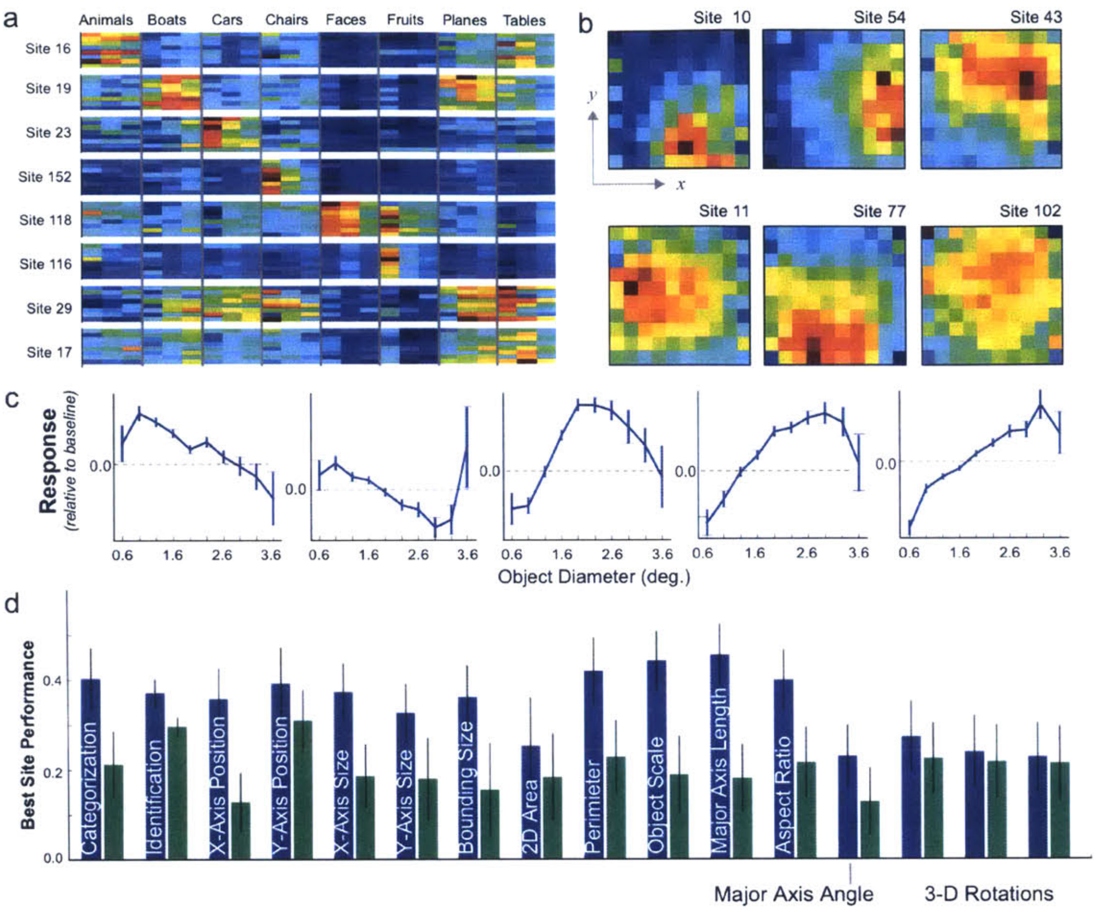


Figure 3.2: **Categorical and non-categorical object property information encoded in single-site responses.** (***a***) Category selectivity heatmaps of the single sites in our IT sample that are best at decoding each of the eight categories present in our stimulus set. Each colored bar represents the response of the indicated site relative to that sites's baseline (blue=low, red=high). The colored bars represent responses averaged over images of each of the eight object exemplars in the indicated category (vertical axis), further broken down into three increasing levels of image parameter variation (horizontal axis, see text). (***b***) Position selectivity response heatmaps for best single sites for object position estimation. Each colored squared position in each heat map represents the average of the indicate site's activity over all images where the objects center is located in that square's position. (***c***) Size selectivity response profiles for the best single units for object size estimation. The x-axis represents the object diameter in degrees as seen by the animal. The y-axis represents response relative to baseline of the indicated unit, averaged over all images whose size falls in the indicated diameter bin. Error bars are standard deviations due to image variation. (***d***) Performance of single best sites from IT (blue bars) and V4 (green bars) on each task. Error bars are over subsets of units to chose best single unit, and over images used to compute performance.

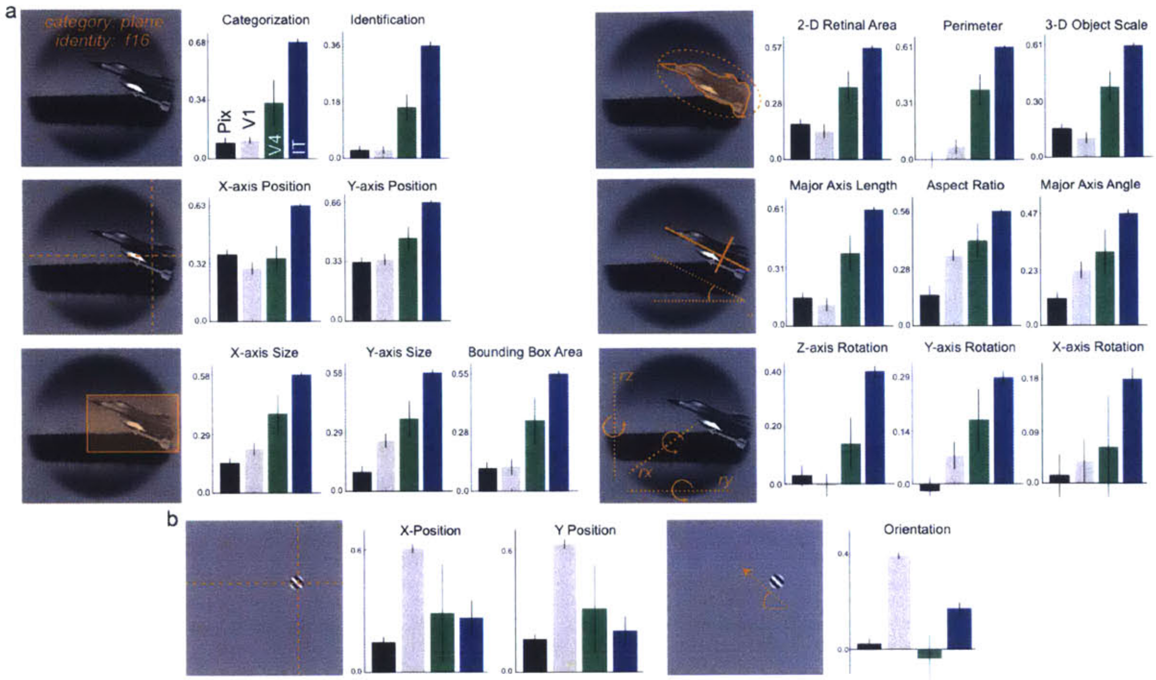
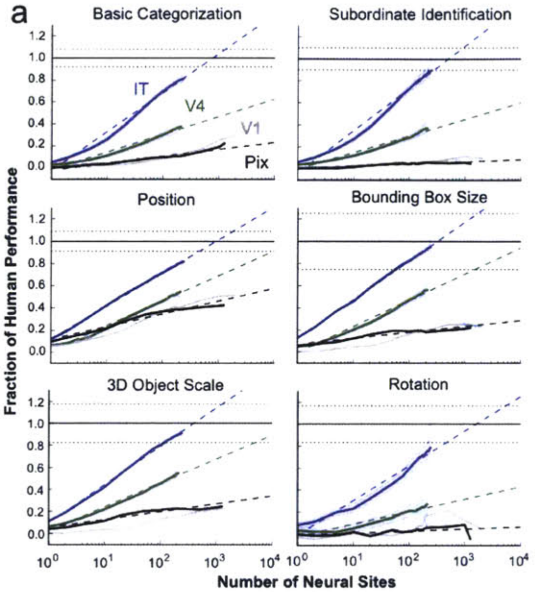


Figure 3.3: **Neural population decoding of a spectrum of categorical and non-categorical properties.** (***a***) For each task, we trained a linear decoder on neural output. For discrete-valued categorization tasks, including object category and exemplar subordinate identity, we used Support Vector Machine (SVM) classifiers with $L2$-loss and $L2$-regularization. For continuous-valued estimation tasks, we used linear regression with $L2$ (Ridge) regularization. We compared decoding performance for our recorded IT population sample (blue bars) and V4 population sample (green bars), as well as for a performance-optimized V1 gabor wavelet model (gray bars) and the trivial pixel control (black bars). For categorical properties, bar height represents balanced accuracy (0 = chance, 1 = perfect). For continuous properties, bar height represents the Pearson correlation between the predicted value and the actual ground-truth value. All values are shown on cross-validated testing images held out during classifier and regressor training. Error bars represent standard deviation over cross-validation image splits. All evaluations are performed with $n = 126$ units, and a fixed number of training/testing examples (see text and Methods for details). (***b***) Population decoding results for position and orientation tasks defined on a simpler stimulus set consisting of grating patches placed on gray backgrounds. y-axis, bar colors, and error bars are as in panel (a).



b

	IT	V4	V1	Pix
Basic Categorization	773 ± 185	2.2×10^6	—	—
Subordinate Identification	496 ± 93	4.4×10^6	—	—
X-axis Position	1414 ± 403	5.2×10^5	3.0×10^7	—
Y-axis Position	918 ± 309	2.5×10^4	8.7×10^6	—
Bounding Box Size	322 ± 90	1.7×10^4	—	—
X-axis Size	256 ± 87	9.8×10^3	3.4×10^7	—
Y-axis Size	237 ± 87	3.8×10^3	9.5×10^6	—
3-D Object Scale	401 ± 90	3.2×10^4	—	—
Major Axis Length	201 ± 70	1.1×10^4	—	—
Aspect Ratio	163 ± 61	951 ± 59	6.5×10^3	—
Major Axis Angle	804 ± 136	3.2×10^6	—	—
Z-axis Rotation	1932 ± 1061	—	—	—
Y-axis Rotation	369 ± 115	2.8×10^5	—	—
X-axis Rotation	1570 ± 530	—	—	—

— = more than 10 billion sites required

Figure 3.4: **Comparison of neural population decoding performance to human psychophysical measurements.** (***a***) Human-relative performance as a function of number of subsampled sites used to decode the property, for selected tasks. The x-axis represents the base-10 logarithm of the number of sites. For each task, the y-axis represents the performance of the decoder with the indicated number of sites, as a fraction of median human performance for that task. A value of 1 would mean that the neural decoder achieve 100% of human performance level. As in Fig. 3, balanced accuracy was used for both neural decoders and humans for the categorical properties, while estimate/actual correlation was used for continuous-valued properties. Solid lines represent measured data; dotted lines represent log-linear extrapolations based on the measured data. We evaluated our measured IT (blue lines) and V4 (green lines) neural populations out to the data limited 266 and 126 sites respectively, and evaluated V1 model (gray lines) and pixels (black lines) out to 2000 units. Human performance for each indicated task was measured using large-scale web-based psychophysics (see text and Methods). The variation in human performance between individuals in our psychophysical studies is indicated by the dotted horizontal lines flanking $y = 1$ (the median human performance level). (***b***) Estimated number of neural sites that would be required to match median human performance. Error bounds are due to variation in site subsamples. Value is shown as "—" when more than 10^{10} sites would be required.

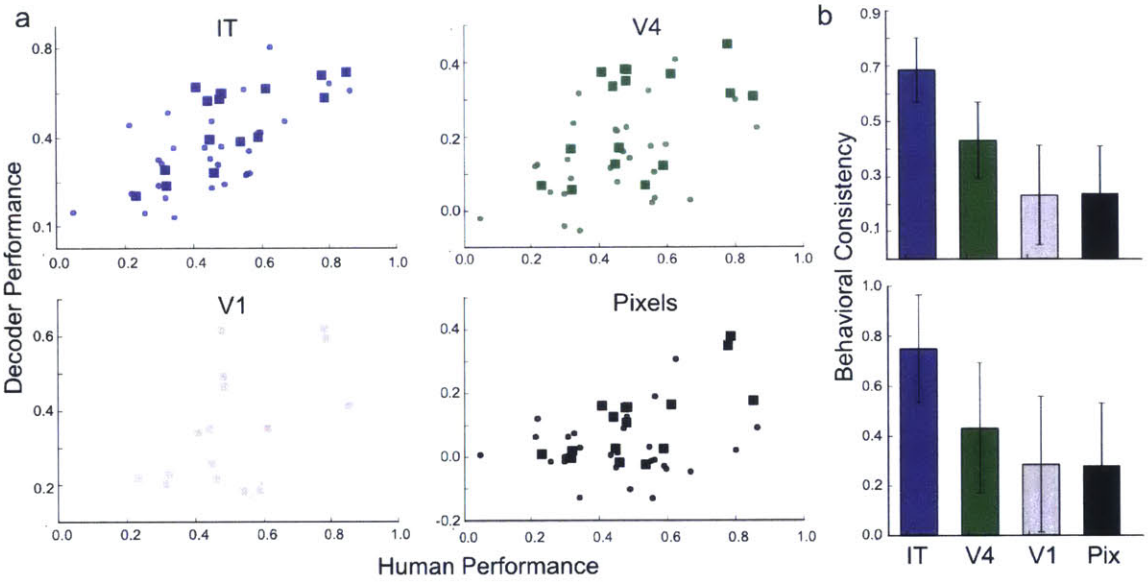


Figure 3.5: **Consistency of neural population decoding with human performance pattern.** (*a*) Scatters show human performance (x-axis) versus neural performance (y-axis) for a variety of tasks. Large squares show the aggregated tasks (n = 14) indicated in Fig. 4b. Small circles (n = 30) indicate values for further breakdown of the data into subordinate identification and pose estimation tasks on a per-category basis (see text and Methods for details). (*b*) Summary of data from panel a. Bar height represents Spearman's R correlation between human and neural decode for the 14 aggregated tasks (top panel) and 44 disaggregated tasks (bottom panel). Error bars are standard deviations due to be task and image variation (see Methods for details).

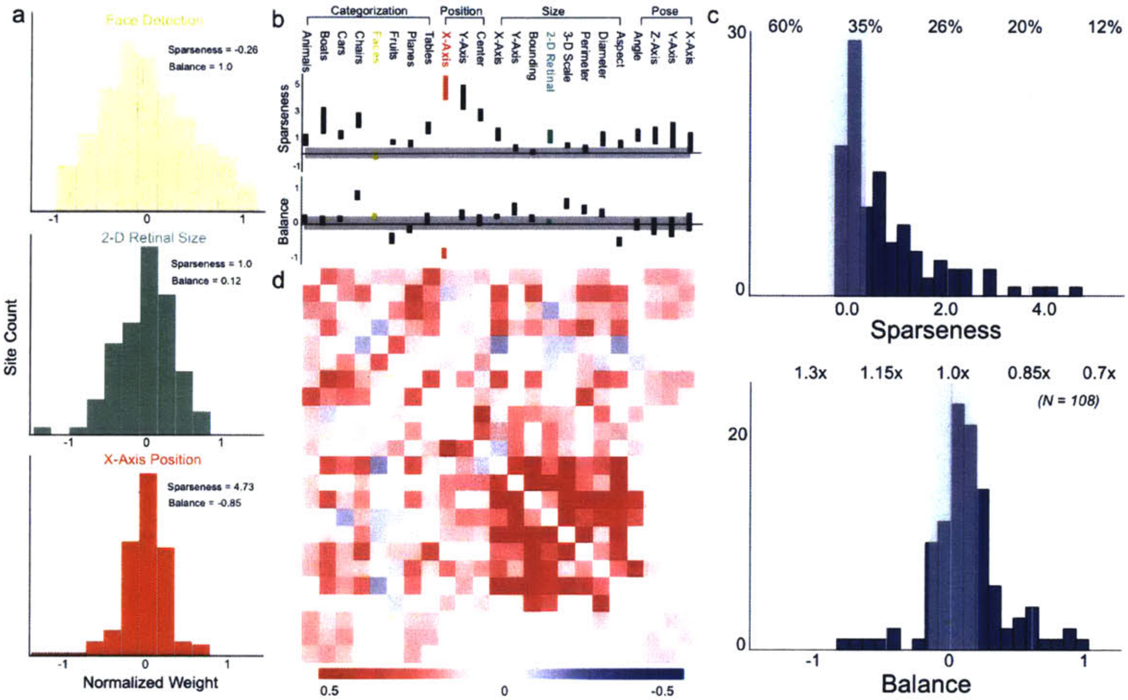


Figure 3.6: **Distribution and overlap of site contribution across tasks.** (***a***) Distributions of decoder weights across neural sites for several tasks. x-axis represents normalized decoder weight, with y-axis is site bincount. (***b***) Sparseness (top panel) and balance (bottom panel) of weight distributions for selected tasks. Sparseness is measured via excess kurtosis (γ_2, see Methods for details), while balance is measured via skewness (γ_1). Error bars are standard deviations over image splits on which weights were determined. Gray band represents 1 standard deviation of distribution of values taken on size-matched samples from a gaussian distribution. (***c***) Histograms of values of sparseness (top panel) and balance (bottom panel) over all 108 tasks. Reference values on sparseness panel show fractions of "high-relevance" sites in 3-point distributions (see text and Methods). Reference values at top of balance panel show fractions of values above vs. below means, ranging from 1.3 to 0.7. (***d***) Quantification of weight pattern overlap for pairs of tasks. Each colored square in the heatmap is the Pearson correlation between the absolute value of the weight vectors for a pair of tasks. A high value (red color) indicates that the weight pattern for the pair of tasks is similar; a low value (blue color) indicates the opposite. White indicates a value that is not statistically significantly different from zero. The order of tasks is the same as in panel b.

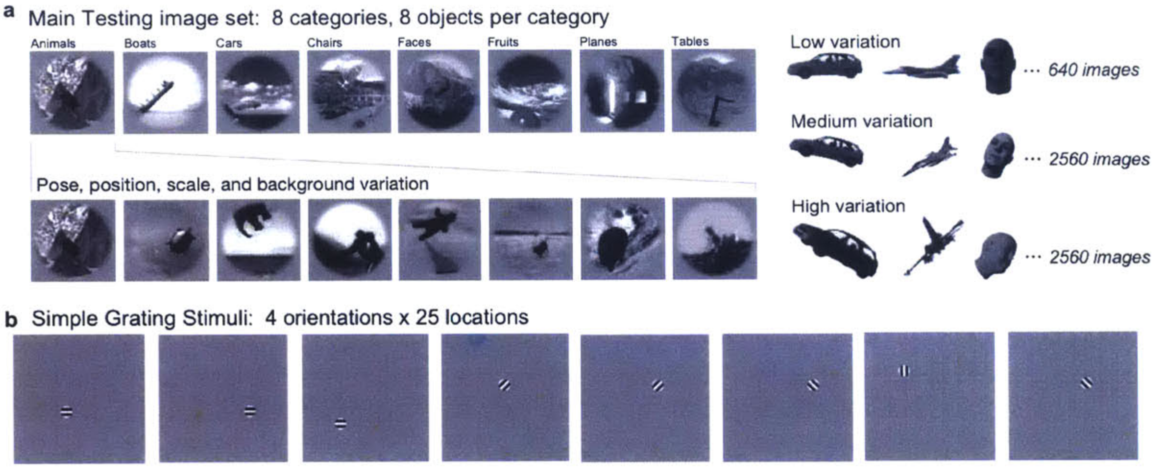

a Main Testing image set: 8 categories, 8 objects per category
Animals
Boats
Cars
Chairs
Faces
Fruits
Planes
Tables
Pose, position, scale, and background variation
Low variation
··· 640 images
Medium variation
··· 2560 images
High variation
··· 2560 images
b Simple Grating Stimuli: 4 orientations x 25 locations

Figure 3.7: **Image sets.** (***a***) High-variation testing image set on which we collected neural data and evaluated models contained 5760 images of 64 objects in 8 categories. The image set contained three subsets, with low, medium and high levels of object view variation. Images were placed on realistic background scenes, which were chosen randomly to be uncorrelated with object category identity. As discussed in the Methods, this dataset supported a wide range of categorical and non-categorical tasks, on which we evaluated population performance of V4 and IT neural populations, as well as computational models. (***b***) Simple grating set stimuli used to estimate V4 and IT receptive fields. This stimulus set supported three simple tasks, including x and y position estimation of the center of grating object, as well as grating orientation.

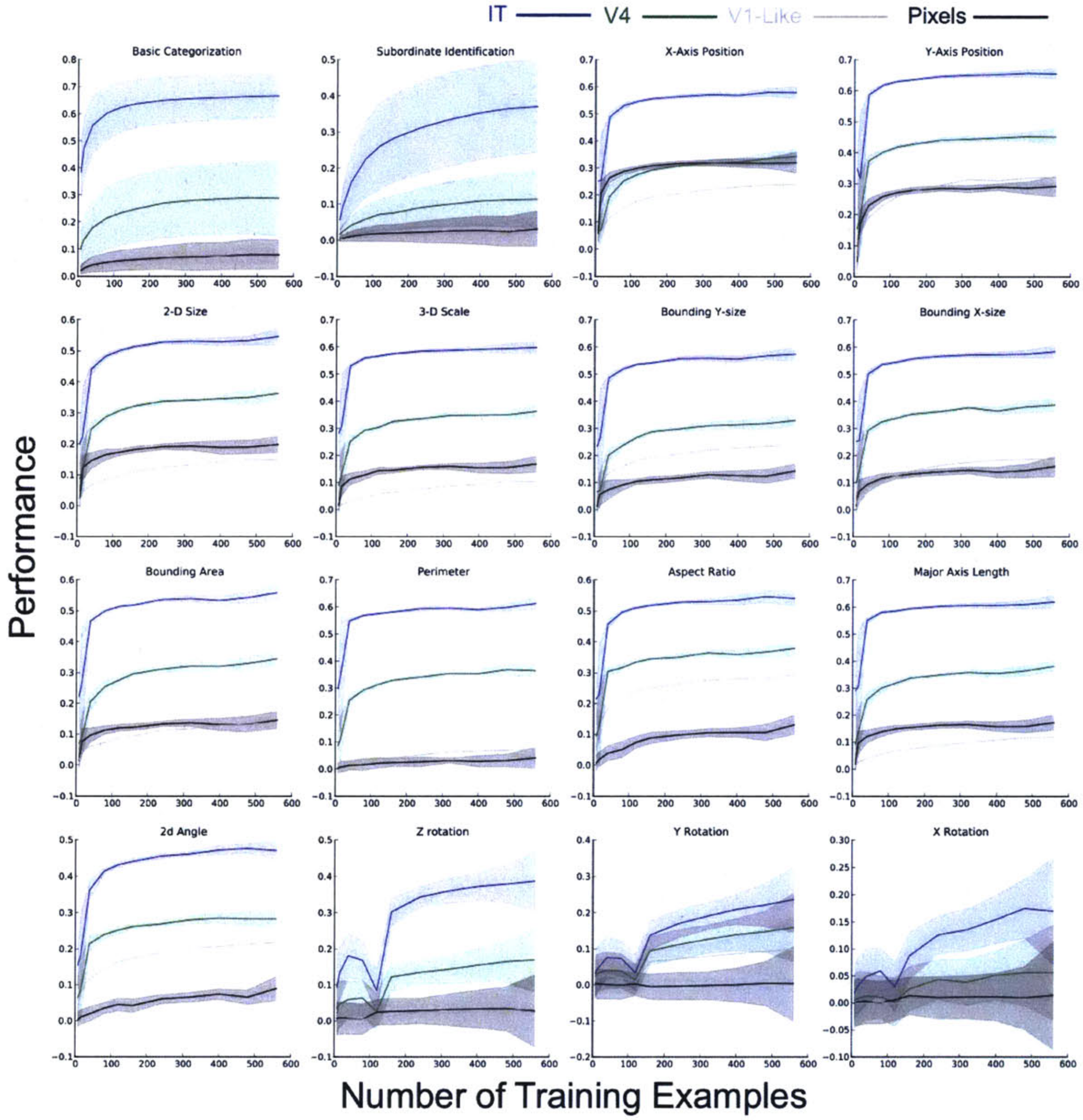


Figure 3.8: **Performance training curves.** For each task, performance of population decodes as a function of number of training examples used to train linear classifiers and regressors. Error bars are over samples of units and image splits. Blue lines are for IT neural population, green lines are V4 neural population, gray is V1-like model population, and black is pixel control.

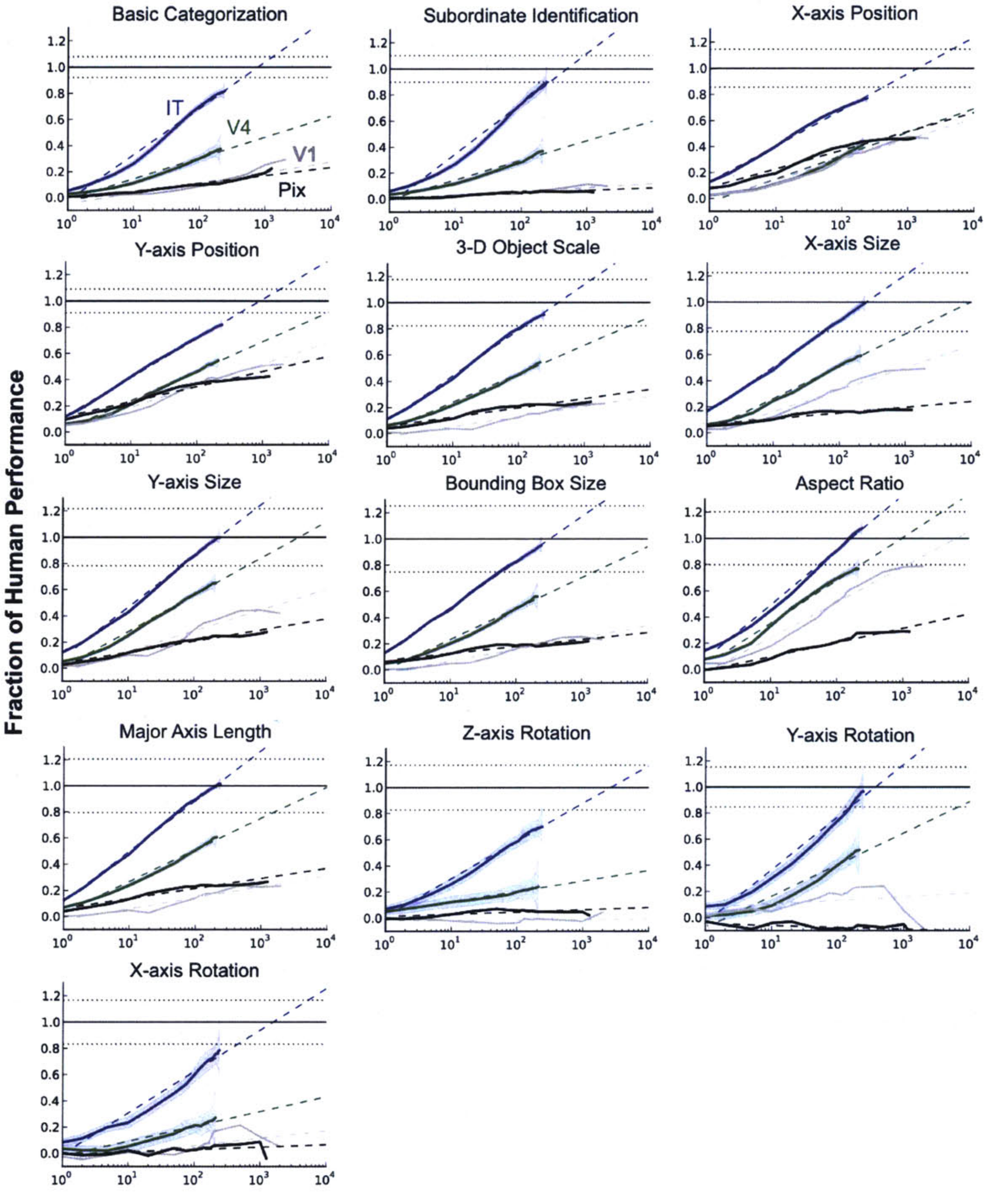


Figure 3.9: **Performance extrapolation.** This figure shows extrapolations for all measure tasks. The x and y axes are as in Figure 3.4a.

