

Computation and representation in the primate visual system

Jeremy Freeman

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Center for Neural Science
New York University
January 2013

Eero P. Simoncelli

© Jeremy Freeman
All Rights Reserved, 2013

Dedicated to Alan D. Freeman

Acknowledgements

Much of the work described in this thesis, in particular the work described in Chapters 3 and 4, was performed in close collaboration with Corey Ziemba: a remarkably talented physiologist who has also become one of my closest friends.

Thanks to my advisors and mentors, in chronological order: Denis Pelli, David Heeger, Michael Landy, Eero Simoncelli, Tony Movshon, and E.J. Chichilnisky. Each fundamentally and permanently shaped my approach to science.

Thanks to my friends: Deep Ganguli, for teaching me how to chill out; John Tuthil, James Golden, and Arpiar Saunders, for showing me other kinds of neuroscience; Peter Baker, Dan Hammer, Adam Hunt, Scott Young, Ben Sussman, Emily Allen, and Meredith Leich, for their gifted insights into what the world can offer outside science; Matt Haxby and Mike Brotzman, for the nights playing video games; and Anne Lind, for support, companionship, and the comforts of dairy.

Thanks to my collaborators: Eli Merriam, Gijs Brouwer, Helena Wang, Tobias Donner, Luke Hallum, Rama Chakravarthi, Yan Karklin, Umesh Rajashaker, Deep Ganguli, Greg Field, and Peter Li. Each taught me to think in a new way.

And thanks to my family, for encouraging my passions and tolerating the late-night Amtrak arrivals: my siblings James, Josh, John, Jennifer; my extended family Steve, Greg, Jake, Joe, Mariana, Brian; my pets Happy, Norman, Tammie, Esther, Mr. Xerxes, Flopsy; and most of all my mom, Betty.

Preface

I started graduate school at New York University in the fall of 2008. The subsequent four years were the most intellectually exciting and stimulating of my life. Many of the scientific avenues I pursued were fruitful and productive, and many were not. Topics ranged from high-level visual perception and object recognition to nonlinear processing in retinal ganglion cells, and lots in between. This thesis consolidates what I consider the most internally cohesive and conceptually important subset of my work, concerning a new computational framework for understanding the representation of information in the intermediate stages of the primate visual system, and how that framework can guide perceptual and physiological investigation. Some of the contents of this thesis have appeared previously in published journal articles [75], conference abstracts, or manuscripts currently under review or in preparation. The contents of other published work beyond the scope of this thesis have been left out in part or in full [74, 167, 231, 73, 72], but my thinking about them has greatly influenced the work presented here. As such, this thesis reflects the intellectual influence of all of my academic advisors and collaborators. This work would not exist without them.

Jeremy Freeman

New York, New York

Abstract

The purpose of vision is to find behaviorally relevant structure in the ever-flowing chaos of sensory input. In the primate, this goal is achieved by a hierarchy of cortical areas that extract increasingly complex forms of information from the light arriving at the retina. Despite success characterizing the early stages of this pathway – the retina, the lateral geniculate nucleus, and primary visual cortex (V1) – we have a poor understanding of how transformations in later stages yield selectivity for the complex shapes and objects that primates readily recognize.

According to a classical, constructionist view, the later stages of the visual system assemble elementary inputs – like the oriented features encoded by V1 – into larger and more complex combinations, capturing the structural relationships that determine the visual world. But this approach has stumbled on the enigmatic second visual area, V2, whose neurons defy our intuitions about how to begin segmenting scenes and encoding the shapes of objects.

In this thesis we develop a framework for the study of intermediate visual processing in the primate, focused on computation and representation in area V2. Rather than try to predict the responses of visual neurons to arbitrary inputs, we test hypotheses about their function by generating targeted experimental stimuli. The stimuli we use reflect the messy statistical reality of natural images, rather than intuitions about object construction. We identify novel responses properties in

macaque and human V2 that robustly differentiates it from V1. We propose mechanistic explanations for these properties by contextualizing them among existing models of hierarchical computation. And we link these properties to several perceptual capabilities – and limits – that appear to depend specifically on processing in V2, and imply striking consequences for everyday vision.

Contents

Dedication	iii
Acknowledgements	iv
Preface	v
Abstract	vi
List of Figures	x
1 Introduction	1
1.1 The problem of vision	1
1.2 Structure of the thesis	7
2 Background	9
2.1 Summary of visual pathways	9
2.2 The second visual area	14
2.3 Computation and theory	22
2.4 Perception	35
2.5 Outlook	39

3	Responses to naturalistic stimuli differentiate V2 from V1	41
3.1	Introduction	41
3.2	Synthesis of naturalistic stimuli	43
3.3	Electrophysiology in macaques	51
3.4	fMRI in humans	65
3.5	Possible mechanisms	78
3.6	Discussion	85
4	Linking perception and physiology through V2	89
4.1	Introduction	89
4.2	Perceptual discrimination	91
4.3	Crowdsourcing psychophysics	98
4.4	Perceptual invariance	114
4.5	Discussion	125
5	Population representations in V2 predict visual metamers	129
5.1	Introduction	129
5.2	Model structure and image synthesis	133
5.3	Perceptual determination of critical scaling	140
5.4	Estimation of physiological locus	152
5.5	Relationship to visual crowding	155
5.6	Discussion	164
6	Conclusion	169
6.1	Summary of contributions	170
6.2	Future work	172
	Bibliography	177

List of Figures

2.1	Anatomy of macaque visual pathways	10
2.2	Hierarchy of the primate visual system	11
2.3	Models of early visual processing	12
2.4	Sensitivity to shape and curvature in V2	16
2.5	Simulations of shape selectivity	17
2.6	Border ownership signaling in V2	19
2.7	Spatial diversity of orientation preference in V2	20
2.8	Magnitude dependencies in natural images	24
2.9	Normalization eliminates magnitude dependencies	26
2.10	Model neurons that learn magnitude dependencies	28
2.11	Comparing texture models	31
2.12	Synthesis-by-analysis algorithm	32
2.13	Portilla and Simoncelli example successes	33
2.14	Portilla and Simoncelli example failures	34
2.15	The filter-rectify-filter model	36
3.1	Steerable pyramid	44
3.2	Texture synthesis	48
3.3	Experimental stimuli	50

3.4	Stimulus sequence in the physiology experiment	52
3.5	Differential responses to naturalistic images in V2	55
3.6	Distribution of modulation	56
3.7	Variability across categories	57
3.8	Modulation and receptive field size	58
3.9	Modulation and basic properties	60
3.10	Marginal control experiment	62
3.11	Discriminating visual areas	64
3.12	Stimulus sequence in the fMRI experiment	67
3.13	fMRI responses to naturalistic stimuli in V2	71
3.14	Reliability of fMRI responses	72
3.15	Event-related modulation	74
3.16	Correlations between macaque and human	75
3.17	Evidence for feedback in V1: fMRI	76
3.18	Evidence for feedback in V1: electrophysiology	77
3.19	LN mechanisms do not differentiate naturalistic images	79
3.20	A mechanism uniquely sensitive to naturalistic images	81
4.1	Interpolating between noise and naturalistic	92
4.2	3AFC psychophysical “oddity” task	94
4.3	Example psychometric functions	95
4.4	Correlations between psychophysics, fMRI, and physiology	96
4.5	Correlations in single neurons	97
4.6	Comparing the laboratory and the crowd	105
4.7	Distribution of sensitivity	106
4.8	Validating the correlation between perception and V2 physiology . .	107

4.9	PCA reduces the dimensionality of texture parameters	110
4.10	Predicting sensitivity from texture parameters	111
4.11	Statistically-matched samples of a texture category	115
4.12	Relating variance within and across categories	116
4.13	The F -statistic is related to receptive field size.	117
4.14	t -SNE applied to V1 population responses	120
4.15	t -SNE applied to V2 population responses	121
4.16	Comparing low-dimensional maps in V1 and V2	122
4.17	High-dimensional distances capture clustering	124
5.1	Receptive field size and eccentricity	131
5.2	Model pooling regions	135
5.3	ABX psychophysical task	141
5.4	Metamers for the V2 model	144
5.5	Metamers for the V1 model	145
5.6	Metamer psychometric functions	147
5.7	Scaling is robust to manipulations	151
5.8	Relating perceptual and physiological scaling	155
5.9	Simulations of crowding	157
5.10	Directly relating the model to crowding	160
5.11	Dyslexia and crowding	162
5.12	Consequences of metamerism for visual search	163

Chapter 1

Introduction

1.1 The problem of vision

Here lay a way to formulate the purpose of vision – building a description of the shapes and positions of things from images. Of course, that is by no means all that vision can do; it also tells about the illumination and about the reflectances of the surfaces that make the shapes – their brightness and colors and visual textures – and about their motion. But these things seemed secondary; they could be hung off a theory in which the main job of vision was to derive a representation of shape.

– David Marr

These words articulate an intuitive notion of how vision works. We look around the world and see shapes and objects that interest us. We point to them, talk about them, interact with them. And we segment the world into these objects by using

the cascade of neuronal processing in our brain's visual system, which implements the computations and algorithms required for the task.

Marr championed this constructionist view of how the world is assembled and how the visual system segments it, and visual neuroscience has been endowed with it ever since. Adelson put it crisply: "Our world contains both things and stuff, but things tend to get the attention" [1].

But the focus on individuation did not begin with Marr. It has appeared across several philosophical and epistemological traditions, a few of which bear mentioning. The philosophers of the high middle-ages, especially Thomas Aquinas, inherited from Aristotle a rigid naturalism that linked understanding an object to perceiving its material form. Unlike Plato, for whom the forms of things were abstract concepts to be apprehended through argument and logic, Aristotle, and later Aquinas, emphasized that forms were to be found in the material, perceptible world. In *A Portrait of the Artist as a Young Man*, one of James Joyce's characters, the university student Stephen, provides a flowery account of Aquinas' epistemology: "Look at the basket, he said . . . In order to see that basket, said Stephen, your mind first of all separates the basket from the rest of the visible universe which is not the basket. The first phase of apprehension is a bounding line drawn about the object to be apprehended . . . the esthetic image is first luminously apprehended as selfbounded and selfcontained upon the immeasurable background of space and time which is not it. You apprehend it as *one* thing. You see it as one whole. You apprehend its wholeness. That is *integritas*."¹ For Aquinas, *seeing* material objects was crucial to understanding them – an early epistemological cornerstone of the scientific method – and he particularly emphasized seeing objects.

An emphasis on individuation of objects is also rooted in traditional accounts of how humans use language. In explaining his "picture theory of language," Saint

Augustine describes his memory of learning, as a small child, to identify objects with words; "When they (my elders) named some object, and accordingly moved towards something, I saw this and I grasped that the thing was called by the sound they uttered when they meant to point it out. Their intention was shown by the bodily movements, as it were the natural language of all peoples; the expression of the face, the play of the eyes, the movement of other parts of the body, and the tone of voice when expressed our state of mind in seeking, having, rejecting or avoiding something. Thus, as I heard words repeatedly used in their proper places in various sentences, I gradually learnt to understand what objects they signified; and after I had trained my mouth to form these signs, I used them to express my own desires. "2 Wittgenstein was drawn to this intuitive notion that "individual words in language name objects," and his early work articulated a theory according to which the meaning of a statement about objects in the world is evaluated via a correspondence between the ontological organization of the world into individuated elements, and the logical structure of the words we use to describe those elements and the relationships among them.

In later work, however, Wittgenstein recognized as a naive temptation both the assumption that the world is so organized and individuated, and the assumption that language and perception glom onto its precise structure. Crucial to his critique was the idea that the meaning of words are as rooted in their correspondence to things in the world as they are to one another, and the complex social discourses in which they are *used*. Concepts do not delineate precise categories, but rather form fluid networks or *families* in which ideas, and words, resemble one another.

The fluidity of linguistic concepts mirrors the fluid ontology of objects. Natural images – the ordinary input to the visual system – primarily contain complicated mixtures of stuff, like textures and colors, and only occasionally the sharp outline

of an object. In real images, patterns can appear both distinct and object-like or statistical and textural depending on the scale at which we perceive them: an isolated face may be a distinct object, but a crowd of faces becomes a texture. Most ordinary objects are also not defined purely by their shape or by their texture, but by some fluid mixture of the two. Adelson describes such a tension when struggling to capture the defining features of a tree: it “is not a shape that can be template matched, since the particular branches are different for every tree”. But “it is not a texture either, since there is a top and a bottom, and a textural quality that changes from one part to another” [1].

The crisp organization of the world into objects may thus be a distracting intuition, abstracted from the mechanisms of vision. Indeed, animals much simpler than primates are highly visual: a bee presumably navigates the garden by seeking out textures and colors in its sensory input. Nevertheless, it may not carefully delineate an ontology of different plants and flowers, nor does its survival depend on doing so. This idea is echoed in the “ecological psychology” of Merleau-Ponty and Gibson who, among others, argued that vision is a process of continually interrogating the world in order to extract properties of interest. The function of the brain, Gibson claimed, is to “seek and extract information about the environment from the flowing array of ambient energy.”³ But whether or not the extracted information is a set of object categories, or a rich mess of texture and stuff, may be irrelevant.

Most significantly, and most relevant to this thesis, is that the constructionist program has stumbled when trying to map its computations onto the structure and function of the primate visual system. The goal of visual neuroscience is to “understand vision in physiological terms” and delineate how the physiology supports and constrains visual perception and action [112]. For decades, experimenters have traced the flow of visual signals through the neurophysiology of the early visual

system, from the retina, to LGN, to primary visual cortex (V1). The responses of neurons in each of these areas can be reasonably well described using compact phenomenological, or *functional*, models that relate the visual input to the response. Neurons in these early areas encode information about local spatial and temporal contrast, and orientation. In particular, Hubel and Wiesel's success identifying the basic properties and mechanisms of orientation selectivity in V1, and capturing the construction of these properties with simple feed forward models [110], held great promise for characterizing later stages. But carrying this project forward has been a surprisingly formidable challenge. As Hubel wrote in a retrospective on current major problems of neuroscience, "We have almost no examples of neural structures in which we know the difference between the information coming in and what is going out-what the structure is for. We have some idea of the answer for the retina, the lateral geniculate body, and the primary visual cortex, but that's about it" [109].

Most efforts have taken the individuation of objects as a key goal. In some downstream areas of the ventral stream, like V4 and IT (inferotemporal cortex), neurons have been identified that exhibit selectivity to particular object categories or complex shapes, with responses that tolerate variation in the size, location, or surrounding context of the input [14, 117, 208, 138, 60, 115, 241, 189]. But it remains unclear how the stages of processing between V1 and these later areas achieve such complex responses. Perhaps most enigmatic is the second visual area, V2. Given its location, between V1 and downstream areas, it is tempting to imagine that V2 begins constructing selectivity to the elements that make up objects. Many experiments in V2 have been directly motivated by that intuition – for example, measuring the responses of V2 neurons to angles, curves, and other features that resemble the bits and pieces of objects [102, 103]. These studies have identified subpopulations of V2 neurons with distinctive response properties, but none have

identified properties that robustly differentiate neurons in V2 from those in V1 [171, 116, 225, 129, 141, 102, 103, 63].

Parallel efforts in computer vision have constructed hierarchical systems that mimic perceptual tasks performed by humans, like object recognition [90, 78, 49, 179, 181, 197]. Many of these systems perform well in limited problem domains, but fail when confronted with the rich input variability encountered during real-world recognition [61, 173]. Notably, although the high-level representations of some of these models have been linked to properties of IT neurons, and related hierarchical models have predicted forms of selectivity found in V4 [32], the intermediate computations of these models have neither described nor predicted properties of neurons in V2. New insights into how V2 begins the biological solution to pattern recognition could have a profound impact on representational strategies in computer vision, as well as implications for human perception.

This thesis describes a series of investigations – physiological, perceptual, and computational – probing the function of primate V2. As will be explained, our progress in this endeavor reflects, in large part, a critique of Marr’s guiding principle of object individuation. At the same time, the inter-disciplinary nature of the work is indebted to Marr. He postulated that any information processing system must be understood at three distinct levels – computational theory, representation and algorithm, and implementation. The complexity of a problem like V2 has compelled us to engage all three of these levels, simultaneously wherever possible. This methodological combination has been crucial to our success in V2, and will likely be important to success describing intermediate computation in other systems.

1.2 Structure of the thesis

In the next Chapter, we review in more detail the visual pathways of the primate, emphasizing previous experimental efforts in V2. We also describe computational and psychophysical investigations into possible intermediate stages of visual representation, emphasizing those that have directly motivated the experiments described in this thesis.

Chapter 3 introduces a new approach to studying V2. We describe the generation of synthetic, stochastic experimental stimuli that are based on naturally-occurring images of visual texture. We establish that neurons in V2 are robustly distinguished from neurons in V1 in their responses to these stimuli. Parallel experiments in humans using functional magnetic resonance imaging (fMRI) establish sensitivity to the same image features in human V2 but not V1.

Chapter 4 presents specific links between the physiological responses described in Chapter 3, and perceptual sensitivities to the same stimuli. We establish a form of perceptual sensitivity to stimuli containing naturalistic features (compared to stimuli lacking them), and also a form of perceptual invariance in which distinct images appear similar because they share the same statistical properties. We describe physiological correlates of both perceptual properties, implying the behavioral relevance of neuronal responses in V2.

In Chapter 5, we describe a perceptual consequence of the V2 representation established in the preceding chapters. We show that it predicts novel visual *metamers* – heterogeneous images that are physically different, but appear similar because they yield similar neuronal responses in V2 populations. We describe behavioral experiments documenting the perception of these metamers and linking them specifically to V2, and end by discussing the consequences of these metamers for everyday vision

and the phenomenon of visual crowding.

Chapter 6 summarizes the contributions of this thesis, emphasizing both what we have learned about V2 but also the puzzles that remain, and guidelines for how to explore them further.

It bears mention that the work described in this thesis was performed in an order opposite to how it is presented here. We performed the metamer experiments described in Chapter 5 first. In those experiments, we identified an indirect link between a perceptual model and known physiological properties of V2 neurons, specifically, the size of their receptive fields. This motivated us to perform the physiological and fMRI experiments exploring V2 described in Chapters 3 and 4. The order presented here is more conceptually coherent, but we consider the order in which the work was done a testament to the power of computation and psychophysics in guiding physiological investigation.

Notes

¹From *A Portrait of the Artist as a Young Man* by James Joyce, as quoted by Jeremiah Hackett in his essay “Duns Scotus: A Brief Introduction to his Life and Thought”, *Studies in Scottish Literature*, 26(1):37, 1991.

²From *The Confessions* by Saint Augustine, as quoted by Ludwig Wittgenstein in the intro to *The Philosophical Investigations*, New York, NY: Pearson, 1973.

³From *The Senses Considered as Perceptual Systems* by J. J. Gibson, as quoted by David Marr in *Vision*, New York, NY: W. H Freeman and Co, 1982.

Chapter 2

Background

2.1 Summary of visual pathways

Visual processing in primates is hierarchical [68], with a sequence of areas that process increasingly complex forms of information, from the retina, to the lateral geniculate nucleus (LGN), to primary visual cortex (V1) and the second visual area (V2), and then the extrastriate areas of the dorsal and ventral streams. Areas in the dorsal stream, most notably MT and MST, are involved in the processing of visual motion, whereas ventral areas, including V4, and inferotemporal cortex (IT), have been linked to the representation of complex visual patterns and objects [218, 59]. All of these areas – their sizes, and the relationships among them – are depicted anatomically in Figure 2.1 and schematically in Figure 2.2.