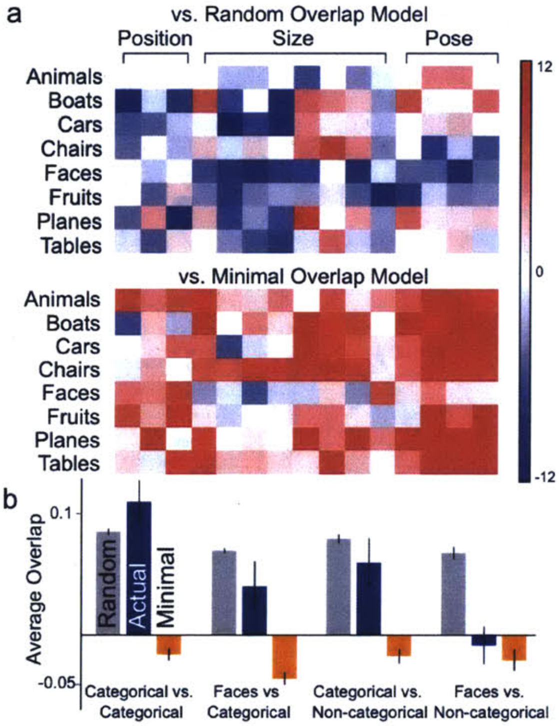


Figure 3.10: **Comparison of task overlap to random and minimal control models.** (***a***) Comparison of weight overlap for categorical vs non-categorical tasks, relative to Random overlap (top) and minimal overlap (bottom) models. (***b***) Average overlap for (1) (non-face) categorical tasks, (2) faces-vs-non-face categorical tasks, (3) non-face categorical tasks vs non-categorical tasks and (4) faces vs non-categorical tasks. Shown are actual neural overlap (blue bars) in comparison to random overlap (gray) and minimal overlap (orange) models. Errorbars for neural data overlap are due to variation in unit sampling and classifier training split. Errorbars in model overlaps are due to variation of model input data (per-task and per-unit weight constrains) due to unit sampling and classifier training split, as well as random initial conditions of model weights.

Chapter 4

Performance-Optimized Hierarchical Models Predict Neural Responses in Higher Visual Cortex*

The ventral visual stream underlies key human visual object recognition abilities. However, neural encoding in the higher areas of the ventral stream remains poorly understood. Here, we describe a modeling approach that yields a quantitatively accurate model of Inferior Temporal (IT) cortex, the highest ventral cortical area. Using high-throughput computational techniques we discovered that, within a class of biologically-plausible hierarchical neural network models, there is a strong correlation between a model's categorization performance and its ability to predict individual IT neural unit response data. To pursue this idea, we then identified a high-performing neural network that matches human performance on a range of recognition tasks.

This chapter is modified from a study published as [Yamins, Hong*, Cadieu, Solomon, Seibert, and DiCarlo, 2014]. This also includes unpublished preliminary results of a recent work with Dan Yamins, Najib Majaj, and James DiCarlo.

Critically, even though we did not constrain this model to match neural data, its top output layer turns out to be highly predictive of IT spiking responses to complex naturalistic images at both the single site and population levels. Moreover, the model's intermediate layers are highly predictive of neural responses in V4 cortex, a mid-level visual area that provides the dominant cortical input to IT. These results show that performance optimization — applied in a biologically appropriate model class — can be used to build quantitative predictive models of neural processing.

4.1 Introduction

Retinal images of real-world objects vary drastically due to changes in object pose, size, position, lighting, non-rigid deformation, occlusion, and many other sources of noise and variation. Humans effortlessly recognize objects rapidly and accurately in spite of this enormous variation, an impressive computational feat [DiCarlo and Cox, 2007]. This ability is supported by a set of interconnected brain areas collectively called the ventral visual stream [Grill-Spector et al., 2001; Malach et al., 2002], with homologous areas in non-human primates [Kriegeskorte et al., 2008b]. The ventral stream is thought to function as a series of hierarchical processing stages [Tanaka, 1996; Logothetis and Sheinberg, 1996; Gross, 1994] that encode image content (e.g. object identity and category) increasingly explicitly in successive cortical areas [Vogels and Orban, 1994; DiCarlo and Cox, 2007; DiCarlo et al., 2012]. For example, neurons in the lowest area, V1, are well-described by Gabor-like edge detectors that extract rough object outlines [Carandini et al., 2005a], though the V1 population does not show robust tolerance to complex image transformations [DiCarlo et al., 2012]. Conversely, rapidly-evoked population activity in top-level inferior temporal (IT) cortex can directly support real-time, invariant object categorization over a wide range of tasks [Hung et al., 2005a; Rust and DiCarlo, 2010]. Mid-level ventral areas — such as V4, the dominant cortical input to IT — exhibit intermediate levels of object selectivity and variation tolerance [Rust and DiCarlo, 2010; Freiwald and Tsao, 2010; Connor et al., 2007].

Significant progress has been made in understanding lower ventral areas such as V1, where conceptually compelling and quantitatively accurate models have been discovered [Carandini et al., 2005a]. These models have the ability to predict the response magnitudes of an individual neuronal unit to novel image stimuli based on its responses to a fixed number of sample images. Higher ventral cortical areas, especially V4 and IT, have been much more difficult to understand. While first-principles-based models of higher ventral cortex have been proposed [Fukushima, 1980; Riesenhuber and Poggio, 2000; Serre et al., 2007a; Lecun et al., 2004; Bengio, 2009; Pinto et al., 2009], these models fail to match important features of the higher ventral visual neural representation in both humans and macaques [Kriegeskorte et al., 2008b; Kiani et al., 2007]. Moreover, attempts to directly fit V4 and IT neural tuning curves on general image stimuli have shown only limited predictive success [Rust et al., 2006; Gallant et al., 1996]. Explaining the neural encoding in these higher ventral areas thus remains a fundamental open question in systems neuroscience.

As with models of V1, any effective model of higher ventral areas should be neurally predictive. But because the higher ventral stream is also believed to underlie sophisticated behavioral object recognition capacities, models must also match IT on *performance* metrics: the ability to equal (or exceed) the decoding capacity of IT neurons on object recognition tasks. A model with perfect neural predictivity in IT will necessarily exhibit high levels of performance, because IT itself does. Here we demonstrate that the converse is also true, within a biologically appropriate model class. Combining high throughput computational and electrophysiology techniques, we explore a wide range of biologically plausible hierarchical neural network models and then assess them against measured IT and V4 neural response data. We show that there is a strong correlation between a model's performance on a challenging high-variation object recognition task and its ability to predict individual IT neural unit responses.

Extending this idea, we used optimization methods to identify a high-performing hierarchical neural network model that matches human performance on a range of recognition tasks. We then show that the top output model layer is highly predictive of neural responses in IT cortex, providing a first quantitatively accurate model of this highest ventral cortex area — even though the model was never explicitly constrained to match neural data. Moreover, analysis of the intermediate layers of the model show that they are highly predictive of V4 neural responses, confirming the importance of performance as a biologically-meaningful cortical constraint.

4.2 Results

4.2.1 Invariant object recognition performance strongly correlates with IT neural predictivity

We first measured IT neural responses on a benchmark testing image set that has been shown to expose key performance characteristics of visual representations [Cadieu et al., 2013]. This image set consists of 5760 images of photorealistic three-dimensional objects drawn from eight natural categories (Animals, Boats, Cars, Chairs, Faces, Fruits, Planes and Tables). The image set contains high levels of the object position, scale and pose variation that make recognition difficult for artificial vision systems, but to which humans are robustly tolerant [DiCarlo and Cox, 2007; Pinto et al., 2008a]. The objects are placed on cluttered natural scene backgrounds that are randomly selected to ensure that background content is uncorrelated with category identity (Fig. 4.7a).

Using multiple electrode arrays, we collected responses from 168 IT neurons to each image (see Methods). We then used high throughput computational methods to evaluate thousands of candidate neural network models on these same images, measuring object categorization performance as well as IT neural predictivity for each model (see Fig. 4.1a; each point represents a distinct model). To measure categorization performance, we trained standard Support Vector Machine (SVM) linear classifiers on model output layer units [Hung et al., 2005a], and computed cross-validated testing accuracy for these trained classifiers. To assess models' neural predictivity, we used a standard linear regression methodology [Carandini et al., 2005a; Cadieu et al., 2007; Sharpee et al., 2012]: for each target IT neural site, we identified a "synthetic neuron" composed of a linear weighting of model outputs that would best match that site on fixed sample images, and then tested response predictions against actual neural site's output on novel images. See Methods for additional details on performance and predictivity metrics.

In our initial high-throughput experiments, models were drawn from a large parameter space of Convolutional Neural Networks (CNNs) expressing an inclusive version of the hierarchical processing concept [Lecun et al., 2004; Serre et al., 2007a; Mutch and Lowe, 2008; Pinto et al., 2009]. CNNs approximate the general retinotopic organization of the ventral stream via spatial convolution, with computations in any one region of the visual field identical to those elsewhere. Each convolutional layer is composed of simple and neuronally plausible basic operations, including linear filtering, thresholding, pooling and normalization (Fig. 4.8a). These layers are stacked hierarchically to construct deep neural networks.

Each model is specified by a set of 57 parameters controlling the number of layers and parameters at each layer, (e.g.) fan-in and fan-out, activation thresholds, pooling exponents, and local receptive field sizes at each level of the network. Network depth ranged from one to three layers, and filter weights for each layer were chosen randomly according to a bounded uniform distribution whose bounds were model parameters (see Methods). These models are consistent with the Hierarchical Linear-Nonlinear hypothesis that higher level neurons (e.g. IT) output a linear weighting of inputs from intermediate-level (e.g. V4) neurons followed by some simple additional nonlinearities [Connor et al., 2007; Brincat and Connor, 2004].

Models were then selected for evaluation by one of three procedures: (1) random draws from the uniform distribution over model parameter space (Fig. 4.1a, $n =$ 2016, green dots); (2) optimization for performance on the high-variation eight-way categorization task ($n = 2043$, blue dots); and (3) optimization directly for IT neural predictivity ($n = 1876$, orange dots). (See Methods and Fig. 4.9 for more details on these optimizations.) In each of these experiments, we observed a large variation in both performance and IT-predictivity across the range of selected parameters. This result demonstrates that, while the HLN hypothesis is consistent with a broad spectrum of particular neural network architectures, choices for these architectural parameters have a large effect on a specific model's ability either to perform object recognition effectively or match neural data.

Performance was significantly correlated with neural predictivity in all three selection regimes. Models that performed better on the categorization task were also more likely to produce outputs more closely aligned to IT neural responses. While the class of HLN-consistent architectures contains many neurally inconsistent architectures with low IT-predictivity, performance provides a meaningful way to *a priori* rule out many of those inconsistent models. No individual model parameters correlated nearly as strongly with IT-predictivity as performance (Fig. 4.10), indicating that the performance/IT-predictivity correlation cannot be explained by simpler mechanistic considerations (e.g. receptive field size of the top layer).

Critically, directed optimization for performance significantly *increased* the correlation with IT-predictivity compared to the random selection regime ($r = 0.78$ vs $r = 0.55$), even though neural data were not used in the optimization. Moreover, when optimizing for performance, the best-performing models predicted neural output as well as those models directly selected for neural predictivity, though the reverse is not true. Together, these results imply that, while the IT-predictivity metric is a complex function of the model parameter landscape, performance optimization is an efficient means to identify regions in parameter space containing IT-like models.

4.2.2 IT cortex as a neural performance target

Fig. 4.1a suggests a next step toward improved encoding models of higher ventral cortex: drive models further to the right along the x-axis — if the correlation holds, the models will also climb on the y-axis. Ideally, this would involve identifying hierarchical neural networks that perform at or near human object recognition performance levels, and validating them using rigorous tests against neural data (Fig. 4.2a). However, the difficulty of meeting the performance challenge itself can be seen in Fig. 4.2b. To obtain neural reference points on categorization performance, we trained linear classifiers on the IT neural population (Fig. 4.2b, green bars) and the V4 neural population ($n = 128$, hatched green bars). To expose a key axis of recognition difficulty, we computed performance results at three levels of object view variation, from low (fixed orientation, size and position) to high (180-deg rotations on all axes, 2.5x dilation and full-frame translations; see Fig. 4.7a). As a behavioral reference point, we also measured human performance on these tasks using web-based crowdsourcing methods (black bars). A crucial observation is that at all levels of variation, the IT population tracks human performance levels, consistent with known results about IT's high category decoding abilities [Hung et al., 2005a; Rust and DiCarlo, 2010]. The V4 population matches IT and human performance at low levels of variation but performance drops quickly at higher variation levels. (This V4-to-IT performance gap remains nearly as large even for images with no object translation variation, showing that the performance gap is not due just to IT's larger receptive fields.)

As a computational reference, we used the same procedure to evaluate a variety of published ventral stream models targeting several levels of the ventral hierarchy. To control for low-level confounds, we tested the (trivial) pixel model, as well as SIFT, a simple baseline computer vision model [Lowe, 2004]. We also evaluated a V1-like Gabor-based model [Pinto et al., 2008a], a V2-like conjunction-of-Gabors model [Freeman and Simoncelli, 2011], and HMAX [Serre et al., 2007a; Mutch and Lowe, 2008], a model targeted at explaining higher ventral cortex and that has receptive field sizes similar to those observed in IT. The HMAX model can be trained in a domain-specific fashion, and to give it the best chance of success, we performed this training using the benchmark images themselves (see Methods for more information on the comparison models). Like V4, the control models that we tested approach IT and human performance levels in the low variation condition, but in the high-variation condition all of them fail to match the performance of IT units by a large margin. It is not surprising that V1 and V2 models are not nearly as effective as IT, but it is instructive to note that the task is sufficiently difficult that the HMAX model performs less well than the V4 population sample, even when pre-trained directly on the test dataset.

4.2.3 Constructing a high-performing model

While simple three-layer hierarchical convolutional neural networks can be effective at low-variation object recognition tasks, recent work has shown that they may be limited in their performance capacity for higher-variation tasks [DiCarlo et al., 2012]. For this reason, we also allowed our model class to contain combinations (e.g., mixtures) of CNN networks (Fig. 4.8b), which correspond intuitively to architecturally specialized subregions like those observed in the ventral visual stream [Downing et al., 2006; Freiwald and Tsao, 2010]. To address the significant computational challenge of finding especially high-performing architectures within this large space of possible networks, we employed Hierarchical Modular Optimization (HMO). The HMO procedure embodies a conceptually simple hypothesis for how high-performing combinations of functionally specialized hierarchical architectures can be efficiently discovered and hierarchically combined, without needing to prespecify the subtasks ahead of time. Algorithmically, HMO is analogous to an adaptive boosting procedure [Schapire, 1999] interleaved with hyperparameter optimization (see Methods and Fig. 4.8c for more information on the HMO procedure).

We applied the HMO selection procedure on a challenging object recognition screening task, analogous to the pre-training performed for the HMAX model (Fig. 4.7b). This screening set was designed so that its relationship to the benchmark testing images is similar to that between any two typical samples of natural images: having some high-level commonalities but otherwise quite different specific content. Like the testing set, the screening set contained images of objects placed on randomly selected backgrounds, but used entirely different objects in totally non-overlapping semantic categories, with none of the same backgrounds and widely divergent lighting conditions and noise levels. Applying the HMO procedure to this screening set identified a high-performing four-layer model with 1250 top-level outputs (Fig. 4.8B and Fig. 4.11), which we will refer to as the "HMO model".[1]

Using the same classifier training protocol as with the neural data and control models, we then tested the HMO model to determine whether its performance transferred from the screening to the testing image set. In fact, the HMO model matched the object recognition performance of the IT neural sample (Fig. 4.2b, red bars), even when faced with large amounts of variation — a hallmark of human object recognition ability [DiCarlo and Cox, 2007]. These performance results are robust to the number of training examples and number of sampled model "neurons", across a variety of distinct recognition tasks (Figs. 4.12 and 4.13).

4.2.4 Predicting neural responses in individual IT neural sites

Given that the HMO model had plausible performance characteristics, we then measured its IT-predictivity, both for the top-level output, as well as for each of the model's three intermediate layers (Fig. 4.3, red lines/bars). We found that each successive layer predicted IT units increasingly well, demonstrating that the trend identified in Fig. 4.1a continues to hold in higher performance regimes (see Fig. 4.1b). Qualitatively examining the specific predictions for individual images, the model layers show that category selectivity and tolerance to more drastic image transformations emerges gradually along the hierarchy (Fig. 4.3a, top four rows). At lower layers, units predict IT responses to objects only at a limited range of poses and positions. At higher layers, variation tolerance grows while category selectivity develops, suggesting that as more explicit "untangled" object recognition features are

[1] We also performed a pre-training of the HMAX model using the screening set used to learn the HMO model, and then re-extracted the learned HMAX on the testing set. We found that this only further decreased final performance and neural fit results of the HMAX model, e.g. the learned parameters did not effectively generalize from the screening to the testing set.

generated at each processing stage, its representations become increasingly IT-like [DiCarlo et al., 2012].

Critically, we found that the top layer of the high-performing HMO model achieves high predictivity for individual IT neural sites, predicting 48.5 ± 1.3% of the explainable IT neuronal variance (Fig. 4.3b and c). This represents a nearly 100% improvement over the best comparison models and is comparable to the prediction accuracy of state-of-the-art models of lower-level ventral areas such as V1 on complex stimuli [Carandini et al., 2005a]. In comparison, while the HMAX model was better at predicting IT responses than baseline V1 or SIFT, it was not significantly different from the V2-like model. Though the high-performing HMO model is deeper and more complex that then three-layer CNNs investigated earlier in Fig. 4.1a, the direct relationship between model categorization performance and IT-predictivity for hierarchical network models nonetheless extends across the entire range of performance levels and model complexities (Fig. 4.1b).

To control for how much of the model's prediction capacity could be expected to be reproduced by *any* algorithm with high categorization performance, we also assessed semantic ideal observers [Geisler, 2003], including a hypothetical "model" which has perfect access to all category labels and other image parameters. The ideal observers do predict IT units above chance level (Fig. 4.3c, left two bars), which is consistent with the hypothesis that IT neurons are partially categorical in their responses. However, the ideal observers are significantly *less* predictive than the HMO model, showing that high IT-predictivity does not automatically follow from category selectivity, and that there is significant non-categorical structure in IT responses that is attributable to intrinsic aspects of the hierarchical network structure (see, e.g., Fig. 4.3a, last row). In sum, our results suggest that high categorization performance and the hierarchical model architecture class work in concert to produce

IT-like populations, and neither one of these constraints is sufficient on its own to do so.

4.2.5 Population representation similarity

Characterizing the IT neural representation at the population level may be equally important for understanding object visual representation as individual IT neural sites. The Representation Dissimilarity Matrix (RDM) is a convenient tool comparing two representations on a common stimulus set in a task-independent manner [Kriegeskorte et al., 2008b; Pasupathy and Connor, 2002]. Each entry in the RDM corresponds to one stimulus pair, with high/low values indicating that the population as a whole treats the pair stimuli as very different/similar. Taken over the whole stimulus set, the RDM characterizes the layout of the images in the high-dimensional neural population space. When images are ordered by category, the RDM for the measured IT neural population (Fig. 4.4a) exhibits clear block-diagonal structure — associated with IT's exceptionally high categorization performance — as well as off-diagonal structure that characterizes the IT neural representation more finely than any single performance metric (Fig. 4.4a and Fig. 4.14). We found that the neural population predicted by the output layer of the HMO model had very high similarity to the actual IT population structure, close to the split-half noise ceiling of the IT population (Fig. 4.4b). This implies that much of the residual variance unexplained at the single-site level may not be relevant in the IT population level code. Just as with individual unit neural predictivity, the HMAX model is approximately as effective as a V2-like model.

We also performed two stronger tests of generalization: (1) object-level generalization, in which the regressor training set contained images of only 32 object exemplars (4 in each of 8 categories), with RDMs assessed only on the remaining 32 objects, averaging results across many such object splits, and (2) category-level generalization, in which the regressor sample set contained images of only half the categories (8 objects in each of (e.g.) animal, boat, car, and chair categories), but assessed only on images of the other categories (8 objects in face, fruit, plant and table categories), averaged across many such category splits (see Fig. 4.15 and 4.14). We found that the prediction generalizes robustly, capturing the IT population's layout for images of completely novel objects and categories (Fig. 4.4b-c and Fig. 4.14).

4.2.6 Predicting responses in V4 from intermediate model layers

Cortical area V4 is the dominant cortical input to IT, and the neural representation in V4 is known to be significantly less categorical than that of IT [Rust and DiCarlo, 2010]. Comparing a performance-optimized model to these data would provide a strong test both of its ability to predict the "internal" structure of the ventral stream as well as to go beyond the direct consequences of category selectivity. We thus measured the HMO model's neural predictivity for the V4 neural population (Fig. 4.5). We found that the HMO model's penultimate layer is highly predictive of V4 neural responses ($51.7 \pm 2.3\%$ explained V4 variance), providing a significantly better match to V4 than either the model's top or bottom layers. These results are strong evidence for the hypothesis that V4 corresponds to an intermediate layer in a hierarchical model whose top layer is an effective model of IT. Of the control models that we tested, the V2-like model predicts the most V4 variation ($34.1 \pm 2.4\%$). Unlike the case of IT, ideal observer semantic models explain effectively no response variance in V4, consistent with V4's lack of category selectivity. Together these results suggest that performance optimization not only drives top-level output model layers to resemble IT, it also imposes strong biologically consistent constraints on the intermediate feature representations that can support performance at higher levels.

4.2.7 Emergence of non-categorical visual properties

Visual perception involves estimating a variety of other properties besides an object's category or identity [Edelman, 1999; Koenderink and van Doorn, 1979]. Here, we tested whether *categorization performance optimized* networks can also solve *non-categorical* visual tasks in Chapter 3. In order to answer the question, we further optimized a model for category recognition performance on a very large database of natural images containing approximately 1 million images in 1000 every-day object categories [Deng et al., 2009] with filter value tuning [LeCun and Bengio, 1995; Krizhevsky et al., 2012], stopping the optimization when recognition performance reached saturation. To ensure a sufficiently strong test of generalization could be performed, from the beginning of model training we removed categories from the training set that overlapped with those that appeared in the testing image set used in the neural and behavioral experiments discussed above. For an evenly-spaced series of timepoints during model training, we then extracted model unit responses from each layer on the test image set. This procedure for training and testing is somewhat analogous to performing a time-course of neural samples on a developing animal, in the context of our simplified *in silico* visual model. For more details, see Section 4.4.4.

Consistent with results from previous sections, even though no neural data were used to learn the model parameters, and even though the semantic content of the training images was quite different from that of the testing images, the final "adult" model state was nonetheless highly predictive of neural responses in the test images on an image-by-image basis. Specifically, the models' top layer was predictive of detailed neural response patterns in IT cortex, its intermediate layers were predictive of neural response patterns in V4 cortex, and its lowest layers evidence V1-like Gabor edge tuning (Fig. 4.18b; Section 4.4.4).

We then investigated the model's performance on the tasks in Chapter 3 on which we had measured neural population performance. We found that test categorization performance increased throughout the course of training (Fig. 4.19, grayed panels), indicating effective generalization, since because the model was trained on a completely different image set containing non-overlapping categories of objects. Unexpectedly, however, we also found that performance on all non-categorical tasks also increased during training, and that performance on non-categorical tasks was highly correlated with categorization performance across training timepoints (Fig. 4.6a-b, Fig. 4.19 white panels, and 4.20). This may be somewhat surprising, because not only was the model not supervised explicitly for these category-orthogonal estimation tasks, the categorization task for which it was supervised explicitly sought to become invariant to these very same tasks.

We also investigated performance on each task for each layer within the model hierarchy. We found that performance increased with each successive layer of the network, both for categorization tasks as well as identity-orthogonal tasks (Fig. 4.6c and 4.21). This observation may also be somewhat unexpected, both since the higher layers of the model were more directly exposed to the nuisance-variable invariance training criterion, and since the lower model layers are significantly more retinotopic that higher layers. Nonetheless, this result was in direct accord with the neural results seen in Fig. 3.3a. Moreover, the performance pattern across tasks of the fully-trained networks' top layer is highly consistent with the IT neural performance pattern (Fig. 4.6d–e).

Together, these results indicate that this computational model is a plausible description of a hierarchical computational mechanism by which the ventral stream could simultaneously represent both categorical and non-categorical image properties. Moreover, the fact that non-categorical information for a wide spectrum of tasks emerged in the model without being explicitly built in suggest that our observations of the same properties in the neural data (described mainly in Chapter 3) is likely to be a non-accidental feature of how the ventral stream builds high-level object representations.

4.3 Discussion

In this work, we demonstrate a principled method for achieving greatly improved predictive models of neural responses in higher ventral cortex. Our approach embodies a working hypothesis for two concrete biological constraints that shaped visual cortex: (1) the *functional* constraint of recognition performance, and (2) the *architectural* constraints imposed by the convolutional neural network hierarchy.

4.3.1 A generative basis for higher visual cortical areas

Our modeling approach has common ground with existing work on neural response prediction [Sharpee et al., 2012], e.g. the Hierarchical Linear-Nonlinear hypothesis. But in a departure from that line of work, we do not tune model parameters (the non-linearities or the model filters) separately for each neural unit to be predicted. In fact, with the exception of the final linear weighting, we do not tune parameters using neural data at all. Instead, the parameters of our model were independently selected to optimize functional performance at the top level, and these choices create fixed bases from which any individual IT or V4 unit can be composed. This yields a *generative model* that allows the sampling of an arbitrary number of neurally consistent units. As a result, the size of the model does not scale with the number of neural sites to be predicted — and since the prediction results were assessed for a random sample of IT and V4 units, they are likely to generalize with similar levels of predictivity to any new sites that are measured.

4.3.2 What features do good models share?

What characteristics contribute to making certain neural networks (e.g. the HMO-trained model) so much better than others at object recognition performance or predicting neural data? While highly IT-predictive models often had certain characteristics in common (e.g. more hierarchical layers), many poor models also shared these same characteristics, so no one architectural parameter strongly correlated with neural predictivity (see Fig. 4.10). To gain further insight, we performed an initial exploratory analysis of the parameters of the learned HMO model, evaluating each parameter both for how sensitively it was tuned and how diversely it was tuned between model mixture components. We found two classes of model parameters that were both especially sensitive and diverse (Methods and Figs. 4.16, 4.17): (1) filter statistics, including filter mean and spread, and (2) the exponent trading off between max-pooling and average-pooling [Riesenhuber and Poggio, 2000]. These observations hint at a computationally rigorous explanation for what underlies heterogeneity that is observed in the receptive fields of ventral stream neurons both at the unit and sub-area levels [Martin and Schroder, 2013; Downing et al., 2006; Freiwald and Tsao, 2010], but much work remains to be done to confirm such a hypothesis.