Neurons at early stages of visual processing have been characterized by describing the functional relationship between the visual input and their response. Many neurons in both the retina and LGN can be characterized using a linear-nonlinear (LN) model [42, 35] (see Figure 2.3). The LN model describes a neuron as encoding a visual feature in a particular location of the visual field, determined by its linear

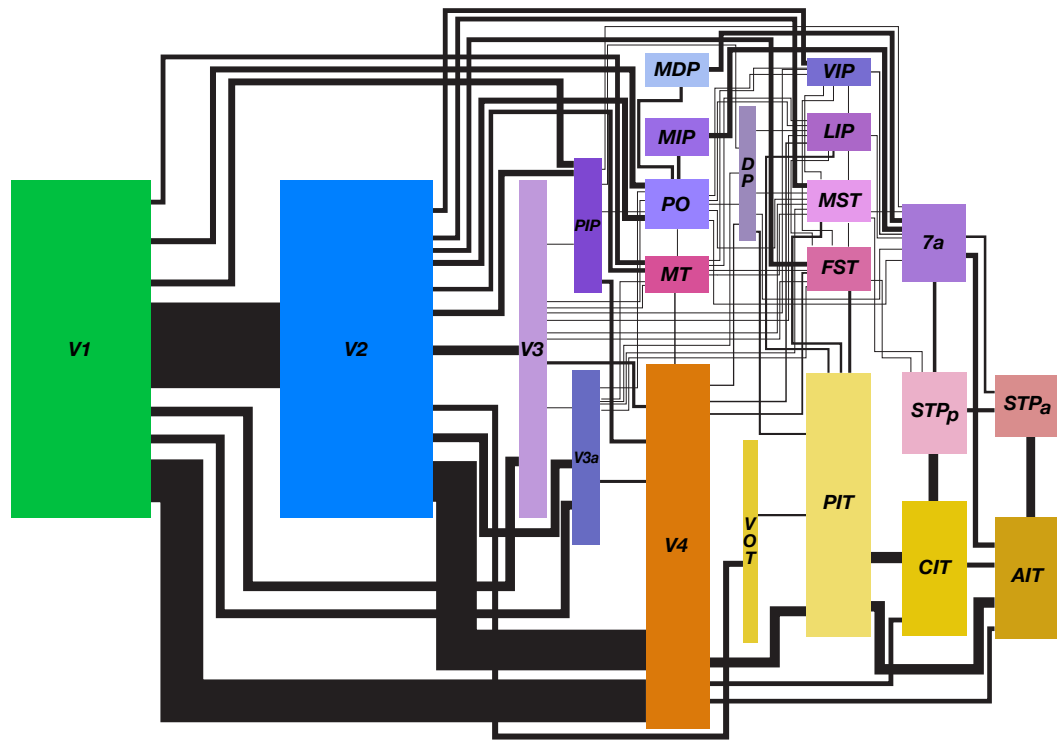


Figure 2.2: Hierarchical organization of the primate visual system, adapted from [228] after [68]. The size of each box is proportional to cortical surface area, and line thickness is proportional to the number of fibers connecting each area.

response to high spatial frequency contrast-modulating gratings, stimuli that ought to cancel within the neuron’s linear receptive field according to the LN model. These non-linear properties have been characterized using hierarchical “subunit” models with two stages of linear integration and non-linearity [106, 224, 223].

The first visual cortical area, V1, is also relatively well characterized in terms of its circuitry and function [111, 113, 151, 150, 130]. Its neurons exhibit sensitivity to the local orientation and spatial frequency of a visual stimulus. Some “simple” V1 neurons can be well described with a linear receptive field followed by a nonlinearity, and the properties of the linear filter can be recovered by measuring responses to

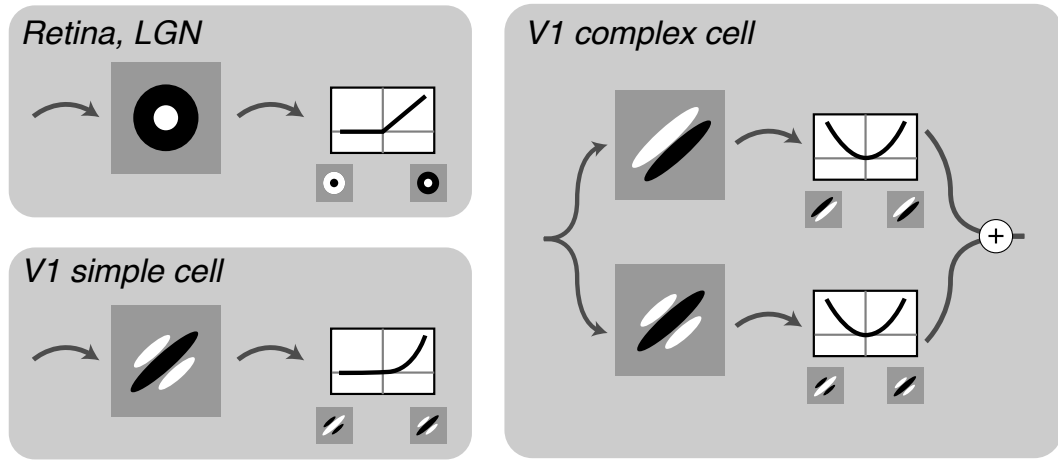


Figure 2.3: Models commonly used to describe responses in the retina, LGN, and V1. The image in each box depicts a linear receptive field, and the functions depict point-wise nonlinearities (e.g. rectification or squaring).

simple noise or sinusoidal grating stimuli. “Complex” cells are similarly orientation-tuned, but exhibit insensitivity to the precise position of a stimulus; their responses to gratings are, at least to some extent, *invariant* to spatial phase [111, 150, 151]. This property is elegantly captured by the “energy model”, which sums and squares the output of two phase-shifted oriented receptive fields to capture insensitivity to phase [2] (Figure 2.3). Real V1 neurons, however, are typically neither perfectly simple nor complex. Their diversity can be captured by more general “quadratic” or “subunit” models that employ a feedforward combination of rectified or squared linear filter responses, analogous to the hierarchical model described above for nonlinear retinal ganglion cells [150, 151, 97, 37, 187, 41, 207, 225]. Finally, adding divisive gain control to these models helps capture other non-linear effects, such as cross-orientation masking, surround suppression, and response saturation [7, 20, 37, 227, 55, 40, 85, 97, 31, 207]. Together, these efforts have yielded compact functional models that successfully capture selectivity in V1 for orientation, spatial

frequency, binocular disparity, and color [150, 151, 54, 187, 213, 158, 107, 191], and can account for most of the explainable variance in V1 responses to simple stimuli [187] and much of the variance in responses to natural stimuli [51, 237].

According to the classical hierarchical view [68, 110], higher extra-striate areas, including V2, V4, IT, and MT, integrate inputs from V1 to encode more complex features of the visual input. In the process, cells simultaneously become selective for specific image features, develop complex forms of invariance, and become more differentiated from one another in their responses [46, 60]. Consequently, linking single cell responses to the input becomes more difficult. The simple one-stage or two-stage models discussed thus far cannot capture the more complex relationship between stimulus and response. Similarly, stimuli used to study earlier stages (e.g. white noise and gratings) are insufficient because they rarely contain the special features that drive these cells to respond.

Two areas of investigation in extrastriate cortex have been notably successful. First is the characterization of how neurons in extrastriate area MT encode visual motion information. The understanding of MT began with the discovery of pattern-motion selectivity [149], and this robust phenomenon provided a basis for subsequent modeling and characterization. A two-stage cascade model accounts for this property of MT neurons and explains a variety of related psychophysical and neuronal phenomena [201]. The same model can also be used to predict single neuron responses to rich ensembles of motion stimuli [186]. Some of these electrophysiological findings in macaques have been corroborated in humans using functional magnetic resonance imaging (fMRI) [99, 114].

There has also been success characterizing responses of extrastriate ventral areas V4 and IT. Neurons in IT exhibit selectivity to highly complex patterns and objects [208, 138, 60, 61], and their responses also tolerate a variety of physical

transformations, like changes in size and position [14, 117, 115, 241, 189]. Similarly, fMRI studies have identified several areas in human inferotemporal cortex involved in representing complex object categories, including faces, scenes, words, and bodies [92, 216, 122, 9]. In V4, neurons are selective to properties of visual shape, such as the curvature of visual surfaces and contours [79, 165, 166, 25]. A comprehensive and elegant comparative study showed that population responses to natural object stimuli change systematically from V4 to IT, becoming both more selective to object stimuli and more tolerant to particular physically-realized transformations [189, 76]. But few if any of these studies have captured, functionally, how neurons at these later stages achieve their selectivities and tolerances through computations applied to images, as mediated by earlier stages, like V2. Rather, they have demonstrated tuning along predefined and behaviorally relevant feature dimensions, leaving unsolved the problem of how that selectivity is achieved.

2.2 The second visual area

Among extra-striate areas, the second visual area, V2, has been the most enigmatic. This is perhaps expected given its location in the visual hierarchy. V2 is farther removed from the input than V1, making direct characterization difficult or impossible. But it is also far from the “top” of the hierarchy, so unlike in IT or even V4, its neurons may be only indirectly involved in the representation of particular objects or object features.

V2 is the major recipient of feedforward projections from V1, and depends on V1 for its function [193]. V2 neurons receive functionally diverse V1 inputs, including both simple and complex cells with a variety of receptive field sizes and orientation and spatial frequency preferences [63]. V2 neurons themselves are selective to local

orientation, and are broadly similar to neurons in V1 in their orientation selectivity and spatial and temporal tuning [133], except that they have receptive fields nearly twice the size of those in V1 [81, 62, 199]. Lesions of V2 preserve normal acuity and contrast sensitivity, but impair the discrimination of visual texture patterns [146], which suggests a role for V2 in processing complex stimuli, but fails to constrain the form of its representation.

Most attempts to discover distinct properties of V2 neurons have taken one of two approaches. The first draws on the constructionist agenda described in Chapter 1: if the goal of the visual system is to encode objects, then V2 neurons ought to respond selectively to the elementary features of objects, more complex than local orientation, but less complex than entire objects (or even entire contours and surfaces). Thus, studies have measured responses in V2 to the local angle between line segments [116], the presence of illusory or anomalous contours [171, 129], the curvature of local line elements [102, 103] (examples in Figure 2.4), or “second-order” patterns containing distinct local sub-regions that are similar in luminance but different in texture [64]. These are visual elements with moderate complexity that would seem to be the building blocks of complex shapes and objects. Subpopulations of V2 neurons exhibit unique responses to these stimulus classes, but when V1 and V2 have been directly compared, differences are small [129, 141, 103, 64]. fMRI studies in humans have used adaptation to demonstrate sensitivity to some of these features in extrastriate cortex, including anomalous contours [147] and second-order texture patterns [128, 95]. In both cases, inferred neuronal selectivity was modest (and comparable) in V1 and V2, and stronger in later extrastriate areas, like V4 and IT. A role for higher areas in processing these features was also suggested by a V4 lesion study, which identified significant deficits in processing texture-defined patterns and anomalous contours [234].



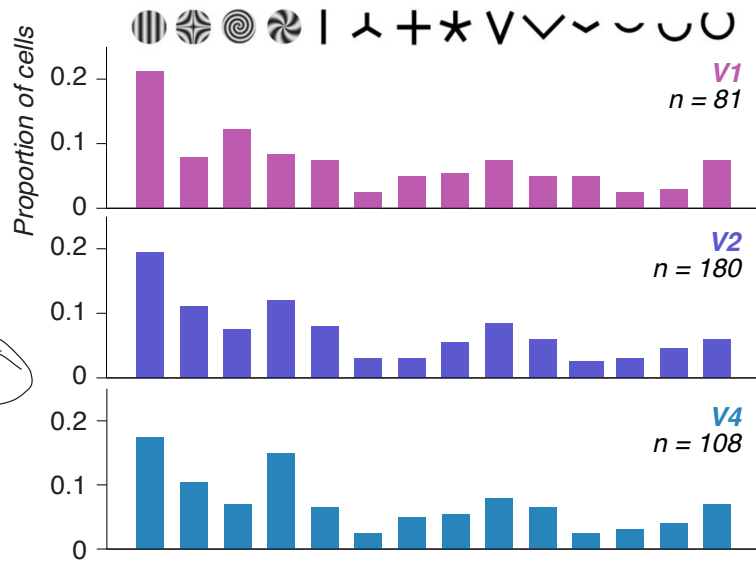
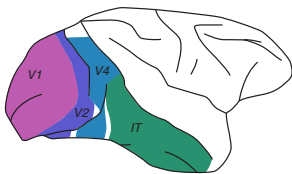
Figure 2.4: Two example V2 neurons that show selectivity to corners (above) and spiral patterns (below) (adapted from [102]).

Why were these approaches unsuccessful in robustly distinguishing V2 from V1? First, if V2 neurons indeed encode complex features, any individual neuron is likely to be selective for a narrow subset, but with tolerances to changes in position (or other dimensions). Without a principled guess, searching for the feature(s) that drive each neuron is difficult given the vast array of possibilities. Curvature and texture-defined borders reflect sensible, but somewhat arbitrary, intuitions, and may not map squarely onto computation in the visual system.

A second concern, particular to the approach of Hegdé and colleagues (e.g. Figure 2.4), is the failure to consider “selectivity” in a feature space with respect to computations performed on the visual input. As a result, it is possible to identify response properties that are superficially interesting, but permit simpler explanations. Consider the family of shapes in Figure 2.4, reproduced in Figure 2.5. Comparative analyses of the distribution of shape preference across many recorded neurons

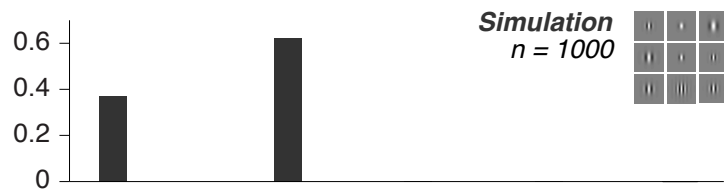
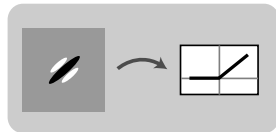
Selectivity to curves, angles, and spirals

(from Hegde and Van Essen, 2007)

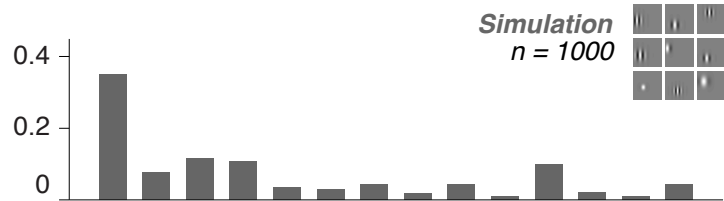
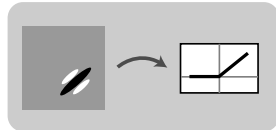


Simulations

Oriented filter,
fixed center



Oriented filter,
random center



Mixture of two
oriented filters

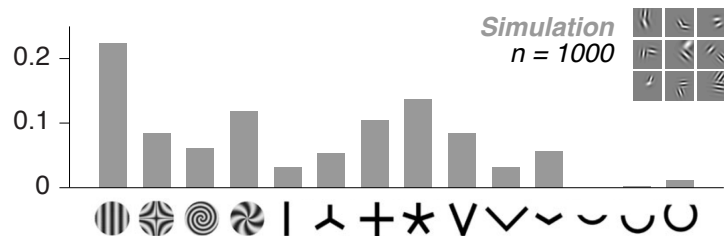
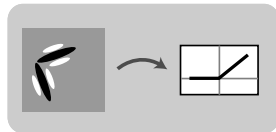


Figure 2.5: Simulations of shape selectivity using LN models (below) capture distributions of shape preference found in areas V1, V2, and V4 (above).

reveal few differences between V1 and V2, or even V4 (Figure 2.5). But in each area, there are significant fractions of neurons that prefer shapes other than gratings or simple bar stimuli – what does that demonstrate about the neuronal response? A simple simulation is instructive (Figure 2.5).⁴ We first compute the outputs of simple V1-like LN model neurons on each stimulus. For each model neuron, we find the stimulus eliciting the largest response, and we report the fraction of model neurons (out of 1000) for which each stimulus was preferred (Figure 2.5). We use only one orientation because, in the experiments, the stimulus set was rotated based on the preferred orientation of the neuron, but we randomize the spatial scale. If the filter is centered precisely on the stimulus, the V1-like model neurons prefer only the grating or the bar, clearly deviating from the measured physiological distributions. If, however, we randomize the position of the filter, we find that nearly 70% of model neurons prefer a stimulus *other* than the simple bar or grating. This arises because of mismatches between the filter center and the stimulus, and any physiologist knows that it is non-trivial to confidently and precisely find the center of a receptive field.

Hegd  et al. tried to control for positional jitter, and reported that it had a minimal effect on selectivity [102]. There also remain differences between the measured distribution and that predicted by the simplest V1-like model. So let us now consider a “simple” V2 model, in which two oriented filters are combined into one. The resulting distribution of preference more closely resembles that of the neurons (e.g. the slightly higher tendency to prefer spirals and orientation combinations) (Figure 2.5). This would seem to suggest that some cells, in each area, linearly combine local orientated inputs. But it could also be a red herring. Randomly combining orientated filters yields a filter with inhomogenous spatial structure; complex forms of spatial inhomogeneity [190, 158, 209], especially when coupled with center-

Figure to the left

Figure to the right

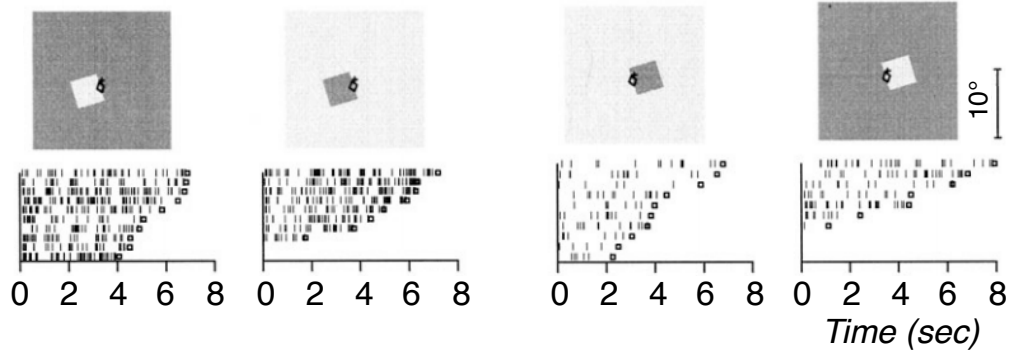


Figure 2.6: Responses of an example V2 neuron that signals border ownership (adapted from [240]). The circle indicates the neuron's receptive field; the figure (the square) is presented to its left or right. Rasters show responses to multiple trials (rows) of different duration.

surround interactions, or cross-orientation suppression [37, 40, 39, 227, 77], could explain the physiological distributions. This brand of exercise is an existence proof at best, limited by the vast array of possible models. Perhaps more deflating is that the stimulus set fails to differentiate V2 from V1, so there may be little worth trying to explain.

One particularly interesting class of stimuli derived from intuitions about surfaces and shapes have more successfully differentiated V2 neurons from those in V1. In several experiments, Von der Heydt and colleagues have measured the responses of V2 neurons to edges induced by presenting a square (“the figure”) on a uniform background (“the ground”) (Figure 2.6). Changing the luminance of the square or the background yields edges that can be specified either by the luminance on either side of the edge, or whether the left (or right) of the edge belongs to the figure (or the ground). A minority of neurons in V2 robustly signal the location of the figure relative to the edge, but are invariant to the edge's polarity. This “border-ownership”

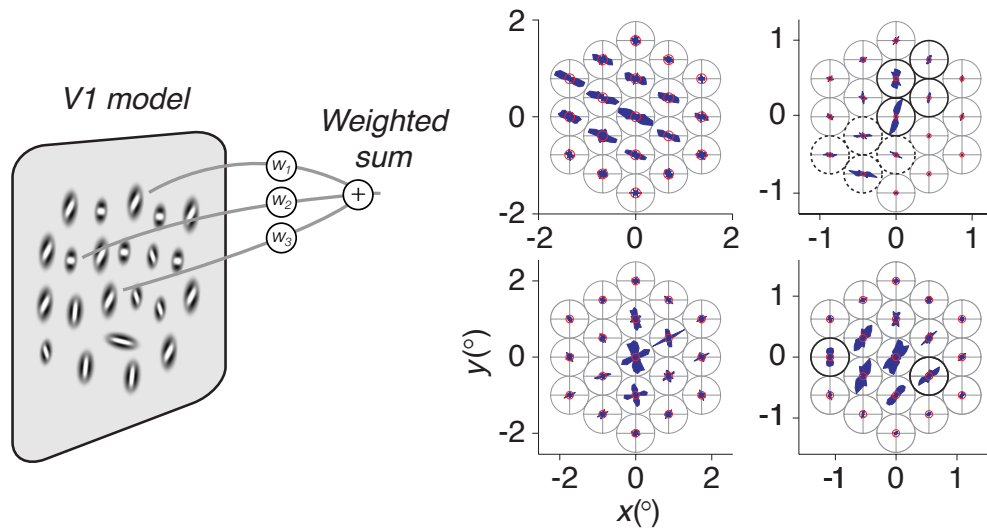


Figure 2.7: *Spatial diversity of orientation preference in an example V2 neuron (adapted from [8]). The underlying model (left) is that V2 neurons linearly combine V1-like orientation-tuned filters at different spatial locations.*

property may be related to feedforward processing, or attentional modulation, or some combination of the two. Some border-related signals can be computed through feed-forward or lateral signaling [48, 86], but border-ownership responses are only present in awake animals, interact with or depend on attentional affects [176, 67], and may reflect feedback from higher areas like V4 that are involved with processing visual surfaces.⁵

A complementary approach to V2 emphasizes that V2 receives its primary inputs from V1, and may perform relatively simple computations on that input. Two studies, specifically, have analyzed V2 receptive fields in terms of V1 inputs, by presenting random mixtures of local oriented elements, and characterizing V2 neurons as linearly combining those elements [8, 225]. For the most part, V2 neurons characterized this way are similar to neurons in V1, exhibiting a global orientation preference across the receptive field. But some neurons exhibit diversity in orientation preference across space that suggests sensitivity to local curvature or other

complex features [8] (Figure 2.7).