4.3.3 A "top-down" approach to understanding cortical circuits

A common assumption in visual neuroscience is that understanding the qualitative structure of tuning curves in lower cortical areas will be a necessary precursor to explaining higher visual cortex. For example, significant work has gone into assessing the extent to which V4 neurons can be understood as a curvature-selective shape representation [Sharpee et al., 2012]. Our results indicate that it is useful to complement this bottom-up approach with a top-down perspective in which behavioral (e.g. performance) metrics are a sharp and computationally tractable constraint shaping individual neural response functions in both higher and intermediate cortical areas. In other words, our "explanation for IT" is that it was selected by an evolutionary and/or developmental process precisely so that it had useful performance characteristics for tasks like those used in our optimization. Similarly, our "explanation of V4" is that it was selected precisely so that it could serve as an effective input for the downstream computation in IT. This type of explanation is qualitatively different from more traditional approaches that seek explicit descriptions of neural responses and brain regions in terms of (e.g.) particular geometrical primitives. However, our results show functionally-relevant constraints can be used to obtain quantitatively predictive models even when such explicit bottom-up primitives have not been identified.

Going forward, we will work to bridge the gap between these bottom-up and top-down explanations, by analyzing model features to build more intuitively interpretative links to lower and intermediate visual cortex, especially in V1 and V2. The results here also suggest that it will be important to explore recent high-performing computer vision systems with architectures inspired by the ventral stream, e.g. [Krizhevsky et al., 2012], to determine whether these specific learning algorithms provide further insight into ventral stream mechanisms. Our results show that behaviorally-driven computational approaches have an important role in understanding the details of cortical processing [Marr et al., 2010]. We anticipate that further work along these lines will uncover more detailed predictions about the underlying constraints that shaped the ventral stream, and speculate that the overall approach may be applicable to other cortical areas and task domains.

4.4 Methods

4.4.1 Experimental data collection

We used the same experimental data collected in Chapter 2. Here, we reproduced the details that have been employed for the reader's convenience.

Array electrophysiology

Neural data were collected in the visual cortex of two awake behaving rhesus macaques (*Macaca mulatta*, 7 and 9 kg) using parallel multi-electrode array electrophysiology recording systems (BlackRock Microsystems, Cerebus System). All procedures were done in accordance with National Institute of Health guidelines and approved by the Massachusetts Institute of Technology Committee on Animal Care guidelines. Six 96-electrode arrays (three arrays each in two monkeys) were surgically implanted in anatomically-determined V4, posterior IT, central IT and anterior IT regions [Felleman and Van Essen, 1991]. Of these, 296 neural sites (168 in IT and 128 in V4) were selected as being visually driven with a separate imageset. Fixating animals were presented with testing images in pseudo-random order with image duration comparable to those in natural primate fixations [DiCarlo and Maunsell, 2000b]. Images were presented one at a time on an LCD screen (Samsung, SyncMaster 2233RZ at 120Hz) for 100ms, occupying a central 8° visual angle radius on top of a gray background, followed by a 100ms gray "blank" period with no image shown. Eye movements were monitored by a video tracking system (SR Research, EyeLink II), and animals were given a juice reward each time central fixation was maintained for 6 successive image presentations. Eye movement jitter within ±2° from a 0.25° red dot at the center of screen was deemed acceptable, while presentations with large eye movements were discarded. In each experimental block, responses were recorded once for each image, resulting in $25 - 50$ repeat recordings of the each testing image.

For each image repetition and electrode, scalar firing rates were obtained from spike trains by averaging spike counts in the period $70 - 170$ms post-stimulus presentation, a measure of neural response that has recently been shown to match behavioral performance characteristics very closely [Majaj et al., 2012]. Background firing rate, defined as the mean within-block spike count for blank images, was subtracted from the raw response. Additionally, the signal was normalized such that its per-block variance is 1. Final neuron output responses were obtained for each image and site by averaging over image repetitions. Recordings took place daily over a period of several weeks, during which time neuronal selectivity patterns at each recording site were typically stable. Based on firing rates and spike-sorting analysis, we estimate that each individual electrode multi-unit site in this study picks up potentials from 1-3 single neural units. To determine whether results would likely differ for direct single-unit recordings, we sorted single units from the multi-unit IT data by using affinity propagation [Frey and Dueck, 2007a] together with the method described in [Quiroga et al., 2004]. Of these units, 21 had internal trial-to-trial consistency with an r-value of 0.3. We assessed the HMO model's prediction ability for these single units, obtaining a median of $50.4 \pm 2.2\%$ explained variance, very close to that obtained directly from the multi-unit data. Moreover, we have supplemented with serially sampled, single-electrode recording [Hung et al., 2005a; Rust and DiCarlo, 2010], and have found that neuronal populations from arrays have very similar patterns of image encoding as assembled single-electrode unit populations.

Test stimulus set

The test stimulus set (Fig. 4.7, a) consisted of 5760 images of 64 distinct objects chosen from one of eight categories (animals, boats, cars, chairs, faces, fruits, planes, tables), with eight specific exemplars of each category (e.g., BMW, Z3, Ford, etc. within the car category). The set was designed specifically to (1) include a range of everyday objects, (2) support both coarse, "basic-level" category comparisons (e.g. "animals" vs. "cars") and finer subordinate level distinctions (e.g. distinguish among specific cars) [Rosch et al., 1976], and (3) require strong tolerance to object viewpoint variation, e.g. pose, position and size. Objects were placed on realistic background images which were chosen randomly so as to prevent correlation between background content and object class identity.

Object view parameters were chosen randomly from uniform ranges at three levels of variation (low, medium, and high), and images were rendered using the photorealistic Povray package [Plachetka, 1998]. The parameter ranges for the three variation levels were:

- **Low variation:** All objects placed at image center ($x = 0, y = 0$), with a constant scale factor ($s = 1$) translating to objects occluding 40% of image on longest axis, and held at a fixed reference pose ($rxy = rxz = ryz = 0$).

- **Medium variation:** Object position varies within one-half multiple of total object size ($|x|, |y| \leq 0.3$), varying in scale between $s = 1/1.3 \sim .77$ and $s = 1.3$, and between -45 and 45 degrees of in-plane and out-of-plane rotation ($\leq 45°$).

- **High variation:** Object position varies within one whole multiple of object size ($|x|, |y| \leq 0.6$), varying in scale between $s = 1/1.6 \sim .625$ and $s = 1.6$, and between -90 and 90 degrees of in-plane and out-of-plane rotation ($\leq 90°$).

Crowd-sourced human psychophysics

Data on human object recognition judgement abilities shown in Figs. 4.2b and 4.12 were obtained using Amazon Mechanical Turk crowd-sourcing platform, an online task marketplace where subjects can complete short work assignments for a small payment. A total of 104 observers participated in one of three visual task sets: an 8-way classification of images of eight different cars, an 8-way classification of images of eight different faces, or an 8-way categorization of images of objects from eight different "basic-level" categories. Observers completed these 30 to 45 minute tasks through Amazon Mechanical Turk. All the results were confirmed in the lab setting with controlled viewing conditions, and virtually identical results were obtained in the lab and web populations (Pearson correlation = 0.94 ± 0.01). For the 8-way basic-level categorization task set, each human observer ($n = 29$) judged a subset of 400 randomly sampled images with blocks for each of the three variation levels (400 out of 640 for low variation and 400 out of 2560 for medium and high variation levels). For the 8-way car ($n = 39$) and 8-way face ($n = 40$) identification task sets, each observer saw all 80 images at the low variation level and all 320 images at both medium and high variation levels. The presentation of images were randomized and counterbalanced so that the number of presentations of each class was the same in the given variation level. Each trial started with a central fixation point that lasted for 500 ms after which an image appeared at the center of the screen for 100 ms, following a 300 ms delay, the observer was prompted to click one of 8 response images that matched the identity or category of the stimulus image. Response images were shown from a fixed frontal viewpoint and remained constant throughout a trial block. All human studies were done in accordance with the MIT Committee on the Use of Humans as Experimental Subjects.

Performance was determined by computing accuracies for each task. For a given 8-way task set and variation level (e.g., high-variation basic-level categorization, medium-variation car subordinate identification, etc.), we constructed the raw 8×8 confusion matrix for each individual observer, and computed the population confusion matrix summing raw confusion matrices across individuals. From the population confusion matrix, we computed accuracy values for each task of recognizing one target class against seven distractor classes (a.k.a. "binary" task). We obtained 72 binary task accuracies by performing this procedure over all combinations of three task sets and three variation levels (3 task sets × 8 targets per task set × 3 levels of variation). We used standard signal detection theory to compute population accuracy from the population confusion matrix definition. The pooled performance scores were highly consistent, with a median (taken over the 72 tasks) Spearman-Brown corrected split-half Pearson-coefficient self-consistency of 0.99. To estimate the subject to subject variability we selected one subject from each task set and combined the task performance of the three task sets to produce 72 "individual" human accuracies.

4.4.2 Neural predictivity and performance metrics

For each IT neural site, we used linear regression to identify a linear weighting of model output units (from the top or intermediate layers) that is most predictive of that neural site's actual output on a fixed set of sample images. Using this "synthetic neuron", we then produced per-image response predictions on novel images not used in the regression training and compared them to the actual neural site's output for those images (Figs. 4.3a and 4.5a). Aggregating over images, we computed the goodness-of-fit r^2 value, normalized by the neural site's trial-by-trial variability to obtain the percentage of explained variance for that site. This was done separately for each measured neural site to obtain an explained variance distribution (Fig. 4.3b and Fig. 4.5b. The overall IT predictivity of a model was defined as the median of this distribution over all measured IT sites (Fig. 4.3c and Fig. 4.5c; see Methods). To measure performance, we trained Support Vector Machine [Pedregosa et al., 2011] classifiers with $l2$ regularization for three types of tasks supported by the testing image set, including 8-way basic category classification (ie. Animals vs Boats vs Cars, etc.), 8-way car identification (Astra vs Beetle vs Clio, etc.), and 8-way face identification, separately for each of the three levels of variation in the testing image set. 8-way task choices were computed as a maximum over margins from 8 binary One-Vs-All (OVA) classifiers [Pedregosa et al., 2011]. Fig. 4.2b shows cross-validated performance accuracies (defined as the fraction of correct predictions averaged over test splits) for the 8-way basic categorization task at the three variation levels. Fig. 4.12 shows accuracies for subordinate identification tasks as well Methods for details.)

Individual neural site predictions

As described in the main text, we used a standard methodology for assessing a model's ability to predict individual sites [Carandini et al., 2005a; Cadieu et al., 2007], in which each site is modeled as a linear combination of model outputs. In this procedure, linear regression was used to determine weightings of top-level model outputs which best fit a given neurons' output on a randomly chosen subset of the testing images. The remaining images were used to measure the accuracy of the prediction. Results from multiple random subsets were assessed independently and averaged to ensure statistical validity. Linear regressor results are reported for 10 splits of cross-validation, using 50%/50% train/test splits. Regression weights were obtained using a simple Partial Least Squares (PLS) regression procedure, using 25 retained components [Helland, 2006; Pedregosa et al., 2011]. For each measured site, separate neural response predictions and cross validated goodness-of-fit r^2 values were obtained. The percentage of explained variance was then computed on a per-site basis by normalizing the r^2 prediction value for that site by the site's Spearman-Brown corrected split-half self-consistency over image presentation repetitions.

To help interpret the meaning of this linear regression technique, consider a hypothetical case in which the responses for all IT neurons in one "source" animal are known on a set of image stimuli, and the goal is to use this data to predict the response of a random sample of IT neurons from a second "target" animal. This is a problem of *neuron identification*, e.g for each target neuron in the target animal, determining which neuron(s) in the source animal correspond to that target neuron. While it is known that at the population code level the IT responses of several different animals (and even different primate species) are similar [Kriegeskorte, 2009], it is not known to what extent there is a 1-to-1 matching of responses between individual neural sites. There is likely to be significant individual variability between the specific tuning curves of units present in different animals, and it is not clear whether the IT units in all animals can be thought of as independent samples from a single master distribution of IT-like neurons. Hence, to explain a given IT unit in the target animal's IT might require linear combinations of multiple source animal IT units, even if a complete sample of neurons from the source animal was available. In more mathematical terms, it is plausible that the best linear fit from one animal's IT to another's would not be particularly sparse. Because it is currently not yet known how sparse the between-animal mapping actually is, in the present work each model's output units is treated a *basis* from which any observed IT must be constructed, with no prior on the expected sparsity of the weighted sums. While in our experiments we did collect responses from units in two animals, we do not have enough units from either animal separately to draw a meaningful conclusion as to what the empirical sparsity distribution is, since accurate estimation would likely require on the order of $\sim 10^3$ units from a single animal. If recordings in multiple animals with enough units and images to assess cross-animal fitting sparsity becomes available, such data will be useful to falsify our — or any — model or IT, because the distributions of sparsenesses of the linear mappings from the model to any one population should match the typical animal-to-animal sparseness distribution.

This observation helps clarify the relationship between our work and some existing work on neural fitting (e.g. [Connor et al., 2007; Brincat and Connor, 2004; Sharpee et al., 2012]). In that line of work, which has provided very useful insight into the units up to the V4 area, a different non-linear model —- roughly equivalent to a single CNN network in our model, described below — is fitted separately for each observed visual neuron. Unlike that work, the present results yield a generative model of a neural population as a whole, one that can fit not just the tuning curves of observed neurons but also predicts what types of neurons a typical sample population should *a priori* contain.

We also implemented two stronger tests of generalization: (1) object-level generalization, in which the regressor sample set contained images of only 32 object exemplars (4 in each of 8 categories), with results assessed only on the remaining 32 objects, averaging results across many such object splits, and (2) category-level generalization, in which the regressor sample set contained images of only half the categories (8 objects in each of (e.g.) animal, boat, car, and chair categories), with results assessed only on images of the other categories (8 objects in face, fruit, plant and table categories), averaged across many such category splits. Fig. 4.15 shows neural fitting results for object and category generalizations.

Prediction accuracy remains high for the object-level generalization, suggesting that the HMO model is effective at the generalizing neural predictions across a wide range of natural image variability. Neuron-level predictions of all models fall off somewhat in the category generalization case, though relative magnitude and ordering between models are preserved. To interpret this, it is again useful to consider the hypothetical animal-to-animal neural identification task described above. Even with *completely comprehensive* source animal response data (e.g, *all* the units in IT — the perfect "model"), the neuron identification task involves some uncertainty. If the training image stimulus set is not comprehensive enough to completely identify the target neuron, predictions from source to target will break down on images outside that image set. When the data used to identify the target neuron is narrowed to a very limited semantic slice of image space (e.g. a fraction of the object categories), it is expected that it will become difficult to identify that specific neuron from responses to just those images. For example, if all the images in the training set were only of simple shapes of a uniform size and geometry, it would be impossible to effectively carry out the neuron identification procedure (via linear regression or any other technique). It is instructive to compare this to the results for the *population* level coding (see section on Representational Dissimilarity Matrices below), where even in the category generalization case, predictions remain accurate.

Linear classifier analysis

Object recognition performance was assessed by training linear classifiers on model and neural output. Linear classifiers are a standard tool for analyzing the performance capacity of a featural representation of stimulus data on discrete classification problems [Hung et al., 2005a; Rust et al., 2006]. For any fixed population of output features (from either a model or neural population), a linear classifier determines a linear weighting of the units which best predicts classification labels on a sample set of training images. Category predictions are then made for stimuli held out from the weight training set, and accuracy is assessed on these held-out images. To reduce the noise in estimating accuracy values, results are averaged over a number of independent splittings of the data into training and testing portions. In our case, the output features of a model on each stimulus are (by definition) the set of scalar values for each top-level model unit when evaluated on that stimulus, a typical procedure from computer vision studies [Mutch and Lowe, 2008; LeCun and Bengio, 1995]. For neuronal sites, the output features are defined as the vector of scalar firing rates for each unit, as is typical in neural decoding studies [Hung et al., 2005a; Rust et al., 2006].

The values shown in Figs. 4.2b and 4.12 are for classifiers trained with 75% train / 25% test splits, averaged over 20 random category-balanced splits. However, the absolute values of performance for a linear classifier depend on the choice of the number of training examples used. To ensure that our conclusions were not dependent this choice, we computed performance curves for varying numbers of training examples. While absolute performances did vary as a function of training examples, we found that the relative ordering of performances did not (see Fig. 4.13a). Moreover, representations that were effective at high variation level (e.g. the IT neuronal population and the HMO model units) achieved most of their performance with comparatively small numbers of training examples.

Absolute performance also varies with the number of features used — the number of neuronal sites sampled in the case of neural data, or the number of top-end units in the case of models. As with the number of training examples, we would like to be sure that our results do not depend strongly on the number of sampled units. However, the analysis of dependence on number of units is somewhat less straightforward than analysis of training set size dependence, because it is not immediately clear how to fairly equate one neural unit with a fixed number of sample model units. Ideally, we would have extremely large numbers of both kinds of units and then simply make comparisons on complete population samples. Given the limitations of neural data collection, the limiting factor in this work is the number of neural sites sampled. We believe, however, that for the three key comparisons that we make, sample sizes issues do not strongly impact our results:

1. IT neural sample vs V4 neural sample: At approximately the same number of neural samples (168 to 128), the performance values at high variation image set are extremely widely separated. While it is unlikely that this difference is due to neural sample size, to ensure that this is true, we computed performance curves for subsamples of the population of different sizes (Fig. 4.13b), averaging over many subsamples of each fixed size. At all sizes, the IT population

strongly outperformances the V4 population. Because these subsample curves appear to have a predictably logarithmic shape, we also fit the data to a logarithmic functional form to extrapolate approximately how many units would be required to achieve the performance measured from the human behavioral experiments. Our estimate suggests that approximately 1050 ± 300 IT units would be consistent with human performance, whereas $\sim 10^7$ V4 units would be required. Such estimates are necessarily very rough, but they illustrate the magnitude of the differences between these neural populations.

2. IT neural sample vs existing comparison models: In all cases the models sampled produced more output features than we had neural sites (4096 in the case of HMAX, and 24316 for the V2-like model, and 86400 for the V1-like model; see below for more information on the these models). The results in Fig. 4.2b show performances computed with the total number of model features in each case. The implication of this is that, even with thousands or tens of thousands of features, these models are not able to equal the performance level of even 168 randomly chosen IT units. Equating the number of features, either by increasing the number IT samples or decreasing the number of model features, would only make the magnitude of the gap larger.

3. IT neural sample vs the HMO model outputs. Our claim is that the HMO model is plausibly correct, i.e. it achieves roughly the right performance for a reasonable number of samples. The HMO model performs at approximately human levels with 1250 top-end outputs, within the sampling error of the number of IT units suggested by extrapolation to achieve human performance. We also subsampled the HMO model to have as many features as our IT sample (168), and found that while the performance degraded somewhat, it did not

drop below measured IT performance levels. However, it is certainly possible that investigating the detailed dependence of model and IT performance on number of samples would allow us to falsify the HMO model. This falsification would be of interest for spurring future work, but for the reasons described above, it would be unlikely to invalidate the claims made in the present work.

Population code representational dissimilarity matrices

Given stimuli $S = s_1, \ldots, s_k$ and vectors of neural population responses $R = \vec{r}_1, \ldots, \vec{r}_k$ in which r_{ij} is the response of the j-th neuron to the i-th stimulus, we following [Kriegeskorte, 2009] by defining the Representational Dissimilarity Matrix as

$$RDM(R)_{ij} = 1 - \frac{\mathrm{cov}(\vec{r}_i, \vec{r}_j)}{\sqrt{\mathrm{var}(\vec{r}_i) \cdot \mathrm{var}(\vec{r}_j)}}.$$

RDM structure is indicative of a range of behavior that a given neural population can support [Kriegeskorte et al., 2008b], and two populations can have similar RDMs on a given stimulus set (and similar population-level classification performance) even if the low-level details of the neural responses are somewhat different. Because they involve correlations over the feature dimension, Representational Dissimilarity Matrices (RDMs) alleviate some of the ambiguities just discussed in analyzing individual units. We produced RDMs for the IT and V4 neural populations, as well as for each of the model-based synthetic IT and V4 neural populations using weights obtained from the regressions for the individual site fits (Fig. 4.3d-e and Fig. 4.14). Following Kriegeskorte [Kriegeskorte et al., 2008b], we measured similarity between population representations by assessing the Spearman rank correlations between the RDMs for the two populations. In addition to the standard image-level RDM, in which each pair of test images gives rise to an element of the RDM, we also computed object-level

RDMs by averaging population responses for each object before computing correlations (so that each pair of objects gives rise to an element of the 64x64 object-level RDM). Similarity of the HMO model object-level RDMs with the IT object-level RDMs are what shown and quantified in Fig. 4.3d-e.

The RDM for the IT neural population we measured has clear block-diagonal structure — associated with IT's exceptionally high categorization performance — as well as off-diagonal structure that characterizes the IT neural representation more finely than any single performance metric. In contrast, the RDM for the V4 population shows how high levels of variation blur out explicit categorical structure for intermediate visual areas, providing a clear visualization of the contrasting population responses underlying the high-variation V4-IT performance gap shown in Fig. 4.2b.

We also computed RDMs for object- and category-level generalizations, using the weightings from the regressions produced as described above in the section on individual neural site predictions. It is instructive to notice that the HMO model maintains high levels of IT similarity even at category-level generalizations (Fig. 4.3d-e), suggesting that while individual IT units may be hard to predict from a semantically narrow slice of image space (e.g. half the categories only), the overall population code structure remains well predicted.

4.4.3 Computational model class

In the initial high-throughput experiments in this paper, the models used are single Convolutional Neural Networks (CNNs). Each individual layer is composed of operations including local pooling, normalization, thresholding and filterbank convolution, which are combined by the following serial composition, where X is a 2-dimensional image and subscripts $\Theta = (\theta_p, \theta_N, \theta_T, \theta_F)$ denote specific parameter choices for the constituent operations:

$$N_\Theta(X) = Normalize_{\theta_N}(Pool_{\theta_p}(Threshold_{\theta_T}(Filter_{\theta_F}(X))))$$

To produce deep CNNs, layers of the form N_Θ are stacked hierarchically by serial composition, taking output of the i-th layer as the input to the $i+1$st layer. In the HMO optimization procedure, we extend the class of convolutional neural networks so that at any stage, networks can consist of mixtures of CNNs where each component has a potentially distinct set of parameters (e.g. pooling size, number of filters, etc), representing different types of units with different response properties [Martin and Schroder, 2013].

Comparison models

We compared results for performance, single site neural fitting, and population-level similarity for a variety of computational models, including:

- The trivial **Pixel** control, in which 256x256 square images were flattened into a 65536-dimensional "feature" representation. The pixel features provided a control against the most basic types of low-level image confounds.
- The baseline **SIFT** computer vision model [Lowe, 2004]. This model provided another control against low-level image confounds.
- An optimized **V1-like** model [Pinto et al., 2008a], built on grid of Gabor edges at a variety of frequencies, phases, and orientations. This model provided an approximation of a comparison point to lower levels in the ventral visual stream.

- A recent **V2-like** model [Freeman and Simoncelli, 2011], composed of conjunctions of Gabors. This model provides an approximation of the second level of the ventral stream.

- **HMAX** [Serre et al., 2007a; Mutch and Lowe, 2008], a multi-layer convolutional neural network model targeted at modeling higher ventral cortex. Because it is a deep network, HMAX has large IT-like receptive fields. HMAX is one of main existing "first-principles"-based models that attempts to build up invariance through hierarchical alternation of simple and complex cell-like layers.

- **PLOS09**, a recent three-layer convolutional neural network [Pinto et al., 2009], which also has large IT-like receptive fields and which was discovered via a high-throughput screening procedure that was a predecessor to the HMO procedure.

Ideal observer semantic "models"

As shown in Figs. 4.3 and 4.5 we also computed the IT- and V4-predictivity for ideal-observer semantic "models" [Geisler, 2003]. Though these ideal observers are *not* image computable models, given our perfect knowledge of image metadata, we were able to compute explained variance percentages using the same linear regression protocol applied to the image-computable models. We evaluated two ideal observers including:

- A **category ideal observer**. This "model" has eight features, one for each of the 8 categories present in the test image set. For each image, the i-th feature is 1 if the image contains an object of category i, otherwise it is 0. For each IT unit, the 8 linear regression weights for this feature set effectively describe how much each category contributes to that unit's response.

- A **all-variable ideal observer**. This "model" is given oracular access to all metadata parameter variables for the images, with one feature for each of 64 object identities (similar to the category ideal observer features above), in additional to features reporting object position, size, scale and image background.

If the IT or V4 explained variance for these (or any) ideal observers were close to 100%, then they would provide a conceptually interpretable explanation of neural variation, a very scientifically desirable result. In fact, the explained variance percentages are significantly less than 100% for the ideal observers we tested (though of course other better ones might be found, e.g. by taking into account 3-d object curvature). These ideal observers therefore serve as useful controls to which other computational models can be compared. For example, the ideal category model serves to control for the minimum amount of IT explained variance that should be expected from *any* model that has high categorization performance. Insofar as a model with high categorization performance explains *more* explained variance than the ideal category model, that additional predictivity can be attributed to the constraints of the model class.

Convolutional neural network model class

Here, we mathematically specify the basic class of Convolutional Neural Network (CNN) models used in this paper. These principles are consistent with a large parameter space of possible networks. The specific parameterized space of networks we use is close to that described in [Pinto et al., 2009], with one, two, or three convolutional layers. Each layer is characterized by a fixed set of parameters, but parameter values can differ between layers. This parameter space expresses an inclusive version of the hierarchical feedforward network concept, and contains models similar to that used in many previous studies — for different parameter values [Pinto et al., 2008a; Freeman and Simoncelli, 2011; Serre et al., 2007a; Mutch and Lowe, 2008].

More specifically, each individual layer is composed of operations including local pooling, normalization, thresholding and filterbank convolution, which are combined as follows:

$$N_\Theta(X) = Normalize_{\theta_N}(Pool_{\theta_p}(Threshold_{\theta_T}(Filter_{\theta_F}(X)))) \tag{4.1}$$

where X is a 2-dimensional input image. The subscripts $\Theta = (\theta_p, \theta_N, \theta_T, \theta_F)$ denote the specific parameter choices for the constituent operations, setting radii, exponents and thresholds, as in [Pinto et al., 2009]. Similar to previous studies, we also use randomly chosen filterbank templates in all models, but additionally allow the mean and variance of the filterbank to vary as parameters. Functions of the form N_Θ are the simplest computational units that we operate on, and are thought to be plausible representations of what happens in a single cortical layer [DiCarlo et al., 2012]. To produce deep CNNs, layers of the form N_Θ are **stacked hierarchically**:

$$\ldots \mathbf{P}^{\ell-1}_{\theta_{P,\ell-1}} \xrightarrow{Filter} \mathbf{F}^{\ell}_{\theta_{F,\ell}} \xrightarrow{Threshold} \mathbf{T}^{\ell}_{\theta_{T,\ell}} \xrightarrow{Pool} \mathbf{N}^{\ell}_{\theta_{P,\ell}} \xrightarrow{Normalize} \mathbf{P}^{\ell}_{\theta_{N,\ell}} \cdots \tag{4.2}$$

where ℓ is layer number and the initial input at the 0-th layer is the image pixel array X. We denote such a stacking operation as $\otimes$, so that the stacked hierarchical model can be written as

$$\mathbf{N} \equiv \otimes_{i=1}^{k} N_{\Theta_i}.$$

Let $\mathcal{N}_k$ denote the space of all stacked networks (N) of depth k or less. In this study, our CNNs are networks of depth $k = 3$ or less.