These efforts focus on a simple class of combination rules. As will be discussed more below, linear combinations of rectified V1 inputs are limited in their representational power. Furthermore, stimuli based on this combination rule resemble filtered, oriented noise, lacking many of the complex features of natural images. If V2 represents such features, distinguishing it from V1 ought to require stimuli that contain them. A final study by Willmore et al. (2010) combined the above modeling approach with natural stimulation; they measured the responses of V2 neurons to natural photographs, and modeled the responses using a feedforward computation with a set of V1-like filter outputs followed by linear combination [237]. They found that V2 neurons exhibited more tuned suppression than V1 neurons, but otherwise identified few notable differences. The tuning they recovered is also complicated by the uncontrolled statistical properties of natural images. When selectivity is itself complex, it can be difficult, when fitting a feedforward cascade model, to distinguish between properties that reflect neuronal selectivity and properties that reflect dependencies within the stimulus [187].

In summary, over decades of physiological investigation, progress in V2 has faced experimental and conceptual difficulties. If V2 neurons are encoding complex features, it is impossible to sample all features in order to find the ones that an individual V2 neuron cares about. Simple combinations of local oriented elements do not elicit differential or selective responses in most V2 neurons. Assessing tuning along more complex feature dimensions like contours and curvature yields results that are difficult to interpret in terms of computations applied to the visual input, because such efforts are motivated more by constructionist intuitions about the building blocks of shapes. Natural images of scenes and objects may contain the features that V2 neurons care about, but are difficult to control experimentally, and

complex features appear in natural photographs of scenes and objects only sparsely, so searching for the one image that drives a V2 neuron is likely searching for a needle in a haystack. Experiments with natural images have also been wedded to simple models based on linear combination of V1 input, which may be insufficient for representing more complex natural image features, as discussed in the next section.

2.3 Computation and theory

In parallel to physiological investigations, theorists have tried to identify normative, computational principles of visual coding. One approach is based on the theory of *efficient coding*, according to which the visual system is optimized to represent natural signals [13, 202]. Models that aim to efficiently represent natural images have revealed key statistical properties and have provided normative explanations of coding in the early visual system.

Many of these models adopt a “generative framework”, representing an image as a linear combination of basis elements. Ignoring color, we denote a vectorized grayscale image as \vec{y} , and represent it as a weighted sum of basis vectors \vec{b}_i :

$$\vec{y} = \sum_i \vec{b}_i x_i \quad (2.1)$$

The weight variables x_i encode the relative contribution of each basis, and different images will induce different values for the x_i s. What is the best basis for natural images? The answer depends on the objective. One solution is based on principal components analysis (PCA), which identifies basis directions that capture maximal variance in the input distribution and for which the coefficients x_i are pairwise decorrelated. Because adjacent pixels in natural images are correlated, and due to

the translation invariance of images, the PCA basis for natural images is the Fourier basis, with larger weights on lower frequency components. This recovers the well described “ $1/f$ ” property of natural images [212, 185], according to which spectral power falls as a function of spatial frequency.

More complicated models seek a linear basis that maximizes the sparsity of the inferred weight variables [163] or makes them as independent as possible [15, 16]. The two optimization procedures are related, and both yield bases that are localized and oriented rather than the large sinusoidal gratings recovered by PCA, and comparable to receptive fields of neurons in primary visual cortex.

Much has been made of this correspondence [221], but it is complicated to interpret. First, it is not clear why a simple linear model operating directly on image pixels should recover receptive field properties fairly deep into the visual system, as opposed to learning filters matched, for example, to the retina. Put another way, why are retinal ganglion cell receptive fields not orientation tuned? Second, these linear models, on their own, fail to reproduce much of the important structure in natural images: random combinations of the basis elements, for PCA, ICA, and sparse coding, contain few of the structures and textures found in natural images, and instead tend to resemble different kinds of clouds [202, 123].

Furthermore, *independent* components analysis fails to recover components that are independent. When applied to natural signals, the outputs of the resulting local oriented filters exhibit substantial residual dependencies. This was first described in the domain of wavelet filters [200], and used to substantially improve image compression [29]. An example of such dependencies is illustrated in Figure 2.8. We took an image of a natural texture, and computed at each location in the image the output of two filters, tuned to vertical orientation, at two different spatial frequencies. Each column of the joint filter histograms (in the middle) shows the

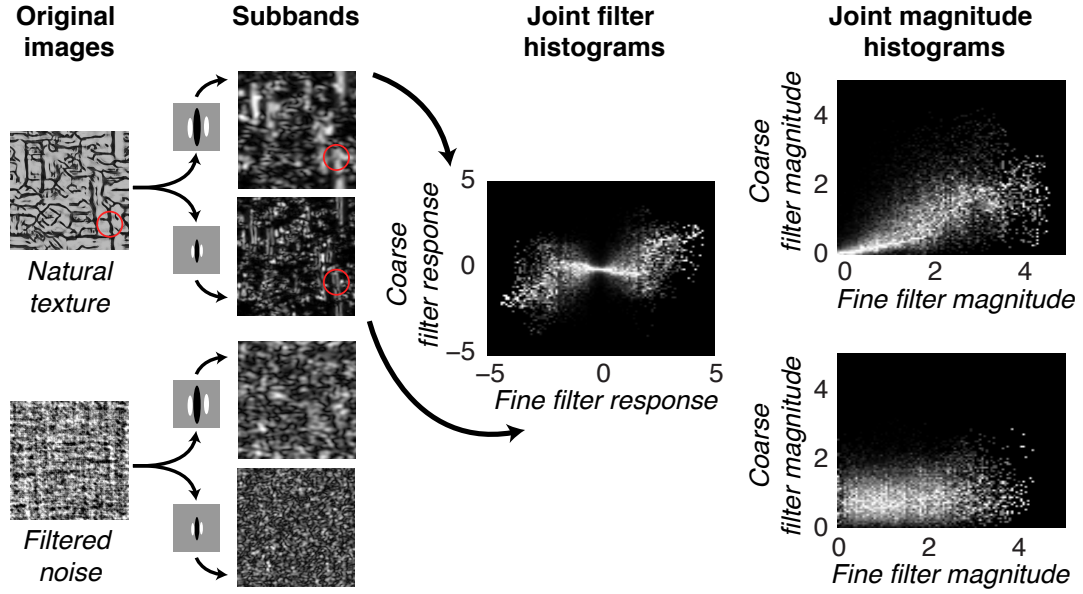


Figure 2.8: The outputs of local filters exhibit strong dependencies when applied to a natural texture image, but not a spectrally-matched image with random phase. When an image is processed with oriented-tuned filters at two spatial frequencies (coarse and fine) (left), the conditional variance of one filter output depends on the value of the other (middle). The relationship becomes a correlation when using a phase-invariant measure of magnitude (right).

distribution of the coarse-scale filter response conditional on fixed values of the fine-scale filter response. Although there is only weak linear correlation between the two filter responses, the “bow-tie” pattern reveals that the variance of the conditional distribution depends on the magnitude of the fine-scale filter response. This can be converted to a simple correlation by working with the magnitudes of the filter responses. We specifically use a phase-invariant measure of magnitude; we evaluate the output of a pair of filters in quadrature pair, tuned to the same orientation and frequency, one symmetric, one anti-symmetric, and compute the square root of the sum of the squared responses. Relating magnitudes of the two filters clearly reveals a correlation (right-most plot in Figure 2.8). This form of dependency occurs for a

variety of filter combinations. Many pairwise comparisons yield positive correlations (e.g. across positions or across scales), though correlations across orientation can be negative when, for example, structure at one orientation tends to occur precisely at locations that do not contain an orthogonal orientation. Importantly, these correlations are properties of natural signals, because they do not appear in images of phase-randomized spectral noise (an example is shown in the bottom row of Figure 2.8).

An intuition for these dependencies is that natural images tend to contain regions with little structure or oriented energy, interdigitated with isolated structures, such as extended and aligned curves or edges, that produce energy across multiple scales at the same spatial location, and also across multiple locations, if the structure is spatially extended. Examine the subbands in Figure 2.8, for example, which show magnitudes at each location in the image for the two filters, fine and coarse. For the natural image, at any given location, large energy for one filter co-occurs with large energy in the other (the red circle depicts one such location). Similar correspondences occur less frequently in spectrally-matched noise.