Mixture networks

We extend the class of convolutional neural networks by a fourth principle, namely that at any stage, networks can consist of mixtures of CNNs where each component has a potentially distinct set of parameters (e.g. pooling size, number of filters, etc), representing different types of units with different response properties [Martin and Schroder, 2013]. Such mixture networks may combine components of differing complexity, which correspond to anatomical bypass connections within the ventral stream [Nakamura et al., 2011]. See Fig. 4.8b.

For a mathematical formulation of this idea, note that because the networks in $\mathcal{N}_3$ are **convolutional**, they can be combined in a standard fashion. Specifically, given a sequence of individual modules $\mathbf{N}(\Theta_{i1}, \Theta_{12}, \ldots, \Theta_{in_i})$ for $i \in [1, \ldots, J]$, possibly of different depths, the mixtur network is defined by aligning the module output layers along the spatial convolutional dimension. Since the outputs of each of the modules is a 3-dimensional tensor, this alignment is well-defined up to a rescaling factor in the spatial dimension. We denote this alignment operation by the symbol $\oplus$, so that a combined mixture network can be written as:

$$\mathbb{N} \equiv \oplus_{i=1}^{J}\mathbf{N}(\Theta_{i1}, \Theta_{12}, \ldots, \Theta_{in_i}).$$

The total output of networks of this form are also a 3-dimensional tensors, so they too can be stacked with the $\otimes$ operation to form more complicated, deeper hierarchies. By definition, the full class $\mathbb{N}$ consists of all the networks formed by iteratively composed the stacking ($\otimes$) operation and the combination ($\oplus$) operation. Conceptually, members of $\mathbb{N}$ are nonlinear mixtures of modules chosen from a "base class" of simpler "homogenous" neural networks (e.g. the elements of $\mathcal{N}$. Schematically, $\otimes$ is a "vertical" composition relationship, increasing the depth complexity of the network. Biologically, it is plausible to think of $\otimes$ as corresponding to producing complex nonlinear representations by feedforward layering. Conversely, $\oplus$ is a "horizontal" composition relationship, increasing the breadth complexity of the network. Biologically, this may correspond to the idea of mixing heterogeneous populations of different types of units in a given area.

Hierarchical modular optimization

The Hierarchical Modular Optimization (HMO) procedure [Yamins*, Hong*, Cadieu, and Dicarlo, 2013] is a computational optimization procedure designed to identify high-performing network architectures from the space $\mathbb{N}$. Intuitively, it is a version of adaptive boosting in which rounds of optimization are interleaved with boosting and hierarchical stacking [Schapire, 1999]. The process first analyses error patterns in the recognition predictions of candidate networks, picking complementary components, e.g. those with optimally non-overlapping errors. Subsequent rounds of optimization attempt to optimize a criteria weighted toward those stimuli that are misclassified by the first-round results. As a result, complementary components emerge without having to pre-specify the corresponding sub-tasks semantically (or in any other way), mapping the complex structure of high-variation recognition problems onto the parameter space of neurally-plausible computations. These components are then aligned along their convolutional dimensions and used as inputs to repeat the same procedure hierarchically to build more complex nonlinearities. While other possible optimization procedures could potentially be used to create high-performing neural networks [Krizhevsky et al., 2012], the HMO process may be particularly efficient because it explicitly takes advantage of the complementary strengths of different components within the large space of network architectures.

This section describes details of the Hierarchical Modular Optimization (HMO) procedure. Suppose that $N \in \mathcal{N}$ and S is a screening stimulus set. Let E be the binary-valued classification correctness indicator, assigning to each stimulus image s 1 or 0 according to whether the screening task prediction was right or wrong, where the prediction for each s was made by employing Maximum Correlation Classifiers (MCCs, see e.g. [Buciu and Pitas, 2003]) on the output features of N with 3-fold cross-validation (see Methods section of main paper describing screening set metric). Let

$$\text{performance}(N, S) = \sum_{s \in S} E(N(s)).$$

To efficiently find N that maximizes performance(N, S), the HMO procedure follows these steps:

1. Optimization: Optimize the performance function within the class of single-stack networks of some fixed depth d_1, obtaining an optimization trajectory of networks in $\mathcal{N}_{d_1}$. (See Figs. 4.8c and 4.11a, left.) The optimization procedure that we use is Hyperparameter Tree Parzen Estimator, as described in [Bergstra et al., 2012]. This procedure is effective in large parameter spaces that include discrete and continuous parameters.

2. Boosting: Consider the set of networks explored during step 1 as a set of weak learners, and apply a standard boosting algorithm (Adaboost) to identify some number of networks $\mathrm{N}_{11}, \ldots, \mathrm{N}_{1l_1}$ whose error patterns are complementary (Fig. 4.8c, right panel).

3. Combination: Form the heterogeneous network $\mathbb{N}_1 = \oplus_i \mathrm{N}_{1i}$ and evaluate $E(\mathbb{N}_1(s))$ for all $s \in S$.

4. Error-based Reweighting: Repeat step 1, but reweight the scoring to give the j-th stimulus s_j weight 0 if N_1 is correct in s_j, and 1 otherwise. That is, the performance function to be optimized for N is now

$$\sum_{s \in S} E(N_1(s)) \cdot E(N(s)).$$

Repeat the step 2 on the results of the optimization trajectory obtained to get models $N_{21}, \ldots N_{2k_2}$, and repeat step 3 (see e.g. Fig. 4.8c and 4.11 A, right). Steps 1, 2, 3 are repeated K times.

After K repetitions of this process, we will have obtained a mixture network $N = \oplus_{i \leq K, j \leq k_i} N_{ij}$. The process can then simply be terminated, or repeated with the output of N as the input to another stacked network. In the latter case, the next layer is chosen using the model class $\mathcal{N}_{d_2}$ to draw from, for some fixed depth d_2, and using the same adaptive hyperparameter boosting procedure. The meta-parameters of the HMO procedure include the numbers of components $l_1, l_2, \ldots$ to be selected at each boosting round, the number of times K that the interleaved boosting and optimization is repeated and the number of times M this procedure is stacked. For the purposes of this work, we fixed the metaparameters $K = 3$, $l_1 = l_2 = l_3 = 10$, and $M = 2$ (with $d_1 = 3$, $d_2 = 1$).

Model screening procedure

To construct a specific model network, we applied HMO to a screening task (Figs. 4.7b and 4.11). Like the testing set, the screening set was designed to be very challenging — having high levels of object pose, position and scale variation [DiCarlo and Cox, 2007]. However, to ensure that a fair test could be made, in all other regards the screening images were distinct from the testing image set, containing objects in totally non-overlapping semantic categories, using none of the same background scenes, lighting, or noise conditions. The image set used for the HMO screening procedure consisted of 4500 images of 36 distinct objects, chosen from one of nine categories, including bodies, building, flowers, guns, musical instruments, jewelry, shoes, tools, and trees. As in the testing set, high-variation subset, objects were shown in varying positions, sizes, and poses, placed in a variety of uncorrelated natural backgrounds scenes. Lighting was provided by ambient environment reflection, and speckle noise was added to simulate natural image distortions. Images were rendered with the Panda3d package [Goslin and Mine, 2004].

The relationship between the screening set and testing set is intended to be similar to that between any two typical samples of natural images: having some high-level natural statistical commonalities, but otherwise quite different specific content. For this reason, any performance increases that could be demonstrated to transfer from the screening to the testing set are likely to also transfer, at least to some extent, to other high-variation image sets.

The screening objective sought to minimize classification performance error on the 36-way object classification task (no categorical semantic information was used), as assessed by training unregularized MCC classifiers with 3-fold cross-validated 50%/50% train/test splits. Using the HMO procedure on this screening set, we generated a network HMO_0, which produces 1250-dimensional feature vectors for any input stimulus. HMO_0 is the model which we refer to throughout the paper as the "HMO Model", and which we used for all testing evaluation.

In the optimization, candidate networks were first evaluated on overall performance metric, and performance gradients in parameter space were identified as seen in the trend toward decreasing screening loss (Step 1 — see Fig. 4.11A, left panel, blue dots). 10 components were identified by Boosting (Step 2) and combined. In subsequent rounds (e.g. Fig. 4.11A, right panel, red dots) the optimization criterion was biased toward weighting more heavily errors of the architectures from earlier rounds (Step 4). Decreasing loss in these later rounds indicates that models are improving at the subset of images that confused the components identified in Round 1. The complementary model components identified in the two different optimization rounds were associated with different directions in the overall large parameter space of possible neural-like computations that effectively solve different subtasks of the overall recognition task (see Fig. 4.11B). As expected, training performance increases as components are combined (see Fig. 4.11C).

Assessment

As described in the main text, we then assessed the HMO_0 model against the testing dataset (Fig. 4.7a). The HMO_0 model showed high performance on testing set, as described in the main text, Fig. 4.2b, and Fig. 4.12. Comparisons to neural data showed that the HMO_0 model also had significantly power to explain neural data, both at the individual site level (Figs. 4.3a-c and 4.5) and the population level (Figs. 4.3d-e and 4.14). The HMO model is a significantly closer match to IT population representations at all variation levels, but the difference is especially evident at the high variation level that most clearly exposes how the high-level IT representation differs from the lower-level V4 representation (Fig. 4.14, black bars).

Subsequently, we determined the stability of the HMO procedure by running it on a variety of alternative screening sets with different choices of objects and categories, varying the numbers of within-category exemplars and varying amounts of semantic similarity to the testing set. Performance and neural fitting ability were largely stable to these changes. Though some of these later models exhibited higher performance and neural explanatory power than the initial HMO_0 model, to prevent domain overfitting we report only the results of the initial model HMO_0 constructed before any testing set results were obtained.

It is important to note that how our screening process connects to the evaluation of other models. In the cases of the SIFT, V1like and V2-like models, we did not pre-train those models using the screening set: this is because those models do not accept pre-training data at all. In the case of HMAX, which does accept pre-training data, we used the *testing* data itself for pre-training, to give that model that highest chance of performance success. Separately, we also performed a pre-training of the HMAX model using the screening set and then re-extracted it on the testing set, but found that this only further decreased final performance and neural fit results of the HMAX model (e.g. learned parameters did not effectively transfer from the screening to the testing set).

Another issue relevant to comparison of models is the question of numbers of total internal units. In the mixture models that we used to create the HMO model, the numbers of filters at each layer were kept very small (≤ 24) to ensure that a total combined model composed of several such components would not be unmanageably large. In the HMO_0 model, the total number of units is approximately the same as that in the HMAX model, and the total number of output features is somewhat smaller (1250 vs. 4096).

Correlation experiments

Performance and neural predictivity results suggest that as performance on high-variation tasks increases, metrics of neural similarity also increase (Fig. 4.1b). To determine whether this correlation is a general feature of the deep feedforward architectures defined here, we ran several additional high-throughput experiments, evaluating a large number of candidate model architectures and measuring categorization performance and IT neural predictivity for each model (Fig. 4.1a and Fig. 4.9). Specifically, we performed three high-throughput searches of the parameter space $\mathcal{N}_3$ described in the above:

1. **Random selection.** We drew several thousand randomly sampled models from the parameter space $\mathcal{N}_3$. For each one, we computed linear classifiers for performance and linear regressors for IT predictivity, as described above. Each green point in Fig. 4.1a corresponds to one such model. In this condition, there is a significant correlation between performance and IT predictivity ($r = 0.55$, $n = 2016$). Negative values on the y-axis correspond to models having negative goodness-of-fit (the r^2 coefficient of determination statistic), due overfitting on the training images. Fig. 4.9, left, shows model performance for as a function of time during the procedure; the lack of any trend corresponds to random sampling of models.

2. **Performance optimization.** Using the recently developed Hyperopt meta-parameter optimization algorithm [Bergstra et al., 2012], we performed a directed search for network parameters that maximized performance on the high-variation 8-way categorization task (Fig. 4.1a, blue points). This optimization was carried out using the recently developed hyperparameter optimization algorithm Hyperopt [Bergstra et al., 2012]. Via this optimization, absolute performance and fitting values were significantly improved compared to the random condition. Moreover, though the optimization was done without reference to any neural data, the correlation between performance and IT predictivity actually increased significantly ($r = 0.78$, $n = 2043$). Fig. 4.9, center panel shows the optimization criterion as a function of timestep during the optimization procedure; the upward trend is due to the optimization process. While the

optimization gains toward the end of the optimization process are slow and appear to be plateau, small improvements are still observed.

3. **IT Predictivity optimization.** In the third experiment, we directly optimized model architecture for IT predictivity, this time without reference to performance (Fig. 4.1a, orange dots). The correlation is comparable to the performance-optimized condition ($r = 0.80$, $n = 1876$), but the optimization plateau occurs significantly earlier (see Fig. 4.9, right panel; we repeated the optimization multiple times, and obtained the same result each time. This suggests that continued optimization would not be effective.) Moreover, the best-performing models from the performance-optimization experiment predict IT neural output as well as the models explicitly optimized for the predictivity objective, while the reverse does not hold.

The results of these experiments support three inferences. First, model performance is modestly correlated with neural predictivity in a random selection regime. Second, optimization pressure for either metric produces markedly better cross-validated accuracy on the optimized axis, and in doing so significantly strengthens the correlations with the other non-optimized metric. Third, when optimizing for performance, the best-performing models predict neural output approximately as well as the most predictive models selected explicitly for neural predictivity, but not vice-versa. The feedforward model architecture class itself imposes a relationship between high-level behavior (performance) and more detailed neural mechanisms, but directed optimization focuses on a region within network parameter space where this constraint is much stronger.

The inclusion of the category ideal observer (purple square in Fig. 4.3d) shows an effective negative control on the performance-predictivity relationship: it lies significantly off the main trend, making it visually clear how the correlation arises from a combination of architectural and performance constraints working in concert.[2] However, this ideal observer is *not* an image computable model, It would be especially instructive to identify a image-computable algorithm that achieved invariant object recognition high performance but low neural IT neural consistency. If such an algorithm existed, its architecture might illustrate a very non-neural solution to object recognition tasks as a purely computer vision problem. With current understanding, we cannot rule out the possibility that such an algorithm does *not* exist — e.g. recent high-performing computer vision systems are deep convolutional neural networks e.g [Krizhevsky et al., 2012].

Fig. 4.1a also implies that even with intensive optimization, individual models in the $\mathcal{N}_3$ are limited in performance and neural prediction ability, underscoring the need for an enlarged model class. However, further analysis of the results of these optimization experiments provides insight into how to construct a more effective model class. In Fig. 4.16, we show scatter plots of model performance on pairs of binary subtasks, e.g. performance on the 2-way cars-vs-planes task as compared to performance on the 2-way boats-vs-chairs task. These plots show that, as the optimization algorithm explores parameter space, it identifies mutually-exclusive subspaces that are effective for some of the natural subtasks defined in the overall task space. The highest performing architectures for one subtask are often significantly suboptimal for other subtasks, leading to "v-shaped" subtask-vs-subtask scatter plots. In choosing a single architecture that is best for overall performance, the optimization is forced to trade off performance on some of these subtasks.

[2] Note that a converse control, in which a model has very high neural consistency for a population of IT units but low performance, cannot exist. IT units are already known to have high performance, so any model that matches IT units sufficiently well must also have high performance.

The effectiveness of optimized mixture models (such as HMO) may be understood in the context of Fig. 4.16, which suggests that models composed of mixtures from the $\mathcal{N}_3$ class might be significantly more effective than any single model alone. Such mixtures are also suggested by the observation from neurophysiology studies that patches within IT are selectively responsive for distinct object classes [Downing et al., 2006; Kanwisher et al., 1997; Freiwald and Tsao, 2010]. Intuitively, such subregions might correspond to architecturally specialized structures within the larger feedforward class. Mixture models avoid the tradeoffs inherent in individual feedforward structures by combining several pareto-optimal network architectures. By identifying particularly effective mixture combinations, the HMO procedure overcomes these limitations efficiently. In addition, however, a key ingredient for the HMO model's success is that the components constituting the model, which were (by construction) complementary on the original screening set, were still complementary on the testing set. This holds even though the testing set had entirely distinct object categories, so the basis on which the complementarity of the components was originally discovered — non-overlapping error patterns in screening-set object identity judgements — is no longer even applicable. This strongly rules out image domain-specific overfitting and suggests that mixture components discovered by performance optimization may form a generically useful visual representational basis that can be recombined to solve new object recognition problems. In fact, achieving high performance and neural fitting capability appears to require diversity in many of the parameters of the constituent components (Fig. 4.17).

Model parameter diversity analysis

We characterized model parameters in terms of per-component tuning specificity versus inter-component diversity. Tuning specificity is a measure of how specifically each parameter needed to be tuned to produce optimal performance. To compute this, we analyzed the distribution of each parameter's values along the optimization trajectory near the optimal point using the concept of entropy. By definition, the entropy of (N samples from) a distribution P is:

$$E(P) = \log(N) - \frac{1}{N}\sum_i n_i \log(n_i)$$

where N is the number of samples from the distribution, the sum is taken over possible values i of the distribution, and n_i is the number of samples with value i.

Suppose an optimal module component Θ^* occurs at timepoint t^* in the trajectory of one optimization run in the HMO process. Then, let $P_{p,k}(\Theta^*)$ be the distribution of values of parameter p in the k-neighborhood around t^* in the optimization trajectory, e.g.

$$P_{p,k}(\Theta^*) = \{\text{value of parameter } p \text{ at timepoints } t \in [t^* - k, \ldots, t, \ldots, t^* + k]\}.$$

The specificity of parameter p around optimal point Θ^* as, by definition,

$$-E(P_{p,k}(\Theta^*)).$$

Intuitively, this is because, if the distribution $P_{p,k}(\Theta^*)$ had high entropy, this indicated that the value of the parameter near the optimal point did not matter very much, and therefore was not tuned very specifically. If, on the other hand, the distribution had low entropy, it was tightly clustered around one or a few optimal values that the optimization had identified as being important, suggesting it was highly tuned. For the purposes, we took $k = 25$ timesteps, but values were not strongly sensitive to k with the range 10-100. For each parameter p, we report the median tuning specificity of that parameter, taken over all component modules.

Inter-component diversity is a measure of how variable a parameter is between the component modules. This was measured by computing, for each pair of components, how well separated the distributions of the parameter's values around each component were from each other. More formally, the *d-prime discriminability index*, d' for two distributions P_1 and P_2 is defined by

$$d'(P_1, P_2) = \frac{|\langle P_1 \rangle - \langle P_2 \rangle|}{\sqrt{0.5(\text{var}(P1) + \text{var}(P2))}}.$$

(The sample d-prime uses the sample versions of the mean and variances.) Suppose Θ_1^* and Θ_2* are two optimal components chosen by the HMO procedure. Then we measure separability for these two components as

$$d'(P_{p,k}(\Theta_1^*), P_{p,k}(\Theta_2^*)).$$

For each parameter p, we define inter-component diversity as the median of this separation value taken over all pairs of components Θ_1 and Θ_2. The higher the diversity, the more different the components were from each other, and vice versa.

Parameters that have both high tuning specificity and high inter-component diversity are both critical for performance, and required to be heterogeneous. Our results highlight certain types of parameters as being simultaneously highly tuned and diverse. This is particularly true for two broad classes of parameters, as can be seen Fig. 4.17, upper right: 1) local filter statistics, including filter mean and spread, and 2) the pooling exponents trading off between max-pooling and average-pooling [Riesenhuber and Poggio, 2000]. Other types of parameters are highly tuned but less diverse (nonlinear activation thresholds, in the lower right), while some appear less important overall (higher-level pooling and normalization kernel sizes, in the lower left). Interestingly, we observe that the parameter controlling the number of network layers ("depth") is both comparatively highly tuned and diverse suggests that allowing network modules of different levels of complexity in the heterogeneous models is important for achieving high model performance. As a result, the final model has a significant proportion of lower-complexity units projecting directly to the final layer, suggesting that bypass connections (e.g. projects from V1 to V4 or V2 to IT) may be a key functional feature of the ventral stream [Nakamura et al., 2011].

Taken together, these results point to a computationally rigorous explanation for why heterogeneity is observed in the receptive fields of ventral stream neurons both at the unit and sub-area levels [Martin and Schroder, 2013; Chelaru and Dragoi, 2008; Downing et al., 2006; Freiwald and Tsao, 2010].

4.4.4 Modeling with further optimization

Computational modeling done so far in this chapter was "simpler" than typical convolutional neural networks in that it did not involve fine tuning of filter values. While it achieved the human level performance in our visual tasks and well predicted monkey high-level visual neural activity, we also decided to explore filter-value optimization with back-propagation, a popular technique following [LeCun and Bengio, 1995; Krizhevsky et al., 2012]. This further optimization increased the neural predictivity slightly (Fig. 4.6**b**).

Basic definitions

Formally, an *image-like array* is a 3-dimensional dimensional floating-point array whose shape is (s, s, nc), where s is the *image size* and nc is the number of *channels* in the image. Let's begin by defining three basic operations on image-like arrays:

- **Filter**: this is a convolutional filterbank operation [LeCun and Bengio, 1995], which applies the same filter block equally to every point in an image-like array. It's parameters include:
  - The number of filters nf. This is a positive integer.
  - The size of the filter kernel fs, in pixels. This is an odd integer.
  - The stride of the convolution, s_f. This is a positive integer.
  - The specific filter values, denoted F, a floating-point matrix of shape (nc, fs, fs, nf), where nc is the number of channels in the input.
  - A bias vector b, of length nf.

  For any image-like array X of shape (s, s, nc) the output of $\mathbf{Filter}_F$ on X is the image-like array Y of shape $(s/s_f, s/s_f, nf)$ where

  $$Y(i, j, k) = b[k] + \frac{1}{fs^2} \sum F[:, :, :, k] \otimes N_{fs}(X, s_f \cdot i, s_f \cdot j)$$

  where $\otimes$ is pointwise array multiplication, $i, j \in [1, \ldots, s/s_f]$, $k \in [1, \ldots, nf]$, and $N_{fs}(X, i, j)$ denotes the square neighborhood of diameter fs at location i, j in X. The convolution is done with "same" mode, meaning that at the edges the image is padding with 0s to produce an output of the same shape as the input

- **Thres** is a rectified linear clipping operation. Its parameters are:
  - The value of the upper clipping threshold, t^{max}, which can be any floating value.
  - The value of the lower clipping threshold, t^{min}, which can be any floating value less than t_i^{max}.

  By definition,

  $$\textbf{Thres}(X) = max(min(X, t^{max}), t^{min}).$$

- **Pool** is a local pooling operation that aggregates values of the input, within each channel. Its parameters
  - The size of the pooling kernel, ps. This is an odd integer.
  - The pooling order po. This is 1, an even integer, or ∞.
  - The pooling stride s_p. This is a positive integer.

  By definition, for any image-like array X of shape (s, s, nc), the output of **Pool** on X is the image-like array Y of shape $(s/s_p, s/s_p, nc)$ where

  $$Y(i,j,k) = \left(\frac{1}{ps^2}\left(\sum N_{ps}(X^{po}, s_p \cdot i, s_p \cdot j)[:, :, k]\right)\right)^{1/po}$$

  where $i, j \in [1, \ldots, s/s_f]$, $k \in [1, \ldots, nc]$, and $N_{ps}(X, i, j)$ is the square neighborhood of diameter ps in X at location i, j. Notice that when $po = 1$, this is simple local averaging, and when $po = \infty$, this is max-pooling.

A *convolutional layer* is a composition of these three basic operations; that is, a function of the form

$$F_{(\theta_P,\theta_T,\theta_F)} = \mathbf{Pool}_{\theta_P} \circ \mathbf{Thres}_{\theta_T} \circ \mathbf{Filter}_{\theta_F}$$

where $(\theta_P, \theta_T, \theta_F)$ are choice of parameters for the three basic operations. A *hierarchical convolutional neural network* (HCNN) is a composition of convolutional layers, e.g.,

$$\mathcal{F} = F_L \circ F_{L-1} \circ \ldots \circ F_1.$$

The only two restriction that are required for composition to make sense are: (1) that the number of channels in layer i is equal to the number of filters in layer $i-1$, that is $nc_i = nf_{i-1}$ and (2) that the spatial size s_i of the image-like arrays is 1 or greater at every stage. If the spatial size every hits 1, then only thresholding or filtering operations with filter size 1 can be applied from then onwards. When this occurs, we say that the network is "fully connected" at that layer (and from then on).

In our case, the input image-like arrays are RGB images, so that the number of channels in in the first layer is 3, one for each color channel. (When applied to grayscale images we simply copy the grayscale values into the three channels).

Network selection

We divide the parameters that specify the layers of an HCNN into two classes, selected in two phases:

1. **Screening:** In which all the parameters *except* the filterblock and bias values where chosen. These parameters, which we refer to as the "architectural parameters", include the number of network layers, and at each layer, the number

of tilers, the sizes of the filter and pooling kernels, and the pooling order.

2. **Training:** In which, once the non-filter parameters are fixed, the filter-values and bias vectors for each layer are determined via error backpropagation.

Details of Error Backpropagation: For any given setting of architectural parameters, we used a standard neural network backpropagation algorithm [Krizhevsky et al., 2012] to set filter filters for the parameters. The training set that we used was the 2013 ImageNet Challenge set [Deng et al., 2009], which contains approximately 1.3 million images in 1000 natural categories. We filtered out any categories that were animals, boats, cars, chairs, fruits, planes or tables from this set (some of these categories do not appear anywhere in the ImageNet challenge set to begin with), retaining 799 categories containing a total of approximately 1 million images.

Details of Screening: As we have done in this chapter, we used high-throughput screening techniques [Bergstra et al., 2013] to select the architectural parameters. In this process, we randomly selected 50 draws of the number of layers and within-layer architecture parameters from a parameter space (see below), ran error backpropagation on the network with those parameters for 5 epochs of ImageNet, and then recorded the final training error. We then used Tree Parzen Estimation in the Hyperopt parameter optimization framework [Bergstra et al., 2013] to further select 150 additional architectural parameters, and again, ran backpropagation on these networks. After having run 200 networks, we selected the best such network and subjected it to further error backpropagation for 40 epochs. This optimal model had 6 layers. At every epoch of ImageNet training, we saved out checkpoints containing the filter and bias parameters.

The parameter space that we tested was defined by the following bounds:

- Number of layers ranged in [4, 5, 6].
- Filter sizes ranged in [3, 5, 7, 9].
- Pooling kernel sizes ranged in [3, 5, 7, 9].
- Pooling order ranged in [1, 2, 3, 4, 5, ∞].
- Upper clipping thresholds ranged in [1, ∞] and lower clipping thresholds ranged in [1, $-\infty$].

The remainder of the parameters were set to the following fixed values: number of filters at layer 1 was 96, at layer 2 was 256, at layer 3 was 512, and then at 256 for subsequent layers; strides at layer 1 was 1, at layer 2 was 2, at layer 3 was 2, and at 1 in subsequent layers.

Evaluation on the testing set

The model that achieved the best performance on the training set was selected for evaluation on the testing image set discussed earlier in Section 3.2.1 — i.e., the images on which we measured neural data and human performance. For each of the 40 checkpoints saved during model training (see above), and each layer of the network, we extracted features for all the testing images. This lead to six time series of length 40, each point of which is a $(5760, nf_i)$ matrix, where nf is the number of features at layer i. We then computed performance on each tasks on which we had earlier computed neural performance, for each layer and timepoint. For each model layer and each timepoint, we also computer the layer's ability to fit V4 and IT neural data, using procedures identical to the one in Section 4.4.2.

Acknowledgments

This work was partially supported by National Science Foundation grant IS 0964269 and US National Eye Institute R01-EY014970. Ha Hong was supported by the Samsung Scholarship. We thank Diego Ardila, Najib Majaj, and Nancy Kanwisher for useful conversations.