Three notable, and related, models have incorporated these magnitude dependencies. Schwartz and Simoncelli (2001) argued that if the magnitude of a linear filter is correlated with the magnitude of other nearby filters, the dependency can be eliminated through divisive normalization [97], by dividing the magnitude of each filter by a local estimate of the magnitudes of the others [195]. Specifically, Schwartz and Simoncelli modeled each filter response r_i by projecting the image onto the filter, squaring, and dividing by a weighted sum of squared responses to other filters:

$$r_i = \frac{(\vec{y}^T \vec{b}_i)^2}{\sum_j w_{ij} (\vec{y}^T \vec{b}_j)^2 + c} \quad (2.2)$$

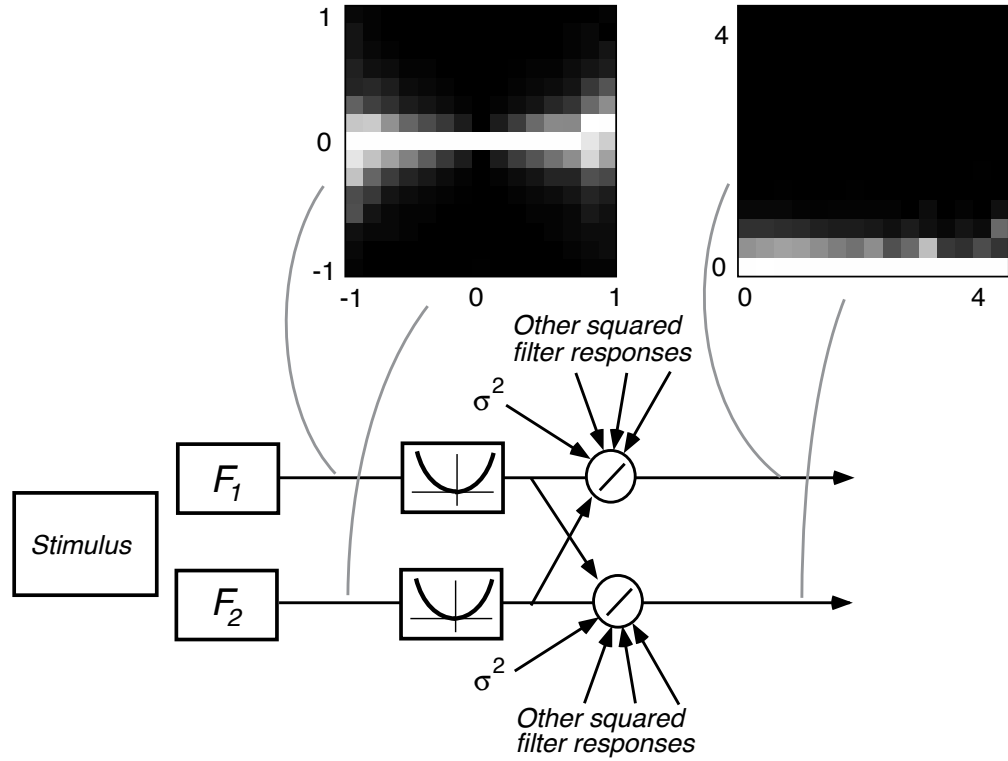


Figure 2.9: Normalizing filter responses by the responses of other neighboring filters eliminates the statistical dependency depicted in Figure 2.8 (adapted from [195])

The weights w_{ij} must be positive to avoid a 0 in the denominator. Optimizing the model involved learning a set of weights w_{ij} such that, when computed on natural images, the distribution of the responses were as close as possible to a distribution, a factorized gaussian, in which the responses were approximately independent. After weight learning, applying the computation to natural signals yielded independent responses, as shown in Figure 2.9. The responses of the learned normalized filters to simple grating stimuli also reproduced non-linear response properties found in V1, such as cross-orientation and surround suppression [195].

Karklin and Lewicki (2008) proposed a different account of magnitude dependencies, working in the framework of efficient coding, but extending those methods

to capture dependencies across filter responses [123]. They specifically modeled image patches \vec{y} as arising from a multivariate Gaussian distribution,

$$P(\vec{y}|\vec{x}) = N(\mu, C) \quad (2.3)$$

where the logarithm of the covariance matrix C is determined by a population of model neurons x_j that weight the relative correlation of different basis filters

$$\log C = \sum_{jk} x_j w_{jk} \vec{b}_j \vec{b}_k^T \quad (2.4)$$

The x_j are the activations of model neurons, and the vectors \vec{b}_k are simple oriented V1-like filters; the matrix logarithm is used to define a basis for the covariance matrix because the inverse operation, the matrix exponential, always yields a positive semi-definite covariance matrix. The key of the model is that the neuronal “activations” x_j tie together weights w_{jk} that control the degree to which combinations of filters \vec{b}_k occur with similar magnitude. Karklin and Lewicki “trained” the model on natural images by finding parameters to maximize the likelihood of an image ensemble, and learned sets of w_{jk} , for each of several model neurons x_j , that reflected typical filter co-occurrences.

One of their learned model neurons is shown in Figure 2.10. It has, in its weights to different filters, encoded a complex mix of dependencies across two orientations and two different locations. In the color plot, each of the lines correspond to a filter \vec{b}_k – the orientation and width of the line depict its orientation and spatial frequency preference. The colors, red to blue, indicate positive or negative weights. The panels on the right show a subset of filters that have positive or negative weights. We can think of this model neuron as having encoded, implicitly in its

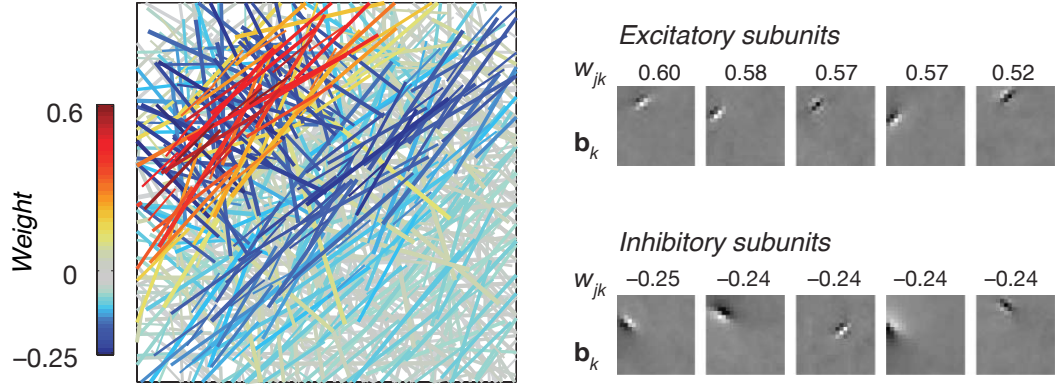


Figure 2.10: Training a statistical model of images yields model neurons that reflect magnitude dependencies by differentially weighting V1-like filters (adapted from [123]).

learned weights, the fact that a combination of orientations occurs frequently in natural images. When presented with an image, this neuron’s *response* depends (approximately) on a weighted sum of squared projections onto each of the filters. As such, these learned “subunits” are analogous to the multiple filters used to describe V1 complex cells [188], and indeed, Karklin and Lewicki found that their model neurons reproduced many nonlinear responses of V1 cells, including position invariance, as well as cross-orientation and surround suppression.

The two models – the normalization model, and the covariance model – are different in their architecture and computations. Fully reconciling them is beyond our scope, but several key similarities and differences bear mentioning. The learned parameters in both models reflect magnitude dependencies in natural images, and the learned dependencies predict interactive effects observed in the responses of more complex V1 neurons. However, the model neurons in each case signal dependencies *indirectly*, through contextual effects. This is particularly clear in Karklin and Lewicki’s model, where the response of each model neuron is (approximately)

a weighted sum of filter magnitudes, notated for two filters as

$$r_j \approx w_{j1}(\vec{b}_1^T \vec{y})^2 + w_{j2}(\vec{b}_2^T \vec{y})^2 \quad (2.5)$$

The learned weights reflect a direction in magnitude space for which the two magnitudes are correlated. But if such a neuron is presented with a distribution of stimuli exhibiting that correlation, compared to a stimulus lacking it, the neuron will not signal, with its mean response, which of the two types of images it sees. Rather, the model neuron's *variance* will be higher for the stimuli exhibiting the correlations, because those stimuli fall along the axis in magnitude space implicitly represented by the neuron.

To explicitly signal the presence or absence of a correlation, we would need to compare the squared projection onto one direction with the squared projection onto another direction. From Equation 2.5, if the direction of correlation is $[1, 1]$, the presence of the correlation can be signaled by computing,

$$r'_j = \left((\vec{b}_1^T \vec{y})^2 + (\vec{b}_2^T \vec{y})^2 \right)^2 - \left((\vec{b}_1^T \vec{y})^2 - (\vec{b}_2^T \vec{y})^2 \right)^2 \quad (2.6)$$

More generally, the activity of the model neurons signal particular image distributions only in their *population* response. Any explicit representation requires downstream computations that compare the outputs of different model neurons, as in Equation 2.6. The above observation also echoes the finding that a cascade of linear filtering, squaring, and linear filtering can explicitly capture correlations, as discussed by Adelson and Bergen [2] when relating the correlational Reichardt model for motion detection with the energy model. Finally, these mechanisms could be appropriately normalized [97] to ensure that they signal correlations rather than overall contrast.

A very different approach to modeling dependencies comes from the domain of analyzing and synthesizing homogenous natural textures. There is no rigid definition of texture, but they are typically considered a subclass of natural images, spatially homogenous with repeated elements, and therefore more tractable for statistical modeling. A longstanding problem is to define a set of computations that, when applied to a texture, would yield a set of outputs (statistics) that sufficiently capture its perceptually important properties. Julesz [118] formalized this goal with the conjecture that there exist a set of statistics such that any two images matched with respect to those statistics (averaged over space) would appear identical. Julesz described, but subsequently disproved [34, 33, 121], the sufficiency of a simple set of pixel statistics, and abandoned the theory.

But since Julesz, there have been impressive demonstrations of statistical parameters that capture textural properties. Bergen (1994), and later Heeger and Bergen (1996), showed that a set of histogram statistics, computed on each of several subbands of a multi-scale [17] or multi-scale and multi-oriented [101] decomposition, captured important properties of visual texture (such decompositions will be discussed in more detail in Chapter 3). Inspired by the Julesz conjecture, they demonstrated the success of their model with a *synthesis-by-analysis* approach, in which statistics are analyzed on an original image, and then iteratively imposed on a new image until it matches the original. If the resulting synthetic image does or does not look like the original (for many different originals), it demonstrates the success or failure of the model.