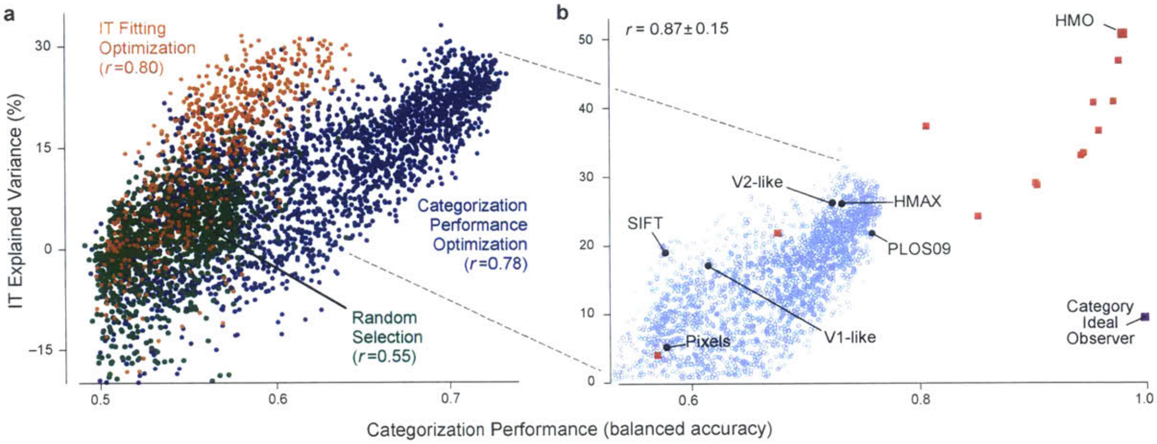


Figure 4.1: **Performance/IT-Predictivity Correlation.** (***a***) Object categorization performance vs IT neural explained variance percentage ("IT-Predictivity") for Convolutional Neural Network (CNN) models in three independent high throughput computational experiments (each point is a distinct neural network architecture). The x-axis shows performance (balanced accuracy, chance is 0.5) of the model output features on a high-variation categorization task; the y-axis shows the median single site IT explained variance percentage ($n = 168$ sites) of that model. Each dot corresponds to a distinct model selected from a large family of convolutional neural network architectures (see text). Models were selected by random draws from parameter space (green dots), object categorization performance-optimization (blue dots) or explicit IT predictivity-optimization (orange dots). (***b***) Pursuing the correlation identified in panel (a), a high-performing neural network was identified that matches human performance on a range of recognition tasks, the HMO model (see text). The object categorization performance vs IT neural predictivity correlation extends across a variety of models exhibiting a wide range of performance levels. Black circles include controls and published models; red squares are models produced during the HMO optimization procedure. The category ideal-observer (purple square) lies significantly off the main trend, but is *not* an actual image-computable model. The r-value is computed over red and black points. For reference, light blue circles indicate performance optimized models (blue dots) from panel (a).

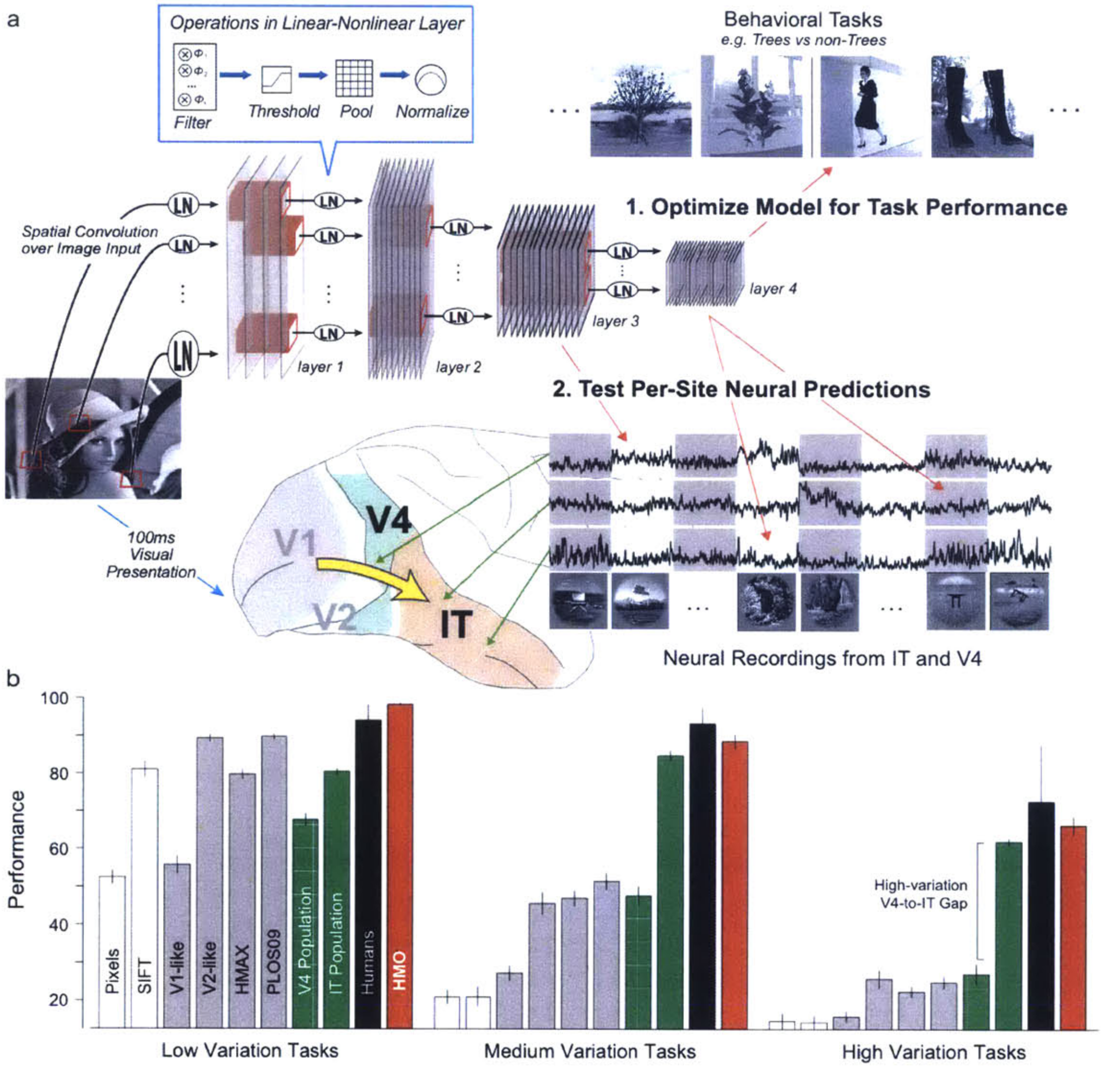

a
Operations in Linear-Nonlinear Layer
Filter
Threshold
Pool
Normalize
Behavioral Tasks
e.g. Trees vs non-Trees
Spatial Convolution over Image Input
LN
layer 1
layer 2
layer 3
layer 4
1. Optimize Model for Task Performance
2. Test Per-Site Neural Predictions
100ms Visual Presentation
V1
V2
V4
IT
Neural Recordings from IT and V4
b
100
80
60
40
20
Performance
Pixels
SIFT
V1-like
V2-like
HMAX
PLOS09
V4 Population
IT Population
Humans
HMO
High-variation V4-to-IT Gap
Low Variation Tasks
Medium Variation Tasks
High Variation Tasks

Figure 4.2 *(preceding page)*: **Neural-like Models Via Performance Optimization.** (***a***) We (1) used high-throughput computational methods to optimize the parameters of a hierarchical convolutional neural network (CNN) with Linear-Nonlinear (LN) layers for performance on a challenging invariant object recognition task. Using new test images distinct from those used to optimize the model, we then (2) compared output of each of the model's layers to IT neural responses, and the output of intermediate layers to V4 neural responses. To obtain neural data for comparison, we used chronically implanted multi-electrode arrays to record the responses of multi-unit sites in IT and V4, obtaining for each neural site the mean visually-evoked response to each of ~6000 complex images. (***b***) Object categorization performance results on the test images for eight-way object categorization at three increasing levels of object view variation (y-axis units are 8-way categorization percent-correct, chance is 12.5%). IT (green bars) and V4 (hatched green bars) neural responses, and computational models (gray and red bars) were collected on the same image set and used to train support vector machine (SVM) linear classifiers from which population performance accuracy was evaluated. Error bars are computed over train/test image splits. Human subject responses on the same tasks were collected via psychophysics experiments (black bars); error bars are computed over individual subjects.

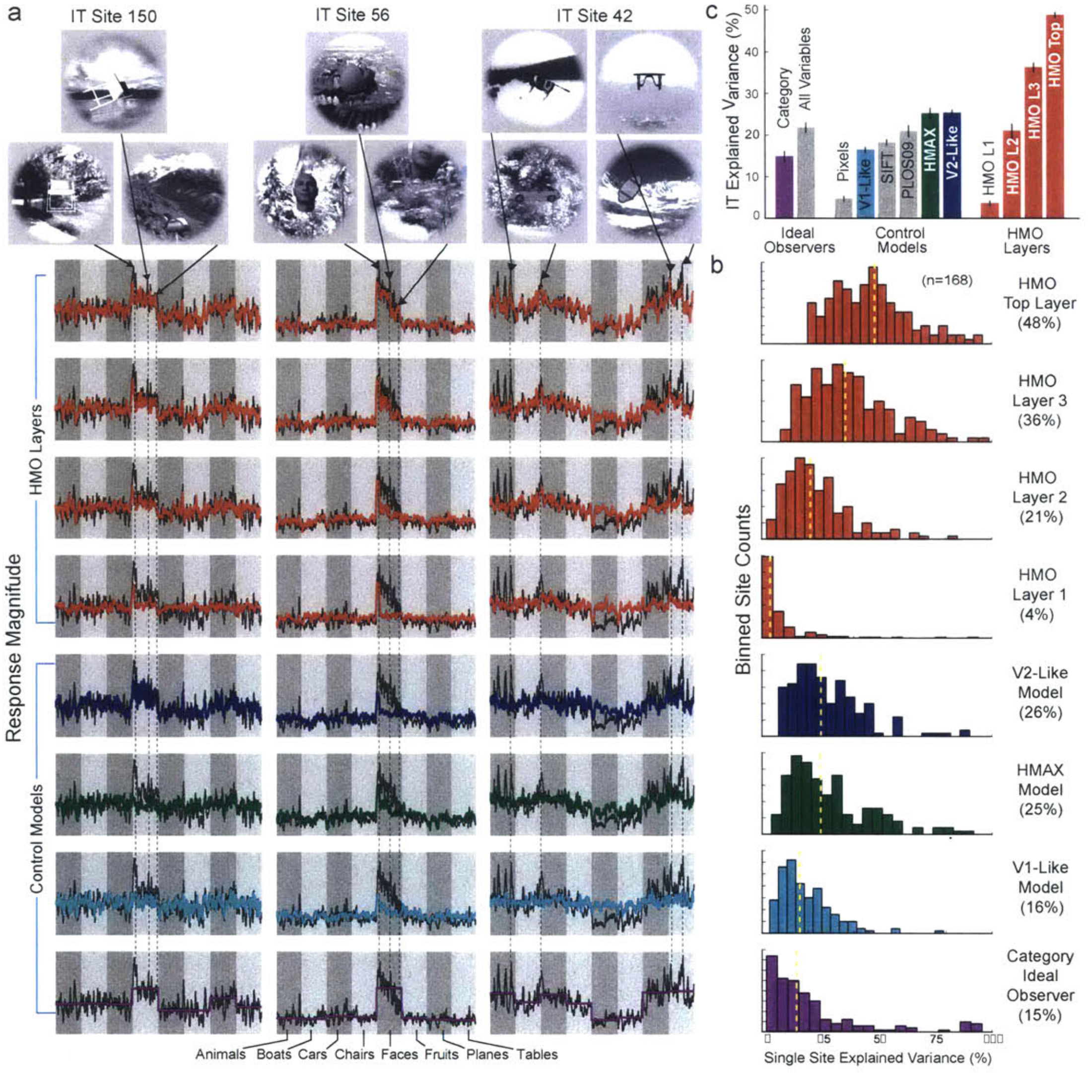

a
IT Site 150
IT Site 56
IT Site 42
Response Magnitude
HMO Layers
Control Models
Animals
Boats
Cars
Chairs
Faces
Fruits
Planes
Tables
b
(n=168)
HMO Top Layer (48%)
HMO Layer 3 (36%)
HMO Layer 2 (21%)
HMO Layer 1 (4%)
V2-Like Model (26%)
HMAX Model (25%)
V1-Like Model (16%)
Category Ideal Observer (15%)
Binned Site Counts
Single Site Explained Variance (%)
0
25
50
75
100
c
IT Explained Variance (%)
0
10
20
30
40
50
Category
All Variables
Pixels
V1-Like
SIFT
PLOS09
HMAX
V2-Like
HMO L1
HMO L2
HMO L3
HMO Top
Ideal Observers
Control Models
HMO Layers

Figure 4.3 *(preceding page)*: **IT Neural Predictions.** (***a***) Actual neural response (black trace) vs. model predictions (colored trace) for three individual IT neural sites. The x-axis in each plot shows 1600 test images sorted first by category identity and then by variation amount, with more drastic image transformations toward the right within each category block. The y-axis represents the prediction/response magnitude of the neural site for each test image (those not used to fit the model). Two of the units show selectivity for specific classes of objects, namely chairs (left) and faces (middle), while the third (right) exhibits a wider variety of image preferences. The four top rows show neural predictions using the visual feature set (i.e. units sampled) from each of the four layers of the HMO model, while the lower rows show the those of control models. (***b***) Distributions of model explained variance percentage, over the population of all measured IT sites (n=168). Yellow dotted line indicates distribution median. (***c***) Comparison of IT neural explained variance percentage for various models. Bar height shows median explained variance, taken over all predicted IT units. Error bars are computed over image splits. Colored bars are those shown in (a) and (b), while gray bars are additional comparisons (see text).

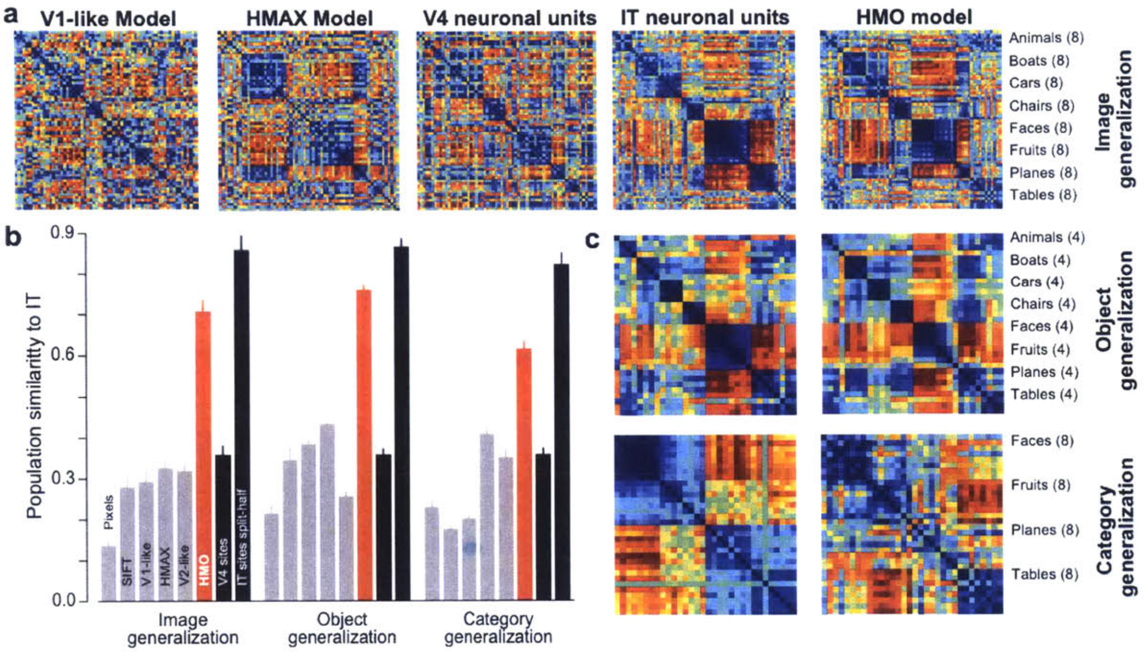


Figure 4.4: **Population-Level Similarity.** (*a*) Object-level Representation Dissimilarity Matrices (RDMs) visualized via rank-normalized color plots (blue=0th distance percentile, red=100th percentile). (*b*) IT population and the HMO-based IT model population, for image, object and category generalizations (see Methods). (*c*) Quantification of model population representation similarity to IT. Bar height indicates the spearman correlation value of a given model's RDM to the RDM for the IT neural population. The IT bar represents the Spearman-Brown corrected consistency of the IT RDM for split-halves over the IT units, establishing a noise-limited upper bound. Error bars are taken over cross-validated regression splits in the case of models and over image and unit splits in the case of neural data.

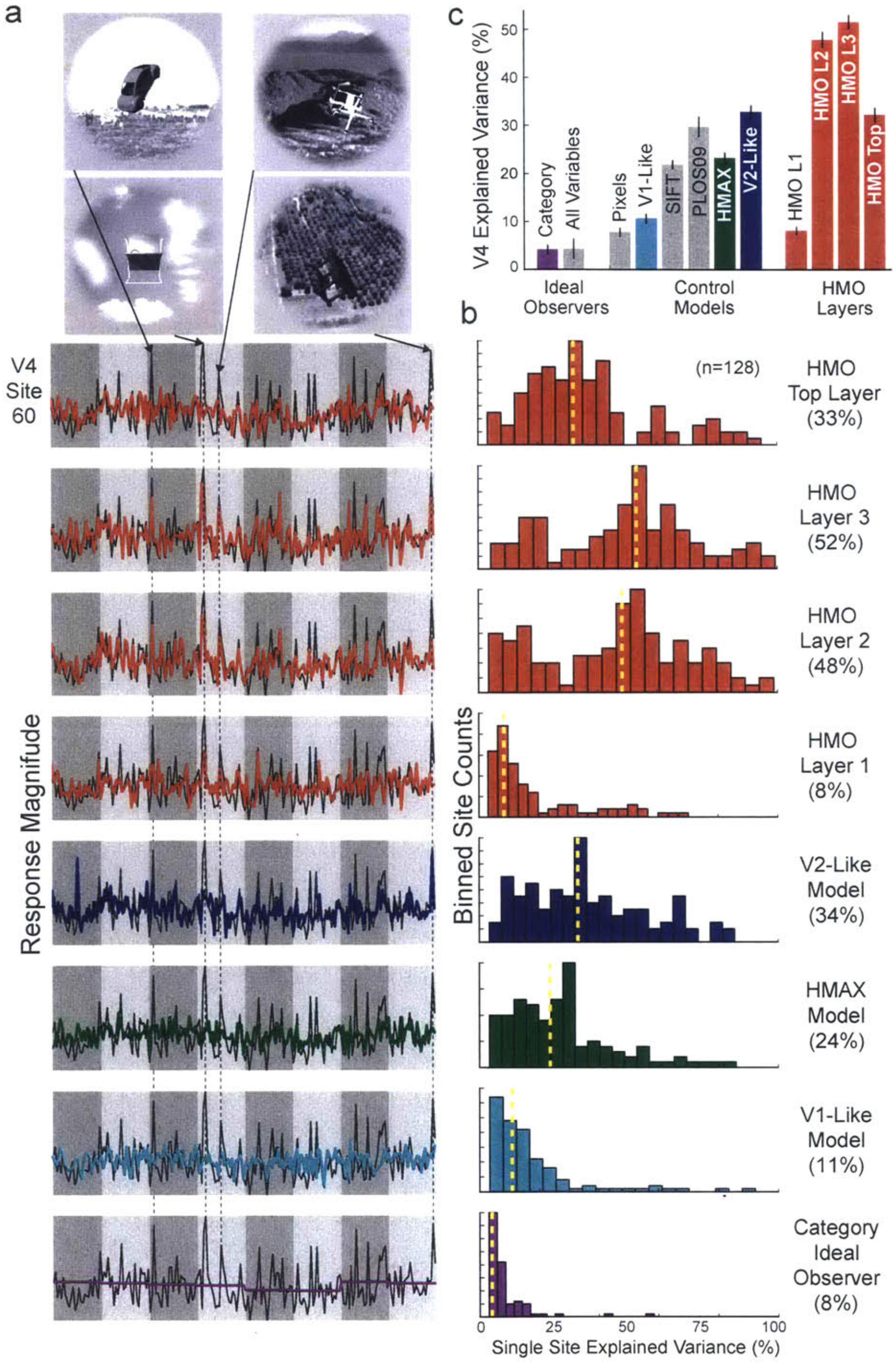


Figure 4.5: **V4 Neural Predictions.** (***a***) Actual vs. predicted response magnitudes for a typical V4 site. V4 sites are highly visually driven, but unlike IT sites show very little categorical preference, manifesting in more abrupt changes in the image-by-image plots shown here. Red highlight indicates the best-matching model (*viz.*, HMO layer 3). (***b***) Distributions of explained variances percentage for each model, over the population of all measured V4 sites ($n = 128$). (***c***) Comparison of V4 neural explained variance percentage for various models. Conventions follow those used in Fig. 4.3.

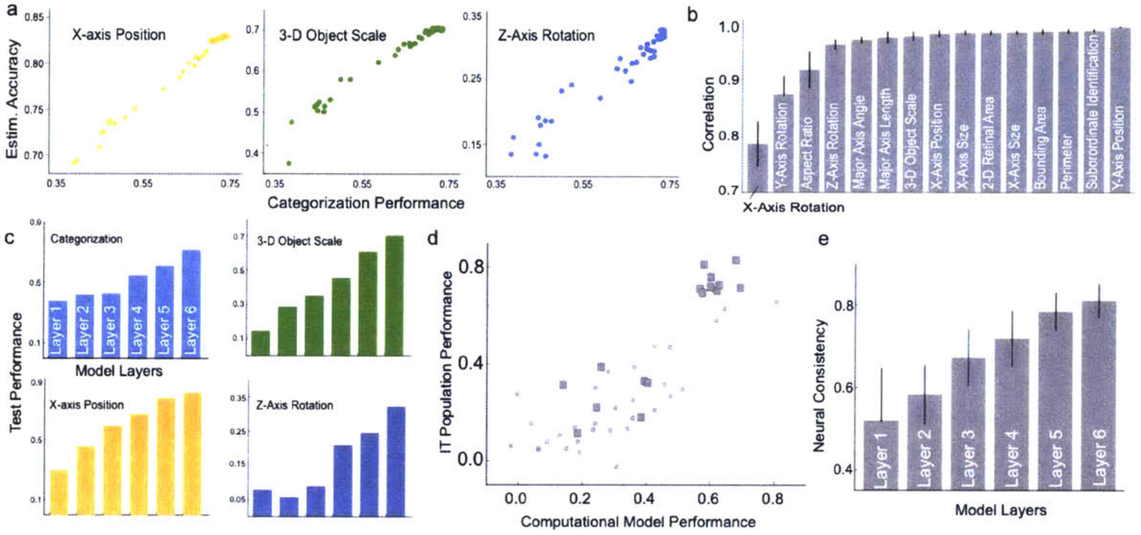


Figure 4.6: **Emergence of Non-Categorical Properties** (The neural network in this figure was further optimized with filter value fine tuning. See Section 4.2.7 and Figure 4.18 for details.) (***a***) Scatter plots of performance on categorization performance vs estimation accuracy for selected non-categorical tasks. Each dot represents a state of the model during training. (***b***) Quantification of relationship in panel a., taken over all tested tasks. Bar height represents Pearson correlation of accuracy on indicated task with test-set categorization performance, taken across training time steps. Error bars are taken across both time-steps as well as performance-assessment splits. (***c***) Performance of fully-trained model at multiple layers. y-axis is as in panel a. (***d***) Scatter plot of performance for top layer of fully-trained model on task battery vs neural performance on the tasks (see Fig. 3.5a). (***e***) Consistency of fully-trained model with neural performance pattern across layers, using the same metric as in 3.5b, bottom panel. y-axis and error bars are as in 3.5b.

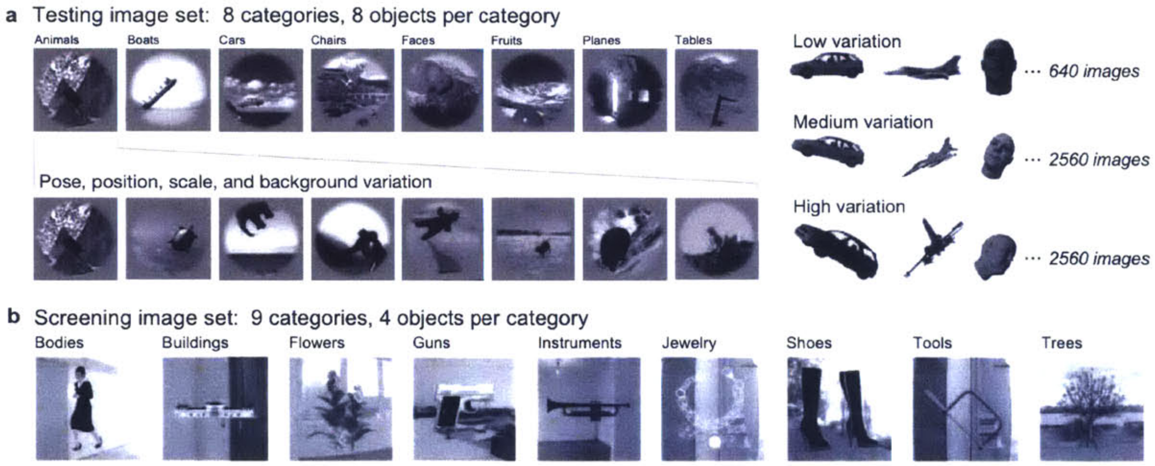


Figure 4.7: (***a***) The Neural Representation Benchmark [Cadieu et al., 2013] testing image set on which we collected neural data and evaluated models contained 5760 images of 64 objects in 8 categories. The image set contained three subsets, with low, medium and high levels of object view variation. Images were placed on realistic background scenes, which were chosen randomly to be uncorrelated with object category identity. (***b***) The screening image set used to discover the HMO model contained 4500 images of 36 objects in 9 categories. As with any two uncorrelated samples of images from the world — such as those images seen during development vs. those seen in adult life — the overall natural statistics of the screening set images were intended to be roughly similar to those of the testing set, but the specific content was quite different. Thus, the objects, semantic categories and background scenes used in screening were totally non-overlapping with those used in the testing set. Moreover, different camera, lighting and noise conditions, and a different rendering software package, were used.

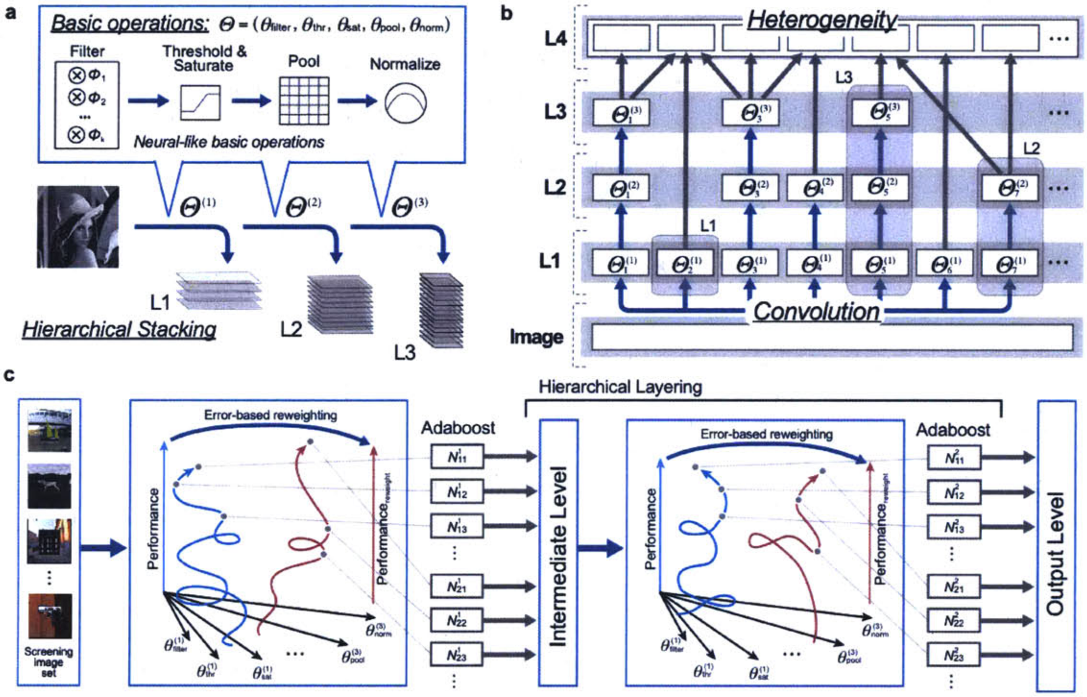
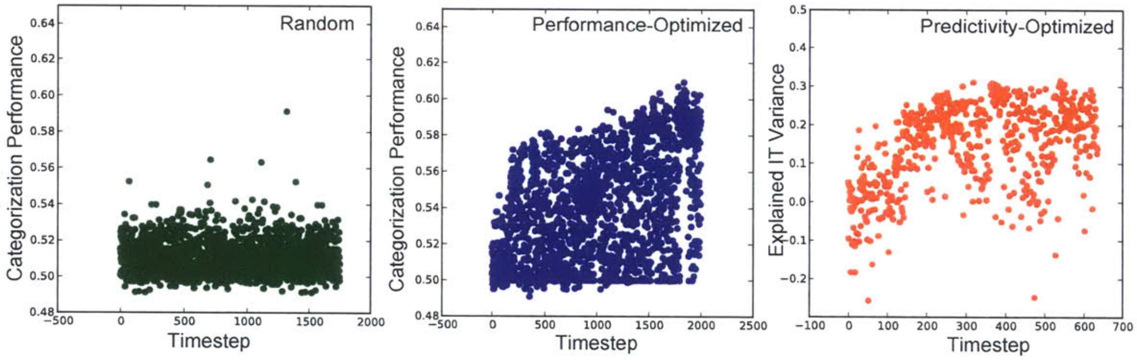


Figure 4.8 *(preceding page)*: In this work, we use Convolutional Neural Networks (CNNs) models. CNNs consist of a series of hierarchical layers, with bottom layers accepting inputs directly from image pixels, with units form the top and intermediate layers used to support training linear classifiers for performance evaluation and linear regressors for predicting neural tuning curves. (***a***) Following a line of existing work, we limited the constituent operations in each layer of the hierarchy to Linear-Nonlinear (LN) compositions including: (1) a filtering operation, implementing AND-like template matching; (2) a simple nonlinearity, e.g a threshold; (3) a local pooling/aggregation operation, such as softmax; and (4) a local competitive normalization. These layers are combined to produce low complexity (L1), intermediate complexity (L2) and high-complexity (L3) networks. All operations are repeated convolutionally at each spatial position, corresponding to the general retinotopic organization in the ventral stream. (***b***) In creating the HMO model, we allow mixture of several of these elements to model heterogenous neural populations, each acting convolutionally on the input image. The networks are structured in a manner consistent with known features of the ventral stream, as a series of areas of roughly equal complexity, but which permit bypass projections. (***c***) Hierarchical Modular Optimization (HMO) is a procedure for searching the space of CNN mixtures to maximize object recognition performance. With several rounds of optimization, HMO creates mixtures of component modules that specialize in subtasks, without needing to prespecify what these subtasks should be. Errors from earlier rounds of optimization are analyzed, and used to reweight subsequent optimization toward unsolved portions of the problem. The complementary component modules that emerge via this process are then combined and used as input to repeat the procedure hierarchically (see Methods).



Random

Performance-Optimized

Predictivity-Optimized

Figure 4.9: Optimization time traces for the high-throughput experiments shown in Fig. 4.1. In the performance and fitting-optimized the y-axis shows the optimization criterion — in the random selection case (left panel), no optimization was done, and the performance data was ignored.

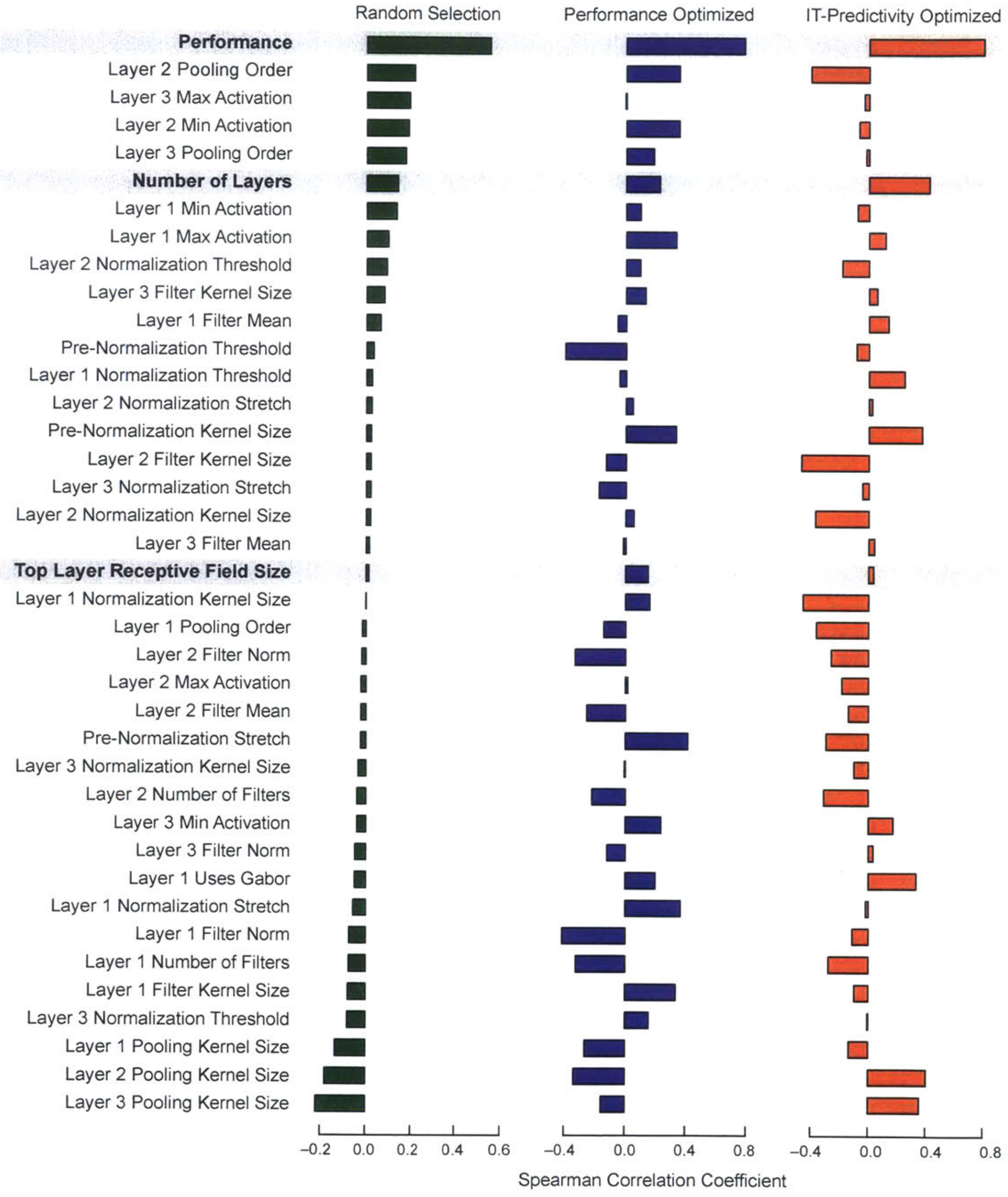

Random Selection
Performance Optimized
IT-Predictivity Optimized
Performance
Layer 2 Pooling Order
Layer 3 Max Activation
Layer 2 Min Activation
Layer 3 Pooling Order
Number of Layers
Layer 1 Min Activation
Layer 1 Max Activation
Layer 2 Normalization Threshold
Layer 3 Filter Kernel Size
Layer 1 Filter Mean
Pre-Normalization Threshold
Layer 1 Normalization Threshold
Layer 2 Normalization Stretch
Pre-Normalization Kernel Size
Layer 2 Filter Kernel Size
Layer 3 Normalization Stretch
Layer 2 Normalization Kernel Size
Layer 3 Filter Mean
Top Layer Receptive Field Size
Layer 1 Normalization Kernel Size
Layer 1 Pooling Order
Layer 2 Filter Norm
Layer 2 Max Activation
Layer 2 Filter Mean
Pre-Normalization Stretch
Layer 3 Normalization Kernel Size
Layer 2 Number of Filters
Layer 3 Min Activation
Layer 3 Filter Norm
Layer 1 Uses Gabor
Layer 1 Normalization Stretch
Layer 1 Filter Norm
Layer 1 Number of Filters
Layer 1 Filter Kernel Size
Layer 3 Normalization Threshold
Layer 1 Pooling Kernel Size
Layer 2 Pooling Kernel Size
Layer 3 Pooling Kernel Size
−0.2
0.0
0.2
0.4
0.6
−0.4
0.0
0.4
0.8
−0.4
0.0
0.4
0.8
Spearman Correlation Coefficient

Figure 4.10 *(preceding page)*: Correlation of model parameters with IT predictivity for the three high-throughput experiments shown in figure 4.1. Parameters for which the correlation is significantly different from 0 are shown. Also included are several additional metrics that are not direct model parameters but that represent measurable quantities of interest for each model, e.g. model object recognition performance. x-axis is spearman-r correlation of the given parameter with IT-predictivity for the indicated model selection procedure, including random (left, green bars), performance-optimized (middle, blue bars), and IT-predictivity optimized (right, red bars). Parameters are ordered by correlation value for the random condition. Performance strongly correlates with IT predictivity in all selection regimes. while Number of layers (model depth) consistently correlates as well, but much more weakly. Interestingly, one "obvious" metric — receptive field size at the top model layer — is only very weakly associated with predictivity, because while the best models tended to have larger receptive field sizes, a large number of poor models also shared this characteristic.

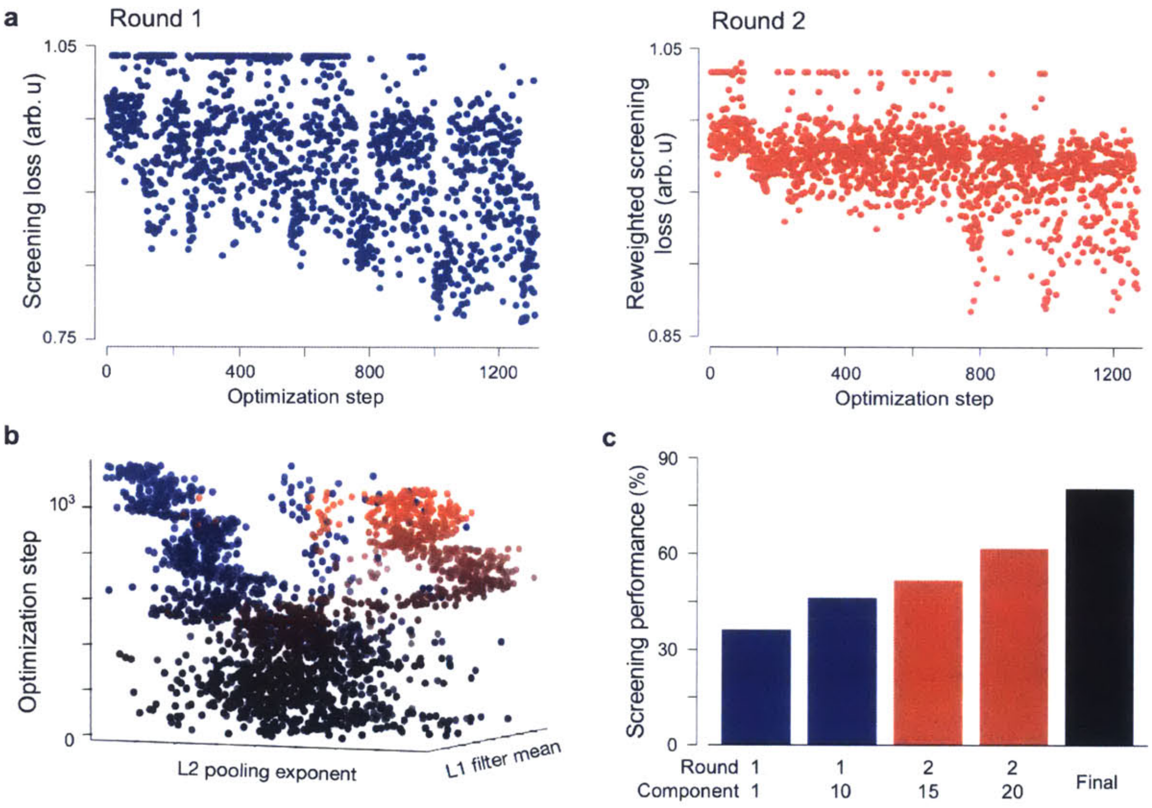


Figure 4.11: A) Optimization loss traces during the HMO procedure showing decreased loss as optimization proceeded. B) Parameter-space trajectories during two optimization rounds shown in panel A (Round 1 are blue dots, Round 2 are red dots). This 3-d plot shows parameter values for two chosen parameters (L1 filter mean and L2 pooling exponent) out of many, but it is evident that subsequent rounds of optimization (e.g. red) gravitate toward different parameter combinations (i.e. different network architectures) than earlier rounds of optimization (e.g. blue). C) Training performance as a function of model complexity, showing dramatic increases as components from Round 1 (blue bars) and Round 2 (red bars) were added. The final model (black bar) consists of 30 components identified with three complementary rounds of optimization, plus one L1 layer that, anatomically, stacks on top of those 30 components and, functionally, produces non-linear combinations of their outputs.

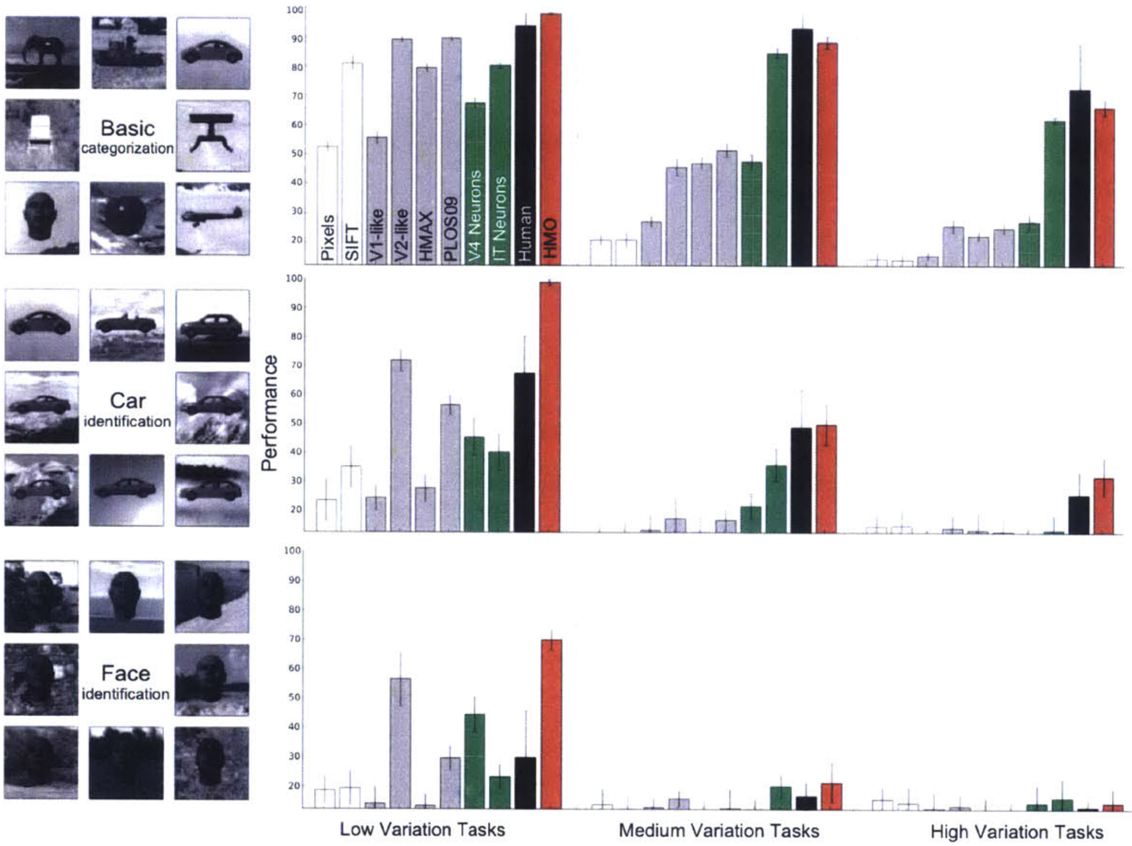


Figure 4.12: Classification results for tasks including basic 8-way basic categorization (top row) 8-way subordinate car identification (middle row), and 8-way subordinate face identification (bottom row). Each task was assessed at low, medium, and high levels of image variation (see Methods). Comparison was made between neural data, human data, existing models from the literature, and the Hierarchical Modular Optimization model outputs. The tasks span a wide range of difficulty, from low-variation basic 8-way categorization where humans perform at greater than 95% accuracy, to high-variation subordinate face identification, where human performance is indistinguishable from chance.

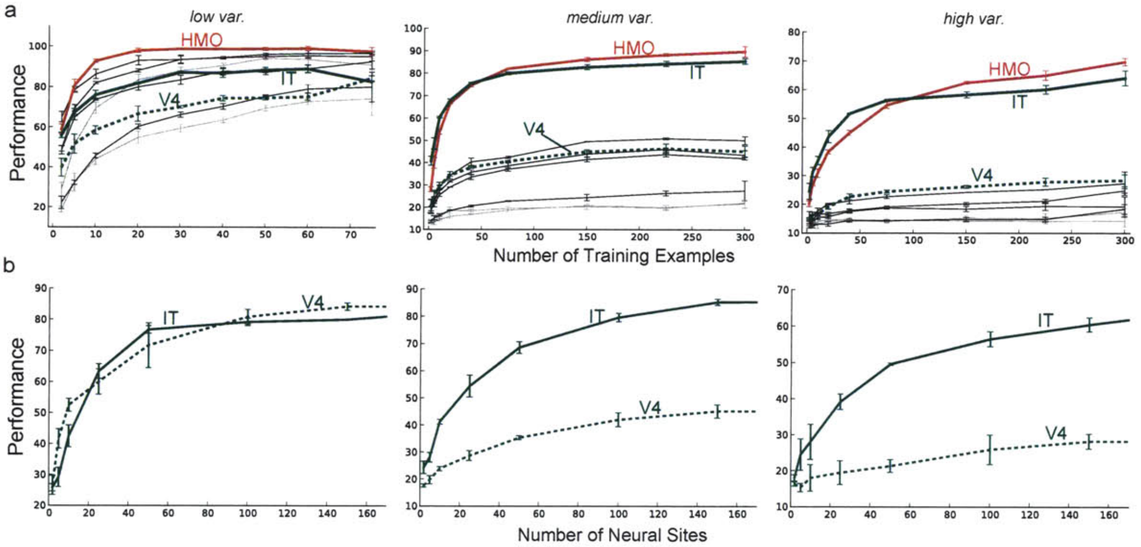


Figure 4.13: a) Dependence of performance on number of training examples for models and neural populations. HMO model is shown in red; IT population in solid green; V4 in dotted green; all other control models are shown in black. b) Direct comparison of dependence of performance on number of neural sites, for the IT (solid green) and V4 (dotted green) neural populations.

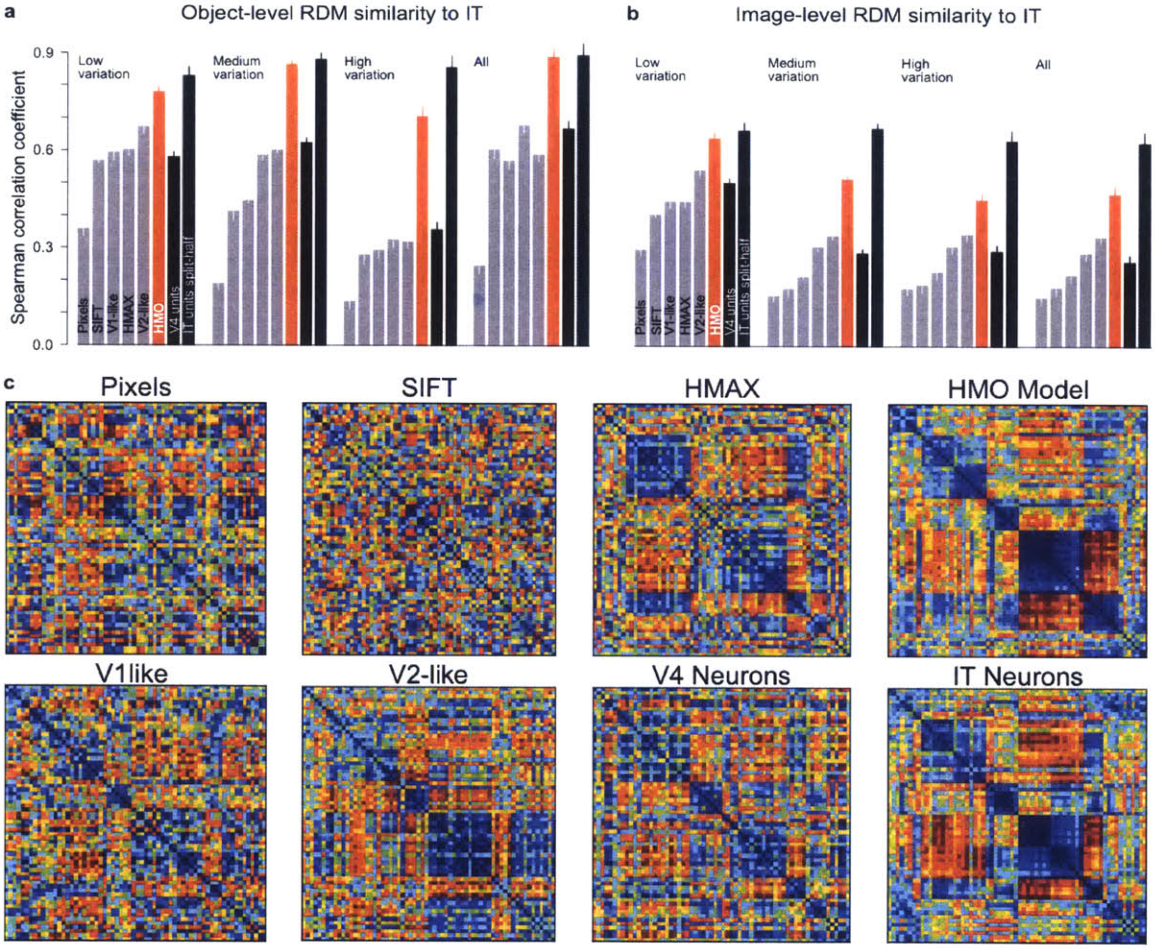


Figure 4.14: Additional RDM comparisons to IT population structure. As in Figure 4.3d-e, each bar shows the Spearman correlation of an RDM for a model (or V4 population) with the RDM for the IT neural population on the same stimulus set. We show comparisons for three subsets of the test image set separated by variation level ("Low", "Medium" and "High"), as well as for the whole stimulus set ("All"). Panel a) shows comparisons of RDMs at the object level, in which population representation vectors are averaged on a per-object basis before taking the pairwise correlations to make the RDM matrices. Panel b) shows more detailed image-level RDMs comparisons, with each stimulus represented separately. c) Object-level RDMs for a variety of models and the V4 and IT neural populations.

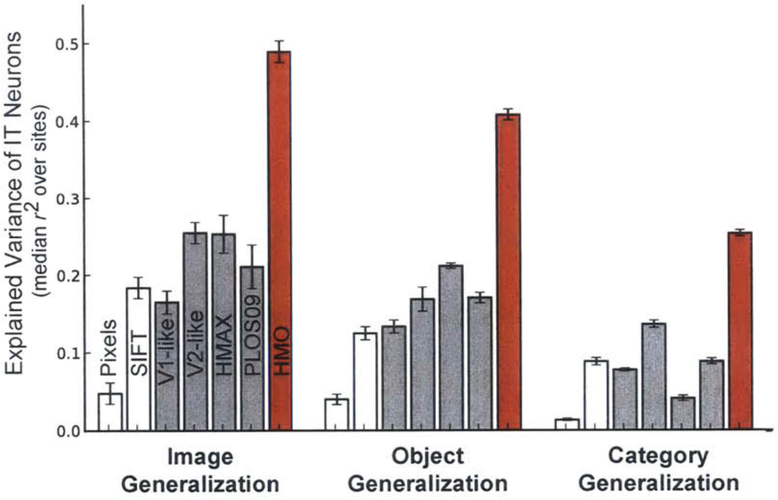


Figure 4.15: IT explained variance for each model, fit with training/test image splits generated by (1) **image generalization**, a random selection process in which train and test splits contain images of the same 64 objects, but on different backgrounds and at widely different poses, positions, and sizes; (2) **object generalization**, in which train and test images are split so that they contain no overlapping objects, so that predictions are tested for generalization across object identity as well as position, pose, size and background variation; and (3) **category generalization**, in which train an test images are split so that they contain no overlapping categories, so that predictions are tested across category boundaries as well. Figure 4.3e shows the corresponding results at the population RDM level.

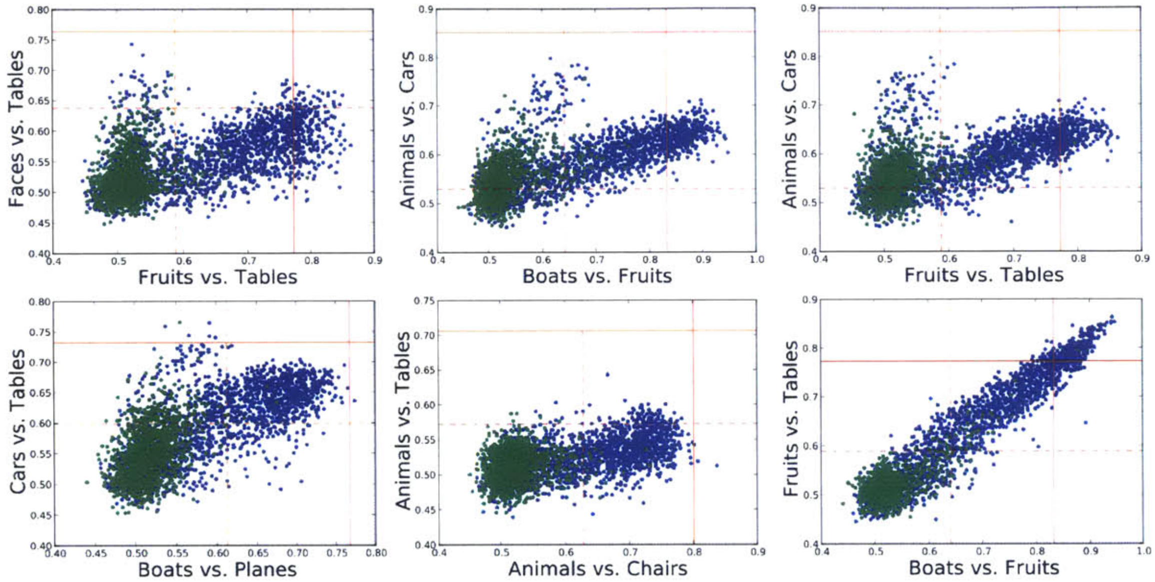


Figure 4.16: **Tradeoffs between subtask-optimal architectures**. Each panel shows pairwise relative performance of the models from the high-throughput experiments in Fig. 4.9a on a variety of binary subtasks. As in that figure, random selections are shown in green and performance-optimized selections are shown in blue. Sometimes performance on one binary subtask — e.g. Boats-vs-Fruits and Fruits-vs-Tables (lower-right-hand corner panel) — directly correlates with performance on another. More commonly, there is a tradeoff between subtask performance in the models explored during optimization, leading to the "V" pattern observed in subtask pairs. Because the procedure was maximizing overall performance (as opposed to performance on any one subtask), one "arm" of the V is heavier than the other, corresponding to the optimization process being forced to make a single "choice" in each of these tradeoffs.