Indeed, failures of the Heeger and Bergen synthesis method demonstrate the importance of magnitude dependencies in visual texture. Consider the two images in Figure 2.11. The images on the left are originals, and the images in the middle are matched to the marginal statistics of each subband of a multi-scale decomposition

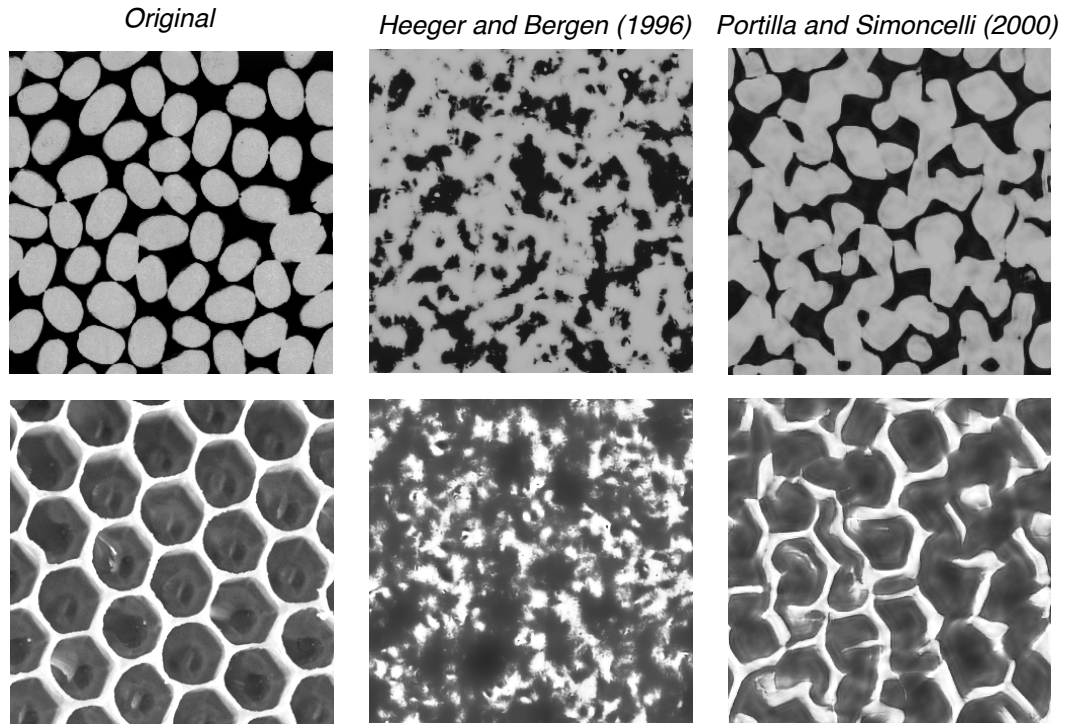


Figure 2.11: *Two different synthesis-by-analysis texture models yield different results when applied to the same original image. The Heeger and Bergen synthesis [101] captures marginal statistics, and the Portilla and Simoncelli synthesis [175] additionally captures extended contours and periodicity.*

(specifically the mean, variance, skew, and kurtosis).⁶ Although they share features of the original, like the relative distribution of uniform patches and highlights, they lack the extended contours and periodicity of the original. Portilla and Simoncelli (2000) showed that such features can be captured by explicitly representing the magnitude dependencies discussed above. They extended Heeger and Bergen’s approach by developing a more general parametric texture model that incorporated, in addition to marginal statistics, an explicit representation of magnitude dependencies across different scales, orientations, and positions. They captured these dependencies by computing pairwise products of magnitudes – e.g. of two oriented filters at

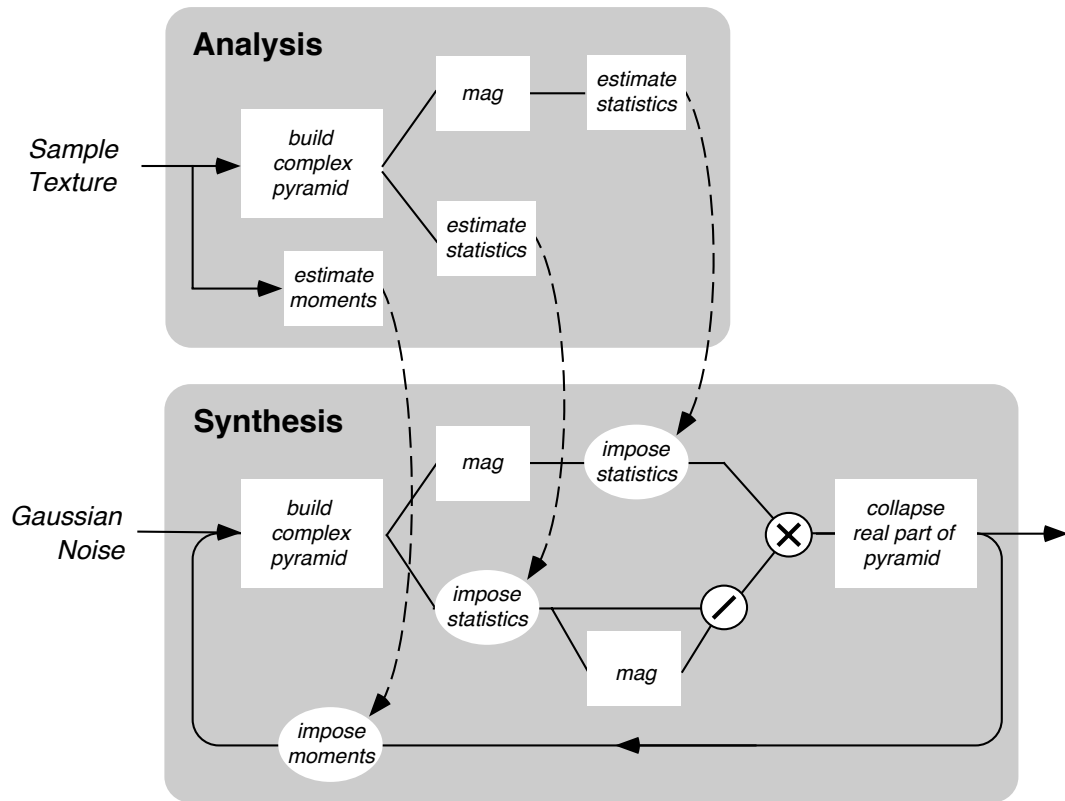


Figure 2.12: The synthesis-by-analysis algorithm used by Portilla and Simoncelli to measure statistical properties of texture images and generate new statistically-matched images by imposing the same properties on an image of Gaussian noise (adapted from [175]).

the same location – and averaging the products over all spatial locations; formally, these are correlations.⁷ They also developed a method for synthesizing textures with these more complex parameters, depicted in Figure 2.12.

Incorporating these correlations yielded synthetic textures that more faithfully capture perceptually striking and naturalistic features of originals, like the contours and periodicity not captured by the Heeger and Bergen model (Figure 2.11). Additional examples of syntheses from the Portilla and Simoncelli model are shown in Figures 2.13 and 2.14, including synthesis “failures” when the model is applied to

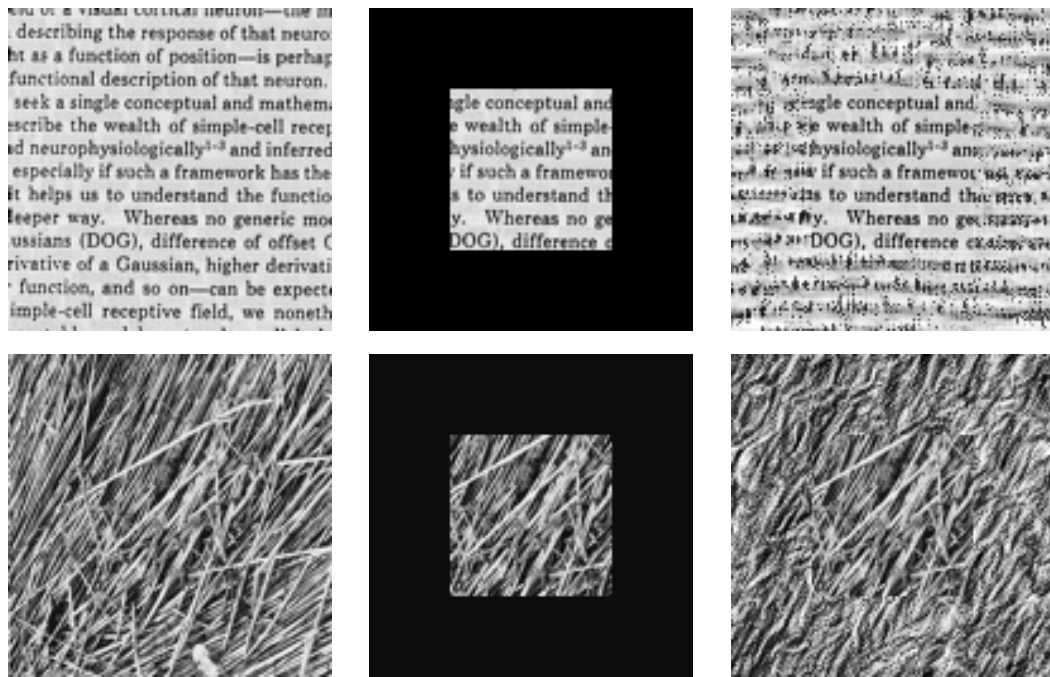


Figure 2.13: Example textures synthesized by the Portilla and Simoncelli model; full original images are shown in left; textures are synthesized (right) to fill in the boundary around the central image patch (adapted from [175]).

non-texture objects; rather than reproduce the object, the model appears to “paint” a new texture with bits and pieces of the original. The most striking property of the synthetic images generated from the Portilla and Simoncelli model is that they contain some of the “features” found in natural images, including features that are presumably the constituent pieces of objects. But the model does not explicitly represent such features, nor are their locations or forms clearly specified. Rather, the model represents them indirectly through its parameters; this is frustrating in that we cannot point to a parameter in the model and say it encodes the presence of a corner-feature at a particular location, but more faithfully captures the murky texture-ness of most features and objects in real images, and perhaps, the way in which the brain represents such features.



Figure 2.14: *Example textures synthesized from inhomogenous natural photographs. Syntheses have similar textural properties, but lack the global organization, of the originals (adapted from [175]).*

The textures generated by the Portilla and Simoncelli model are also far more naturalistic looking than state-of-the-art machine learning models of image structure, even Karklin and Lewicki’s [123]. Two aspects are important to the discrepancy. First, unlike Karklin and Lewicki’s model, which learned its parameters from the statistical structure of natural images, the Portilla and Simoncelli model was designed “heuristically” to reproduce, through synthesis, “natural looking” images. A related point is that statistical learning is slow, and is only tractable for small patches, whereas the Portilla and Simoncelli model operates, and captures statistical dependencies, across large images. There thus remains a substantial gap between rigorous statistical models of image structure and synthesis-by-analysis algorithms for generating naturalistic stimuli.