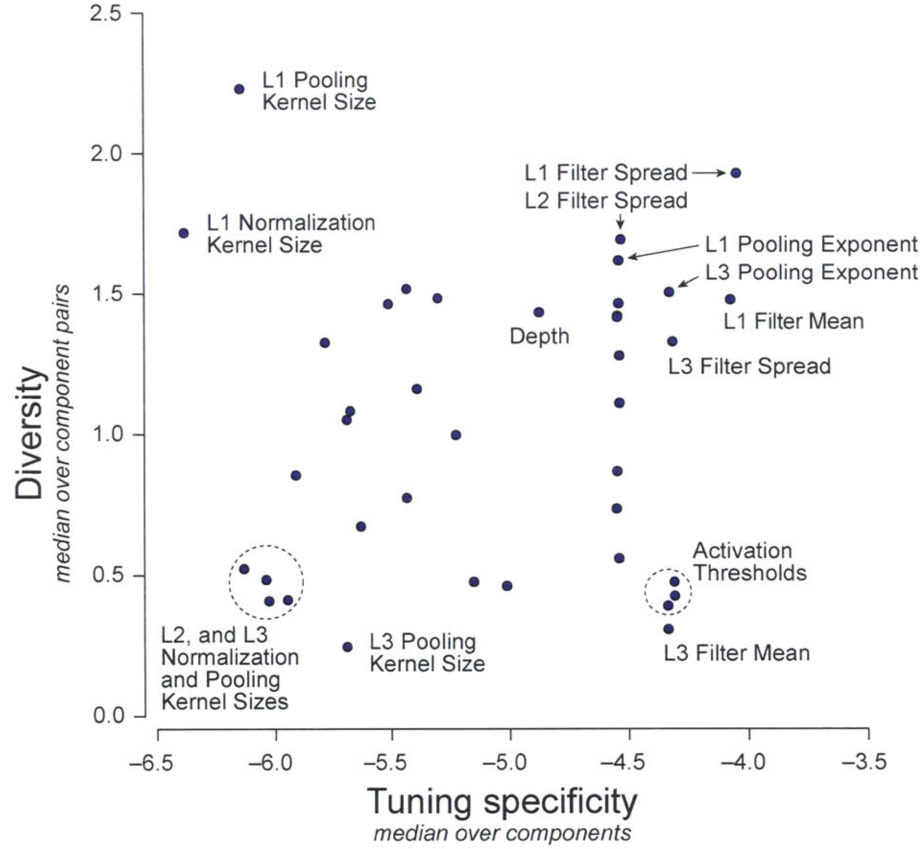


Figure 4.17: Characterization of selected model parameters in terms of per-component tuning specificity versus inter-component diversity. Each point in this plot represents an architectural parameter in the HMO model. Parameters in the upper right corner are highly tuned but also highly diverse in their tunings between model components. See text for the definition of diversity and tuning specificity.

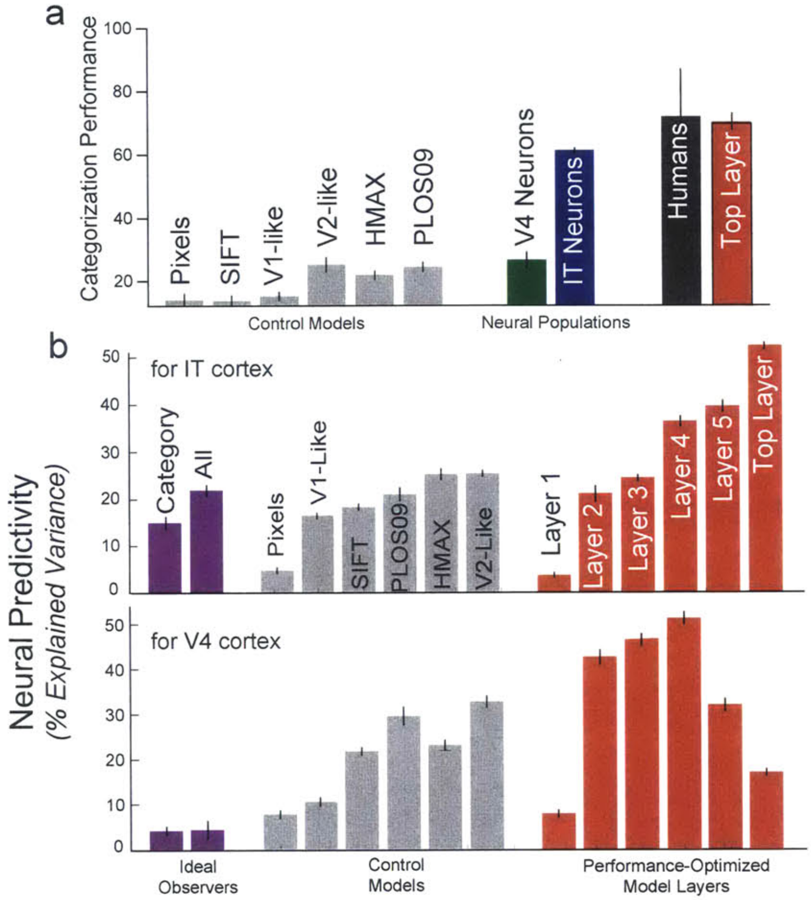


Figure 4.18: **Results of models with fine tuning.** (***a***) Similar to our "standard" convolutional neural networks (termed as HMO models, which stand for Hierarchical Modular Optimization models), the top layer of the filter value tuned, performance-optimized model generalizes from the real photograph training set (red bar), and significantly outperforms control models (gray bars) and the V4 neural population on the 8-way object categorization task (Animals vs Boats vs Cars vs Chairs vs Faces vs Fruits vs Planes vs Tables) in the images shown in Fig. 3.7a. Model performance is comparable to IT neural population (blue bar) and human performance measured via psychophysical experiments (black bar). (***b***) The performance-optimized model is then used predict neural response in IT cortex (top panel) and V4 cortex (bottom bars). Ability to predict IT neural patterns is better with each subsequent model layer, peaking at the top layer (red bars), whereas ability to predict V4 neurons peaks in the middle layers. For both V4 and IT, the performance-optimized model's most predictive layer is significantly better than other control models, including ideal observers that perform perfectly on categorization tasks (purple bars) as well as control models that are also in the general class of neural networks (gray bars).

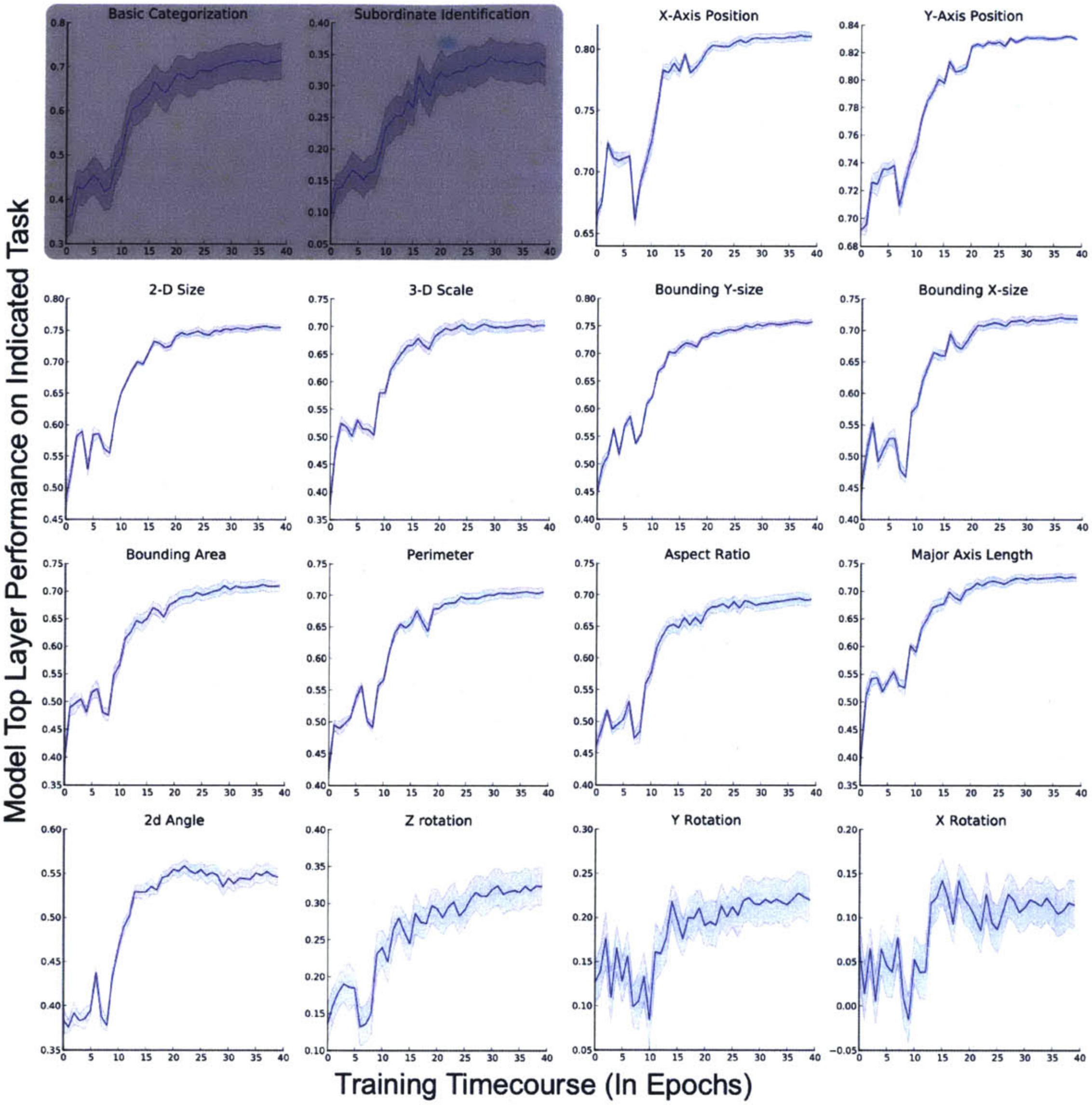


Figure 4.19: **Test set performance timecourses: fine tuned model.** y-axis represents performance of linear regressors and classifiers trained on the top level of the model, on each task defined on the testing set (see Fig 3.7a). x-axis represents timepoints taken during training for categorization on the ImageNet dataset (as described in Methods). Performance was estimated by building top-level regularized classifiers and regressors (as described in the methods text) separately at each time step. Note that the x-axis is the same for all panels, representing the same training trajectory; the various y-axis panels are all based on the single feature set produced by the categorization training. The first two panels, with gray backgrounds, indicate categorical tasks (8-way basic categorization and subordinate category identifications); the remaining white-background panels indicate non-categorical tasks.

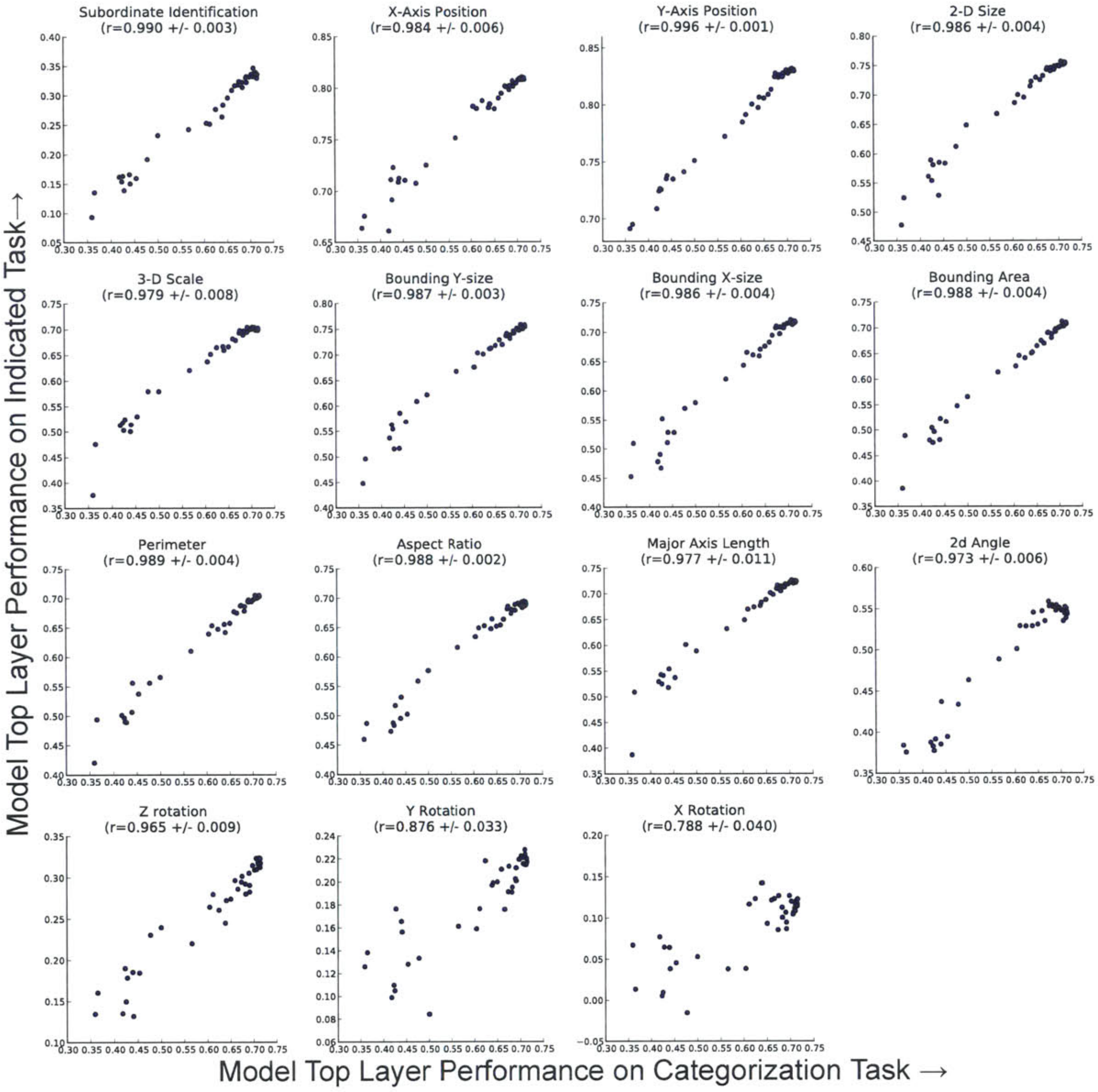


Figure 4.20: **Categorization performance Vs non-categorical task performance: fine tuned model** y-axes in each panel are same as in Fig. 4.19. x-axis is performance on test set categorization task (e.g., the y-axis of the upper-left-most panel in Fig. 4.19). Each dot represents a distinct timepoint as shown on the x axis in Fig. 4.19.

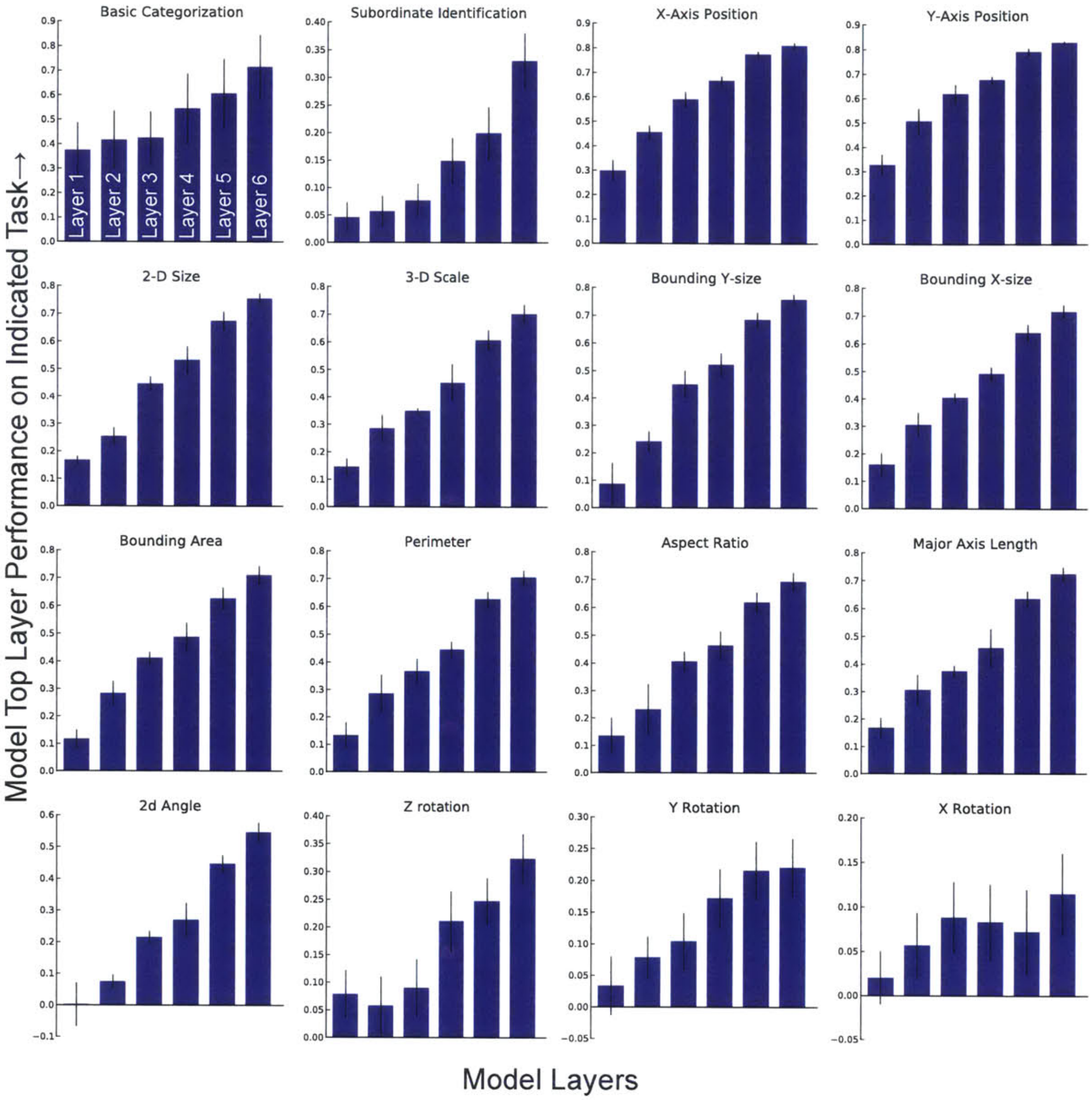


Figure 4.21: **Performance as a function of model layer: fined tuned model** x and y axes in this figure as identical to main text Fig. 4.6c, but showing results for all tasks.

Chapter 5

Conclusions

As a whole, this thesis provides quantitative insights into how the brain, in particular the ventral visual stream, might perform the challenging tasks of core visual perception of objects[1].

In Chapters 2 and 3, we demonstrate that simple, learned weighted sums of randomly-selected average responses of distributed IT neurons robustly predict the human pattern of behavioral performance across a wide range of visual tasks, not only categorization and identification tasks, but also other non-categorical tasks, such as position, size, and viewpoint estimation tasks. This suggests that the ventral visual stream (and in particular, IT cortex), which builds tolerance to identity-preserving transforms, also builds an explicit representation that is sensitive to exactly those properties to which the system is also tolerant. More precisely, the data argue that a simple rate code (integrated between 70 to 170ms post stimulus onset), read out on single-trials, learned from a distributed population of ~60 thousand single IT units,

[1] Defined as the ability to rapidly (<200ms viewing duration) estimating key visual parameters — including object category, identity, size, pose, perimeter, and aspect ratio — of the object presented parafoveally (central $\sim 10^\circ$ of the visual field). See Section 1.3 for details.

reliably explains both the pattern and the magnitude of human performance over a large battery of core object perception tasks.

A fundamental contribution made here is our *quantitative* framework for comparing neural responses to behavior by utilizing a battery of visual tasks that measure the range of human subjects' capabilities in the domain of core object perception. This operational definition of object perception provides a strong *consistency test* by which we could quantitatively evaluate and falsify different neuronal linking hypotheses that might explain behavior. One important advantage of this quantitative framework is that we can iteratively refine our choices of images, tasks, neural recording, and linking hypotheses to continue to deepen our understanding of the link between neurons and high level object perception. A natural extension to our approach would be to study the effects of crowding, clutter, occlusion, and correlated backgrounds. In addition, studying at a more detailed level (e.g., per-image behavioral pattern) would uncover important details in visual object perception.

In Chapters 2 and 3, combined with the modeling results in Chapter 4, we show that, along the hierarchy of the ventral visual stream, the representation explicitness to non-categorical, identity-preserving transformation is not merely retained but rather *increased* in concert with the transformation tolerance, for common objects presented in a visually rich environment with cluttered background. In other words, for photo-realistic images of objects presented parafoveally, IT encodes both the categorical *and* non-categorical parameters of the object *more robustly* than other lower level visual areas do in a way that the IT encoding is readily parsed with simple weighted linear sums. It may be unintuitive that some properties (in particular, position and size estimation) that are typically thought of as low-level visual features are actually more effectively captured in IT neural populations. Our results are nonetheless consistent with the prior studies that formed existing intuitions, in which simple stimuli (e.g., bars and gratings on a gray screen) were mostly employed, whereas our results focus on complex stimuli containing realistic objects on cluttered backgrounds. It should be noted that the stimulus set used in this study is not large enough to show, nor do we mean even to suggest, that the ventral stream builds up an ability to estimate properties of objects presented in the periphery, a function normally associated with the dorsal stream. Instead, we speculate that both the dorsal and ventral stream contain representations for categorical *and* non-categorical visual properties, but at different levels of spatial resolution and scale, the ventral being parafoveally localized fine-scale and the dorsal being large-scale with peripheral coverage. This idea is attractive as it would naturally support behavior in which the dorsal machinery directs foveation based on, for example, an environmental saliency map, while the ventral machinery parses details in each foveation snapshot, to produce an overall scene understanding. Experimental testing of this idea, especially with complex naturalistic stimulus set, would be a rewarding avenue of research.

In Chapter 4, we build a quantitatively accurate computational model of the ventral visual stream by optimizing the classification performance of bio-inspired hierarchical neural networks. Surprisingly, while the model has never been given any neural data to match directly, its top layer shows a remarkable ability to predict IT neural responses to realistic images at both the single site and population levels. Moreover, the model's second top layer turns out to be highly predictive of neural responses in V4, the main cortical input to IT. These results suggest that imposing performance optimization, a behaviorally relevant goal, combined with a sufficiently large set of biologically plausible models, effectively yields quantitatively predictive models of neural processing in the visual system. In addition, as in our experimental observations, the optimized model also shows emerging explicit representation for both categorical and non-categorical properties along its hierarchical layers, even though it has not been directly optimized for it as such. Instead, we find that simple-minded optimization for robust categorization performance brings along performance on all the other non-categorical tasks, in addition to the neural response prediction tasks, "for free." This suggests a series of interesting follow-up studies investigating whether the converse is true — is solving for a non-categorical property (e.g., object position estimation) enough to guarantee categorization performance, or is categorization a much stronger constraint driving the "development" of IT neural responses? In addition, the scale of the space to be search is enormous, even before adding in more sophisticated mechanisms, such as attention and inter-area feedback that are known to exist in nature. In this scenario, it often becomes unclear which models are promising leads that should be followed up more carefully and which are high-performing, but biologically-implausible "dead-ends". While we have not specifically utilized, we argue that biological data could be leveraged to guide and accelerate our search, operating under the hypothesis that models that more closely approximate biological systems are more likely to be on the "right" path, both in terms of machine perception performance, and utility for neuroscience understanding.

Mechanistic modeling of the ventral stream's algorithms is a genuinely challenging problem, and we do not claim that we single-handedly have solved it in this work. Instead, we view this as a demonstration of a robust research approach that can be refined to deepen our knowledge. This work provides comprehensive human benchmarks, makes quantitative perceptual predictions, and establishes a foundation of mechanistic models of human object perception.

In conclusion, this thesis is focused on two related questions in visual neuroscience and machine perception: understanding how different patterns of neural activity give rise to specific human object perception behaviors (the brain-to-behavior link); and developing high-performing computer vision models that in turn predict this neural activity from input images (the image-to-brain link). This work provides answers to these two questions, which is an end-to-end understanding of object perception in the human visual system as a full pipeline from images to behavior. By understanding vision in the brain better, I believe we will be able to discover more effective computer vision and machine perception algorithms, and conversely, such improved algorithms will allow us to gain further insight into how the brain works.

References

S. R. Afraz, R. Kiani, and H. Esteky. Microstimulation of inferotemporal cortex influences face categorization. *Nature*, 442(7103):692–5, 2006.

B. B. Averbeck, P. E. Latham, and A. Pouget. Neural correlations, population coding and computation. *Nature Reviews Neuroscience*, 7(5):358–366, 2006.

Y. Bengio. *Learning Deep Architectures for AI.* Now Publishers, 2009.

J. Bergstra, D. Yamins, and D. Cox. Making a Science of Model Search, 2012. URL `http://arxiv.org/pdf/1209.5111v1.pdf`.

J. Bergstra, D. Yamins, and D. Cox. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *Proceedings of The 30th International Conference on Machine Learning*, pages 115–123, 2013.

I. Biederman and P. C. Gerhardstein. Recognizing depth-rotated objects: evidence and conditions for three-dimensional viewpoint invariance. *Journal of Experimental Psychology: Human Perception and Performance*, 19(6):1162, 1993.

I. Biederman, P. C. Gerhardstein, E. E. Cooper, and C. A. Nelson. High level object recognition without an anterior inferior temporal lobe. *Neuropsychologia*, 35(3): 271–287, 1997.

W. H. Bosking, J. C. Crowley, and D. Fitzpatrick. Spatial coding of position and orientation in primary visual cortex. *Nature Neuroscience*, 5(9):874–882, 2002.

A. A. Brewer, W. A. Press, N. K. Logothetis, and B. A. Wandell. Visual areas in macaque cortex measured using functional magnetic resonance imaging. *The Journal of neuroscience*, 22(23):10416–10426, 2002.

S. L. Brincat and C. E. Connor. Underlying principles of visual shape selectivity in posterior inferotemporal cortex. *Nature Neuroscience*, 7(8):880–6, 2004.

K. H. Britten, M. N. Shadlen, W. T. Newsome, and J. A. Movshon. The analysis of visual motion: a comparison of neuronal and psychophysical performance. *Journal of Neuroscience*, 12(12):4745–65, 1992.

K. H. Britten, W. T. Newsome, M. N. Shadlen, S. Celebrini, and J. A. Movshon. A relationship between behavioral choice and the visual responses of neurons in macaque mt. *Visual Neuroscience*, 13(1):87–100, 1996.

L. E. Brown, B. A. Halpert, and M. A. Goodale. Peripheral vision for perception and action. *Experimental Brain Research*, 165(1):97–106, 2005.

I. Buciu and I. Pitas. ICA and Gabor representation for facial expression recognition. *Image Processing*, 2003.

C. Cadieu, M. Kouh, A. Pasupathy, C. E. Connor, M. Riesenhuber, and T. Poggio. A model of v4 shape selectivity and invariance. *Journal of Neurophysiology*, 98(3): 1733–50, 2007.

C. Cadieu, H. Hong, D. Yamins, N. Pinto, N. Majaj, and J. DiCarlo. The neural representation benchmark and its evaluation on brain and machine. In *International Conference on Learning Representations*, May 2013.

C. F. Cadieu, H. Hong, D. L. Yamins, N. Pinto, D. Ardila, E. A. Solomon, N. J. Majaj, and J. J. DiCarlo. Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS Computational Biology*, 10(12): e1003963, 2014.

M. Carandini, J. B. Demb, V. Mante, D. J. Tolhurst, Y. Dan, B. A. Olshausen, J. L. Gallant, and N. C. Rust. Do We Know What the Early Visual System Does? *Journal of Neuroscience*, 25(46):10577–10597, Nov. 2005a.

M. Carandini, J. B. Demb, V. Mante, D. J. Tolhurst, Y. Dan, B. A. Olshausen, J. L. Gallant, and N. C. Rust. Do we know what the early visual system does? *Journal of Neuroscience*, 25(46):10577–97, 2005b.

C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

M. Chelaru and V. Dragoi. Efficient coding in heterogeneous neuronal populations. *Proceedings of the National Academy of Sciences of the United States of America*, 105(42):16344–16349, 2008.

C. A. Chestek, V. Gilja, P. Nuyujukian, J. D. Foster, J. M. Fan, M. T. Kaufman, M. M. Churchland, Z. Rivera-Alvidrez, J. P. Cunningham, S. I. Ryu, et al. Long-term stability of neural prosthetic control signals from silicon cortical arrays in rhesus macaque motor cortex. *Journal of Neural Engineering*, 8(4):045005, 2011.

S. Chikkerur, T. Serre, C. Tan, and T. Poggio. What and where: A Bayesian inference theory of attention. *Vision Research*, 50(22):2233–2247, 2010.

M. R. Cohen and J. H. Maunsell. Attention improves performance primarily by reducing interneuronal correlations. *Nature Neuroscience*, 12(12):1594–1600, 2009.

M. R. Cohen and J. H. Maunsell. A neuronal population measure of attention predicts behavioral performance on individual trials. *The Journal of Neuroscience*, 30(45): 15241–15253, 2010.

M. R. Cohen and J. H. Maunsell. Using neuronal populations to study the mechanisms underlying spatial and feature attention. *Neuron*, 70(6):1192–1204, 2011.

C. E. Collins, D. C. Airey, N. A. Young, D. B. Leitch, and J. H. Kaas. Neuron densities vary across and within cortical areas in primates. *Proceedings of the National Academy of Sciences*, 107(36):15927–15932, 2010.

C. Connor, S. Hsiao, J. Phillips, and K. Johnson. Tactile roughness: neural codes that account for psychophysical magnitude estimates. *The Journal of Neuroscience*, 10 (12):3823–3836, 1990.

C. E. Connor, S. L. Brincat, and A. Pasupathy. Transformation of shape information in the ventral pathway. *Current Opinion in Neurobiology*, 17(2):140–7, 2007.

S. V. David, B. Y. Hayden, and J. L. Gallant. Spectral receptive field properties explain shape selectivity in area v4. *Journal of Neurophysiology*, 96(6):3492–505, 2006.

G. Deco and E. T. Rolls. A neurodynamical cortical model of visual attention and invariant object recognition. *Vision Research*, 44(6):621–642, 2004.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

R. Desimone, T. D. Albright, C. G. Gross, and C. Bruce. Stimulus-selective properties of inferior temporal neurons in the macaque. *Journal of Neuroscience*, 4(8): 2051–62, 1984a.

R. Desimone, T. D. Albright, C. G. Gross, and C. Bruce. Stimulus-selective properties of inferior temporal neurons in the macaque. *The Journal of Neuroscience*, 4 (8):2051–2062, 1984b.

R. DeValois. Spatial processing of luminance and color information. *Investigative ophthalmology & visual science*, 17(9):834–835, 1978.

R. L. DeValois. Behavioral and electrophysiological studies of primate vision. *Contributions to sensory physiology*, 14:137, 1965.

J. J. DiCarlo and D. D. Cox. Untangling invariant object recognition. *Trends in Cognitive Sciences*, 11(8):333–341, 2007.

J. J. DiCarlo and J. H. Maunsell. Anterior inferotemporal neurons of monkeys engaged in object recognition can be highly sensitive to object retinal position. *Journal of Neurophysiology*, 89(6):3264–3278, 2003.

J. J. DiCarlo and J. H. R. Maunsell. Form representation in monkey inferotemporal cortex is virtually unaltered by free viewing. *Nat Neurosci*, 3(8):814–821, 2000a.

J. J. DiCarlo and J. H. R. Maunsell. Inferotemporal representations underlying object recognition in the free viewing monkey. In *Society for Neuroscience*, New Orleans, 2000b.

J. J. DiCarlo, D. Zoccolan, and N. C. Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–34, 2012.

P. E. Downing, A. Chan, M. Peelen, C. Dodds, and N. Kanwisher. Domain specificity in visual cortex. *Cerebral Cortex*, 16(10):1453–1461, 2006.

S. Edelman. *Representation and Recognition in Vision.* MIT Press, Cambridge, MA, 1999.

A. K. Engel, P. Fries, and W. Singer. Dynamic predictions: oscillations and synchrony in top–down processing. *Nature Reviews Neuroscience*, 2(10):704–716, 2001.

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

D. J. Freedman, M. Riesenhuber, T. Poggio, and E. K. Miller. Categorical representation of visual stimuli in the primate prefrontal cortex. *Science*, 291(5502): 312–316, 2001.

D. J. Freedman, M. Riesenhuber, T. Poggio, and E. K. Miller. A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *The Journal of Neuroscience*, 23(12):5235–5246, 2003.

J. Freeman and E. Simoncelli. Metamers of the ventral stream. *Nature Neuroscience*, 14(9):1195–1201, 2011.

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

W. A. Freiwald and D. Y. Tsao. Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science*, 330(6005): 845–51, 2010.

B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, Feb. 2007a. doi: 10.1126/science.1136800.

B. J. Frey and D. Dueck. Clustering by passing messages between data points. *science*, 315(5814):972–976, 2007b.

K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 1980.

J. Gallant, C. Connor, S. Rakshit, J. Lewis, and D. Van Essen. Neural responses to polar, hyperbolic, and cartesian gratings in area v4 of the macaque monkey. *Journal of Neurophysiology*, 76(4):2718–2739, 1996.

W. S. Geisler. Ideal observer analysis. In L. Chalupa and J. Werner, editors, *The Visual Neurosciences*, pages 825–837. MIT Press, Boston, 2003.

M. A. Goodale and A. D. Milner. Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25, 1992.

M. Goslin and M. R. Mine. The Panda3D Graphics Engine. *Computer*, 37(10): 112–114, 2004.

K. Grill-Spector, Z. Kourtzi, and N. Kanwisher. The lateral occipital complex and its role in object recognition. *Vision Research*, 41(10-11):1409–1422, 2001.

C. G. Gross. How inferior temporal cortex became a visual area. *Cerebral Cortex*, 4 (5):455–69, 1994.

J. Hegdé and D. C. Van Essen. Selectivity for complex shapes in primate visual area v2. *Journal of Neuroscience*, 20(5):61–66, 2000.

I. Helland. Partial least squares regression. *Encyclopedia of statistical sciences*, 9, 2006.

E. J. Holmes and C. G. Gross. Effects of inferior temporal lesions on discrimination of stimuli differing in orientation. *Journal of Neuroscience*, 4(12):3063–8, 1984.

J. A. Horel. Perception, learning and identification studied with reversible suppression of cortical visual areas in monkeys. *Behav Brain Res*, 76(1-2):199–214., 1996.

C. P. Hung, G. Kreiman, T. Poggio, and J. J. Dicarlo. Fast readout of object identity from macaque inferior temporal cortex. *Science*, 310(5749):863–866, 2005a.

C. P. Hung, G. Kreiman, T. Poggio, and J. J. DiCarlo. Fast readout of object identity from macaque inferior temporal cortex. *Science*, 310(5749):863–6, 2005b.

C. P. Hung, G. Kreiman, T. Poggio, and J. J. DiCarlo. Fast readout of object identity from macaque inferior temporal cortex. *Science*, 310(5749):863–6, 2005c.

J. Hyvarinen, H. Sakata, W. H. Talbot, and V. B. Mountcastle. Neuronal coding by cortical cells of the frequency of oscillating peripheral stimuli. *Science*, 162(3858): 1130–1132, 1968.

E. B. Issa and J. J. DiCarlo. Precedence of the eye region in neural processing of faces. *The Journal of Neuroscience*, 32(47):16666–16682, 2012.

B. Jagadeesh, H. S. Wheat, and D. Ferster. Linearity of summation of synaptic potentials underlying direction selectivity in simple cells of the cat visual cortex. *Science*, 262(5141):1901–4, 1993.

P. Janssen, S. Srivastava, S. Ombelet, and G. A. Orban. Coding of shape and position in macaque lateral intraparietal area. *The Journal of Neuroscience*, 28 (26):6679–6690, 2008.

K. O. Johnson, S. S. Hsiao, and T. Yoshioka. Neural coding and the basic law of psychophysics. *Neuroscientist*, 8(2):111–21, 2002.

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

K. N. Kay, T. Naselaris, R. J. Prenger, and J. L. Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, 2008.

J. Keat, P. Reinagel, R. C. Reid, and M. Meister. Predicting every spike: a model for the responses of visual neurons. *Neuron*, 30(3):803–17, 2001.

D. Kersten, P. Mamassian, and A. Yuille. Object perception as Bayesian inference. *Annual Review of Neuroscience*, 55:271–304, 2004.

S. M. Khaligh-Razavi and N. Kriegeskorte. Deep supervised, but not unsupervised, models May explain it cortical representation. *PLoS Computational Biology*, 2014.

R. Kiani, H. Esteky, K. Mirpour, and K. Tanaka. Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of Neurophysiology*, 97(6):4296–309, 2007.

E. Kobatake and K. Tanaka. Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of Neurophysiology*, 71(3):856–67, 1994.

J. J. Koenderink and A. J. van Doorn. The internal representation of solid shape with respect to vision. *Biological Cybernetics*, 32(4):211–216, 1979.

Z. Kourtzi and N. Kanwisher. Representation of perceived object shape by the human lateral occipital complex. *Science*, 293(5534):1506–1509, 2001.

N. Kriegeskorte. Relating population-code representations between man, monkey, and computational models. *Frontiers in Neuroscience*, 3(3):363, 2009.

N. Kriegeskorte, M. Mur, and P. Bandettini. Representational similarity analysis — connecting the branches of systems neuroscience. *Front Syst Neurosci*, 2:4, 2008a.

N. Kriegeskorte, M. Mur, D. A. Ruff, R. Kiani, J. Bodurka, H. Esteky, K. Tanaka, and P. A. Bandettini. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 60(6):1126–41, 2008b.

A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems (NIPS)*, 2012.

S. Lazebnik, C. Schmid, J. Ponce, et al. Spatial pyramid matching. *Object Categorization: Computer and Human Vision Perspectives*, 3:4, 2009.

Y. LeCun and Y. Bengio. Convolutional networks for images, speech, and time series. *The Handbook of Brain Theory and Neural Networks*, pages 255–258, 1995.

Y. Lecun, F.-J. Huang, and L. Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2004.

S. R. Lehky. Fine discrimination of faces can be performed rapidly. *Journal of Cognitive Neuroscience*, 12(5):848–855, 2000.

S. R. Lehky and A. B. Sereno. Comparison of shape encoding in primate dorsal and ventral visual pathways. *Journal of Neurophysiology*, 97(1):307–19, 2007.

P. Lennie and J. A. Movshon. Coding of color and form in the geniculostriate visual pathway (invited review). *Journal of the Optical Society of America A*, 22(10): 2013–33, 2005.

N. Li and J. J. DiCarlo. Unsupervised Natural Experience Rapidly Alters Invariant Object Representation in Visual Cortex. *Science*, 321(5895):1502, 2008.

N. Li, D. Cox, D. Zoccolan, and J. DiCarlo. Flexible and robust object representation in inferior temporal cortex supported by neurons with limited position and clutter tolerance. In *Society for Neuroscience*, 2006.

N. Li, D. D. Cox, D. F. Zoccolan, and J. J. DiCarlo. What response properties do individual neurons need to underlie position and clutter 'invariant' object recognition? *Journal of Neurophysiology*, 102(1):360–376, 2009. ISSN 0022-3077.

S. Liu, J. D. Dickman, S. D. Newlands, G. C. DeAngelis, and D. E. Angelaki. Reduced choice-related activity and correlated noise accompany perceptual deficits following unilateral vestibular lesion. *Proceedings of the National Academy of Sciences of the United States of America*, 110(44):17999–18004, 2013.

N. Logothetis and D. Sheinberg. Visual object recognition. *Annual Review of Neuroscience*, 19:577–621, 1996.

D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91–110, 2004.

R. Luna, A. Hernández, C. D. Brody, and R. Romo. Neural codes for perceptual discrimination in primary somatosensory cortex. *Nature Neuroscience*, 8(9):1210–1219, 2005.

S. P. MacEvoy and Z. Yang. Joint neuronal tuning for object form and position in the human lateral occipital complex. *NeuroImage*, 63(4):1901–1908, 2012.

N. Majaj, H. Hong, E. Solomon, and J. DiCarlo. A unified neuronal population code fully explains human object recognition. In *Computational and Systems Neuroscience (COSYNE)*, 2012.

R. Malach, I. Levy, and U. Hasson. The topography of high-order human object areas. *Trends in Cognitive Sciences*, 6(4):176–184, 2002.

D. Marr, T. Poggio, and S. Ullman. *Vision.* A Computational Investigation Into the Human Representation and Processing of Visual Information. MIT Press, July 2010.

K. Martin and S. Schroder. Functional heterogeneity in neighboring neurons of cat primary visual cortex in response to both artificial and natural stimuli. *Journal of Neuroscience*, 33(17), 2013.

B. W. Mel. Seemore: combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Comput*, 9(4): 777–804, 1997.

W. H. Merigan. The contrast sensitivity of the squirrel monkey (saimiri sciureus). *Vision Research*, 16(4):375–379, 1976.

E. M. Meyers, D. J. Freedman, G. Kreiman, E. K. Miller, and T. Poggio. Dynamic population coding of category information in inferior temporal and prefrontal cortex. *Journal of Neurophysiology*, 100(3):1407–1419, 2008.

M. Mishkin, L. G. Ungerleider, and K. A. Macko. Object vision and spatial vision: two cortical pathways. *Trends in Neurosciences*, 6:414–417, 1983.

J. F. Mitchell, K. A. Sundberg, and J. H. Reynolds. Spatial attention decorrelates intrinsic activity fluctuations in macaque area v4. *Neuron*, 63(6):879–888, 2009.

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

V. B. Mountcastle, W. H. Talbot, H. Sakata, and J. Hyvarinen. Cortical neuronal mechanisms in flutter-vibration studied in unanesthetized monkeys: Neuronal periodicity and frequency discrimination. *Journal of Neurophysiology*, 1969.

J. Mutch and D. G. Lowe. Object class recognition and localization using sparse features with limited receptive fields. *International Journal of Computer Vision (IJCV)*, 2008.

H. Nakamura, R. Gattass, R. Desimone, and L. Ungerleider. The modular organization of projections from areas v1 and v2 to areas v4 and teo in macaques. *Journal of Neuroscience*, 14(9):1195–1201, 2011.

T. Naselaris, R. J. Prenger, K. N. Kay, M. Oliver, and J. L. Gallant. Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63(6):902–915, 2009.

W. T. Newsome, K. H. Britten, and J. A. Movshon. Neuronal correlates of a perceptual decision. *Nature*, 341(6237):52–4, 1989.

A. Nishio, T. Shimokawa, N. Goda, and H. Komatsu. Perceptual gloss parameters are encoded by population responses in the monkey inferior temporal cortex. *The Journal of Neuroscience*, 34(33):11143–11151, 2014.

L. G. Nowak and J. Bullier. The timing of information transfer in the visual system. In *Extrastriate cortex in primates*, pages 205–241. Springer, 1997.

J. O'Kusky and M. Colonnier. A laminar analysis of the number of neurons, glia, and synapses in the visual cortex (area 17) of adult macaque monkeys. *Journal of Comparative Neurology*, 210(3):278–290, 1982.

A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, 155:23–36, 2006.

H. Op de Beeck and R. Vogels. Spatial sensitivity of macaque inferior temporal neurons. *Journal of Comparative Neurology*, 426(4):505–18., 2000.

H. Op de Beeck, J. Wagemans, and R. Vogels. Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nature Neuroscience*, 4 (12):1244–52, 2001.

G. A. Orban. Higher order visual processing in macaque extrastriate cortex. *Physiological Reviews*, 88(1):59–89, 2008.

G. A. Orban, D. Van Essen, and W. Vanduffel. Comparative mapping of higher visual areas in monkeys and humans. *Trends in Cognitive Sciences*, 8(7):315–24, 2004.

M. Pagan, L. S. Urban, M. P. Wohl, and N. C. Rust. Signals in inferotemporal and perirhinal cortex suggest an untangling of visual target information. *Nature Neuroscience*, 16(8):1132–1139, 2013.

A. J. Parker and W. T. Newsome. Sense and the single neuron: probing the physiology of perception. *Annual Review of Neuroscience*, 21(1):227–277, 1998.

A. Pasupathy and C. Connor. Population coding of shape in area v4. *Nature Neuroscience*, 5(12):1332–1338, 2002.

A. Pasupathy and C. E. Connor. Responses to contour features in macaque area v4. *Journal of Neurophysiology*, 82(5):2490–2502, 1999.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research (JMLR)*, 12:2825–2830, 2011.

N. Pinto, D. D. Cox, and J. J. Dicarlo. Why is real-world visual object recognition hard? *PLoS Computational Biology*, 2008a.

N. Pinto, D. D. Cox, and J. J. DiCarlo. Why is Real-World Visual Object Recognition Hard. *PLoS Computational Biology*, 2008b.

N. Pinto, D. Doukhan, J. J. Dicarlo, and D. D. Cox. A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLoS Computational Biology*, 2009.

N. Pinto, Y. Barhomi, D. D. Cox, and J. J. DiCarlo. Comparing state-of-the-art visual features on invariant object recognition tasks. In *Applications of computer vision (WACV), 2011 IEEE workshop on*, pages 463–470. IEEE, 2011.

D. Pitcher, L. Charles, J. T. Devlin, V. Walsh, and B. Duchaine. Triple dissociation of faces, bodies, and objects in extrastriate cortex. *Current Biology*, 19(4):319–24, 2009.

T. Plachetka. POV Ray: Persistence of Vision Parallel Raytracer. *Proceedings the of Springer Conference on Computer Graphics*, 1998.

R. Q. Quiroga, Z. Nadasdy, and Y. Ben-Shaul. Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. *Neural Computation*, 16 (8):1661–1687, Aug. 2004. doi: 10.1162/089976604774201631.

R. C. Reid and J. M. Alonso. The processing and encoding of information in the visual cortex. *Current Opinion in Neurobiology*, 6(4):475–80, 1996.

M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–25, 1999a.

M. Riesenhuber and T. Poggio. Hierarchical Models of Object Recognition in Cortex. *Nature Neuroscience*, 2(11):1019–1025, 1999b.

M. Riesenhuber and T. Poggio. Models of object recognition. *Nature Neuroscience*, 3 Suppl:1199–204., 2000.

C. A. Rishel, G. Huang, and D. J. Freedman. Independent category and spatial encoding in parietal cortex. *Neuron*, 77(5):969–979, 2013.

E. T. Rolls. Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition. *Neuron*, 27(2):205–18., 2000.

E. Rosch, C. B. Mervis, W. D. Gray, and D. M. Johnson. Basic objects in natural categories. *Cognitive Psychology*, 8:382–439, July 1976.

N. C. Rust and J. J. DiCarlo. Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area v4 to it. *Journal of Neuroscience*, 30(39):12978–95, 2010.

N. C. Rust and J. J. DiCarlo. Balanced increases in selectivity and tolerance produce constant sparseness along the ventral visual stream. *The Journal of Neuroscience*, 32(30):10170–10182, 2012.

N. C. Rust, V. Mante, E. P. Simoncelli, and J. A. Movshon. How mt cells analyze the motion of visual patterns. *Nature Neuroscience*, 9(11):1421–31, 2006.

R. Sayres and K. Grill-Spector. Relating retinotopic and object-selective responses in human lateral occipital cortex. *Journal of Neurophysiology*, 100(1):249–267, 2008.

R. E. Schapire. Theoretical views of boosting and applications. *Algorithmic Learning Theory*, 1999.

P. Schiller. Effect of lesion in visual cortical area v4 on the recognition of transformed objects. *Nature*, 376:342–344, 1995.

M. T. Schmolesky, Y. Wang, D. P. Hanes, K. G. Thompson, S. Leutgeb, J. D. Schall, and A. G. Leventhal. Signal timing across the macaque visual system. *Journal of Neurophysiology*, 79(6):3272–3278, 1998.

A. B. Sereno and S. R. Lehky. Population coding of visual space: comparison of spatial representations in dorsal and ventral pathways. *Frontiers in Computational Neuroscience*, 4, 2010.

A. B. Sereno, M. E. Sereno, and S. R. Lehky. Recovering stimulus locations using populations of eye-position modulated neurons in dorsal and ventral visual streams of non-human primates. *Frontiers in Integrative Neuroscience*, 8, 2014.

T. Serre, A. Oliva, and T. Poggio. A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences of the United States of America*, 104(15):6424–9, 2007a. 0027-8424 (Print) Journal Article.

T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2007b.

M. N. Shadlen, K. H. Britten, W. T. Newsome, and J. A. Movshon. A computational analysis of the relationship between neuronal and behavioral responses to visual motion. *Journal of Neuroscience*, 16(4):1486–510, 1996.

T. O. Sharpee, M. Kouh, and J. H. Reyholds. Trade-off between curvature tuning and position invariance in visual area v4. *Proceedings of the National Academy of Sciences of the United States of America*, 110(28):11618–11623, 2012.

D. L. Sheinberg and N. K. Logothetis. The role of temporal cortical areas in perceptual organization. *Proceedings of the National Academy of Sciences of the United States of America*, 94(7):3408–3413, 1997.

D. L. Sheinberg and N. K. Logothetis. Noticing familiar objects in real world scenes: the role of temporal cortical neurons in natural vision. *Journal of Neuroscience*, 21(4):1340–50., 2001.

K. Shoemake. Animating rotation with quaternion curves. In *ACM SIGGRAPH computer graphics*, volume 19, pages 245–254. ACM, 1985.

N. Sigala, F. Gabbiani, and N. Logothetis. Visual categorization and object representation in monkeys and humans. *Journal of Cognitive Neuroscience*, 14(2): 187–198, 2002.

Y. Sugase, S. Yamane, S. Ueno, and K. Kawano. Global and fine information coded by single neurons in the temporal visual cortex. *Nature*, 400(6747):869–873, 1999.

S. K. Swaminathan and D. J. Freedman. Preferential encoding of visual categories in parietal cortex compared with prefrontal cortex. *Nature Neuroscience*, 15(2): 315–320, 2012.

K. Tanaka. Neuronal mechanisms of object recognition. *Science*, 262:685–685, 1993.

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

K. Tanaka. Mechanisms of visual object recognition: monkey and human studies. *Current Opinion in Neurobiology*, 7(4):523–529, 1997.

M. J. Tarr and H. H. Bülthoff. Image-based object recognition in man, monkey and machine. *Cognition*, 67(1):1–20, 1998.

B. S. Tjan and G. E. Legge. The viewpoint complexity of an object-recognition task. *Vision Research*, 38(15-16):2335–50., 1998.

D. Y. Tsao and M. S. Livingstone. Mechanisms of face perception. *Annual Review of Neuroscience*, 31:411–37, 2008.

D. Y. Tsao, W. A. Freiwald, R. B. Tootell, and M. S. Livingstone. A cortical region consisting entirely of face-selective cells. *Science*, 311(5761):670–674, 2006.

L. G. Ungerleider and J. V. Haxby. "what" and "where" in the human brain. *Current Opinion in Neurobiology*, 4(2):157–165, 1994.

L. G. Ungerleider, T. W. Galkin, R. Desimone, and R. Gattass. Cortical connections of area v4 in the macaque. *Cerebral Cortex*, 18(3):477–99, 2008.

B. E. Verhoef, R. Vogels, and P. Janssen. Inferotemporal cortex subserves three-dimensional structure categorization. *Neuron*, 73(1):171–82, 2012.

R. Vogels and G. Orban. Activity of inferior temporal neurons during orientation discrimination with successively presented gratings. *Journal of Neurophysiology*, 71:1428–1451, 1994.

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

L. Weiskrantz and R. C. Saunders. Impairments of visual object transforms in monkeys. *Brain*, 107:1033–1072, 1984.

S. Yaginuma, T. Niihara, and E. Iwai. Further evidence on elevated discrimination limens for reduced patterns in monkeys with inferotemporal lesions. *Neuropsychologia*, 20(1):21–32, 1982.

Y. Yamane, E. T. Carlson, K. C. Bowman, Z. Wang, and C. E. Connor. A neural code for three-dimensional object shape in macaque inferotemporal cortex. *Nat Neurosci*, 2008.

D. Yamins*, H. Hong*, C. Cadieu, and J. Dicarlo. Hierarchical modular optimization of convolutional networks achieves representations similar to macaque it and human ventral stream. *Advances in Neural Information Processing Systems (NIPS)*, 2013.

D. Yamins*, H. Hong*, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 2014.

A. Yarbus. Eye movements during perception of complex objects. *Eye movements and vision*, 7:171–196, 1967.

T. Yoshioka, B. Gibb, A. K. Dorsch, S. S. Hsiao, and K. O. Johnson. Neural coding mechanisms underlying perceived roughness of finely textured surfaces. *The Journal of Neuroscience*, 21(17):6905–6916, 2001.

M. D. Zeiler and R. Fergus. Visualizing and Understanding Convolutional Networks. arXiv.org, Nov. 2013.

S. Zhu and D. Mumford. A stochastic grammar of images. *Foundations and Trends in Computer Graphics and Vision*, 2:259–362, 2006.

D. Zoccolan, N. Oertelt, J. J. DiCarlo, and D. D. Cox. A rodent model for the study of invariant visual object recognition. *Proceedings of the National Academy of Sciences of the United States of America*, 106(21):8748–8753, 2009. ISSN 0027-8424.

E. Zohary, M. N. Shadlen, and W. T. Newsome. Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature*, 370:140–143, 1994.