2.4 Perception

Perceptual investigations into intermediate-level representations have focused largely on the detection and segregation of visual texture patterns (reviewed comprehensively by [127]), and computational mechanisms that might support those tasks. Approaches have largely fallen into two categories, emphasizing either how observers segregate textural subregions, or how observers use statistical properties to discriminate and represent textural properties.

Studies of texture segregation focus on how observers identify boundaries between subregions of an image that are similar in luminance but differ in their texture. These boundaries cannot be detected using linear filters because the luminance in each region is similar. But humans can readily identify the orientation and spatial frequency of such boundaries in so-called “second-order patterns” [18, 88, 124, 222, 43, 126]. A widely used model for capturing this behavior is the filter-rectify-filter, or “backpocket” model. The model consists of an initial stage of linear filtering (e.g. with oriented filters), followed by rectification or another point-wise nonlinearity (analogous to the measures of magnitude discussed above), and a second stage of linear filtering. For example, to detect a vertical boundary between regions of vertical and horizontal oriented elements, an image can be processed first separately with vertical and horizontal filters, and then a “vertical” second-stage “filter” can be used to detect the boundary (Figure 2.15).

Inspired by early successes in spatial vision characterizing orientation and spatial frequency-tuned channels, and linking them to physiological mechanisms of early vision (reviewed in [58, 87, 89]), there have been several efforts to psychophysically characterize components of the FRF model (reviewed in [127]). But it has not been straightforward to map its computations onto intermediate stages of the primate

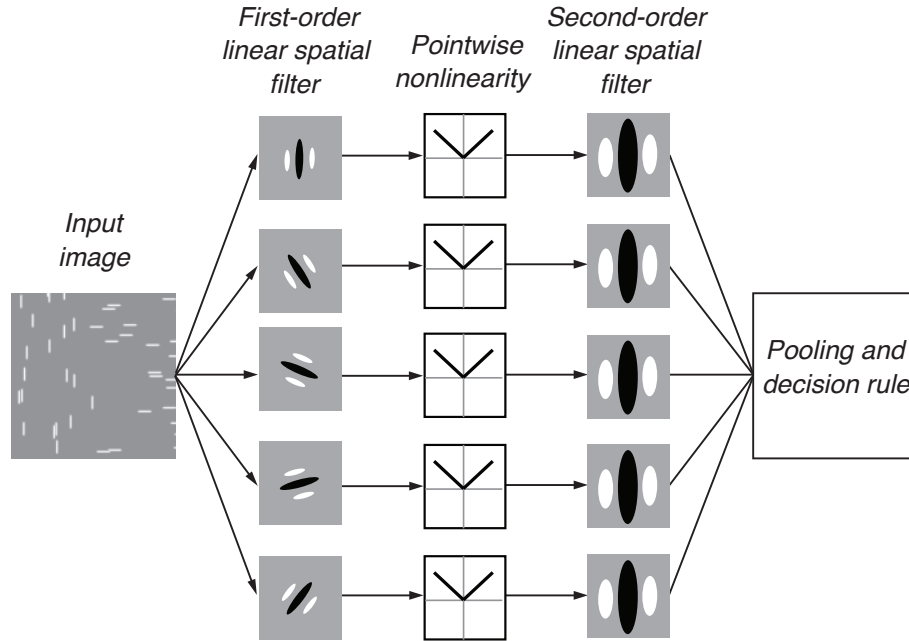


Figure 2.15: *The filter-rectify-filter, or “backpocket” model, proposed to capture sensitivity to second-order patterns, in which boundaries are formed between subregions that contain different textures, like the vertical boundary between regions of vertical and horizontal texture shown here (adapted from [127]).*

visual system. As discussed above, the majority of neurons in V1 and V2 are sensitive to luminance-defined rather than texture-defined patterns, and selectivity to texture-defined patterns is comparable between V1 and V2 [64]. fMRI adaptation experiments have revealed selectivity to both second-order orientation [128] and spatial frequency [95]; in both cases the effects are comparable in V1 and V2, and at least for orientation, effects are more pronounced in higher areas. Furthermore, in so far the FRF model computes linear combinations of V1 responses, it cannot explicitly capture the magnitude dependencies that are important for capturing naturalistic features; recent efforts to incorporate normalization in the second stage of the FRF model, however, may begin to bridge that gap [231, 95].

A second tradition in the study of texture perception has focused on statistics that capture the visual appearance of texture. As discussed above, this concept was introduced and formalized by Julesz (1962), who conjectured that there exist a set of statistics such that images matched for those statistics appear identical [118]. After finding counterexamples to his own theory [34, 33, 121], Julesz instead focused on “textons”, elementary features and feature combinations that can be used to build textures and texture regions [120, 19]. But like the curves and angles of Hegdé et al. (Figure 2.5), textons are difficult to interpret in terms of computations that might be performed on the visual input. Furthermore, unlike approaches based on texture synthesis-by-analysis, the construction of a texture out of textons is notably divorced from the computations that might represent its properties. As Adelson describes, textons embody the turn towards individuation “Julesz speaks of his textons as the quarks of vision” [1]. Related efforts in visual search and cognitive psychology have emphasized the detection of texton-like features or “feature conjunctions” [215]. These studies have identified some perceptual and cognitive mechanisms of textural and attentional processing, but they are similarly difficult to relate to the computations performed by the visual system on naturally occurring signals [184].

The Portilla and Simoncelli model (2000) is notable in that it both represents rich features of natural textures, and its computations could plausibly be mapped onto as yet uncharacterized stages of visual processing beyond V1 [175]. But the model has proved difficult to probe experimentally, partly due to the complexity of its parameterization. Balas (2006) expanded the “heuristic” perceptual validation of the original model by generating synthetic textures after lesioning subsets of parameters, and using a perceptual 3AFC “oddity” task to assess the relative necessity and sufficiency of different parameter groups [12]. The marginal statistics and magnitude correlations were most clearly necessary, but the marginal statistics on

their own were not sufficient, whereas the magnitude correlations were, at least for some texture categories. Balas emphasized that the redundancy of the parameter groups leaves the effects of lesioning difficult to interpret, because imposing other statistics may inadvertently impose the statistic that has been “lesioned”. Imposing marginal statistics, for example, inadvertently induces some magnitude correlations. This issue of redundancy will become important when interpreting the physiological experiments described in Chapters 3 and 4.

A final, notable example of using the Portilla and Simoncelli model to study perception came from Balas et al. (2009), who established a connection between texture processing, the Portilla and Simoncelli model, and the phenomenon of visual crowding. Crowding is a breakdown in object recognition that occurs when closely spaced but non-overlapping flankers hinder the identification of a peripherally viewed target object, such as a letter [21, 168]. Bouma (1970) showed that the spacing required to escape crowding, called the critical spacing, is proportional to eccentricity (distance from fixation). During crowding, the local features of the target remain the same and the target remains visible, but it becomes unrecognizable, appearing as a dynamically changing texture of elementary features that lack identity [164, 168, 73]. Balas et al. noted that, phenomenologically, the textural scrambling that occurs when applying the Portilla and Simoncelli texture synthesis to objects (e.g. Figure 2.14) resembles the jumbled appearance of crowded stimuli. For combinations of target objects and distractors typically used in crowding experiments, they showed that difficulty recognizing objects in synthetic textures predicted difficulty recognizing the same objects when viewed peripherally. They hypothesized that a representation like that proposed by Portilla and Simoncelli, operating within local regions across the visual field, could provide a general account of crowding phenomena, although they did not implement or test such a model. Thus, it remains

unclear whether the representation they proposed can account for the dependence of crowding on spacing and eccentricity – its defining property – as well as how such a representation might relate to the physiology of the visual system.

2.5 Outlook

Despite identifying a minority of V2 neurons with unique response properties, existing physiological studies of V2 suggest the need for a new approach, if we hope to identify basic properties of image encoding that robustly distinguish neurons in V2 from those in V1.

The above considerations suggest a recipe for progress: we need experimental stimuli that are more complicated than local combinations of orientation, but better controlled than photographs of scenes and objects. The Portilla and Simoncelli texture synthesis approach will provide us with a vehicle for presenting controlled homogenous stimuli with naturalistic features that reflect complex non-linear combinations of V1 responses. As we will show, its image synthesis can be adapted and manipulated for generating families of experimental stimuli for both physiological (Chapter 3) and perceptual studies (Chapter 4), including a reformulation of its computations within local regions that tile the visual field (Chapter 5).

Crucially, emphasizing textures in the study of V2, particularly stochastic textures synthesized from a statistical model, represents a departure from the guiding principle of individuation, in so far as the features of textures, while occasionally object-like, are not obviously useful for constructing the shapes of objects.

More generally, the scientific approach described in the following Chapters – characterizing neurons by generating stimuli that reflect hypotheses about their computations, rather than fitting models to characterize responses – is unusual

in sensory neuroscience. In so far as this approach is successful, we hope that it will stimulate similar investigation into intermediate stages of computation in other sensory systems.

Notes

⁴The idea for this simulation, and example code, was provided to me by Geoffrey Boynton.

⁵Rudiger von der Heydt, personal communication

⁶Heeger and Bergen (1996) matched the complete histogram of each pyramid subband, but in practice, similar syntheses are achieved for most textures when matching the full histograms or just the first four moments

⁷Even more formally, they are covariances, because they are not normalized by variance. We continue to call them correlations throughout this thesis for intuition, but note that including normalization by variance may improve synthesis results and stability, and help remove some redundancy across parameter groups [145]

Chapter 3

Responses to naturalistic stimuli differentiate V2 from V1

3.1 Introduction

V2 is the largest extrastriate cortical visual area in primates, and its responses depend on feedforward input from V1 [193, 204]. Neurons in V2 might combine and elaborate signals from V1 to encode image features that V1 does not, but most efforts to differentiate the responses of cells in the two areas have been unsuccessful. As discussed in Chapter 2, most efforts have measured the responses of V2 neurons to artificial stimuli containing features thought to form the building blocks of objects, including gratings, angles, curvature, anomalous contours, and second-order patterns. In most cases, V2 neuronal selectivity for these image attributes is qualitatively and quantitatively similar to that of V1 [171, 116, 225, 129, 141, 102, 103, 64]. Other efforts have measured V2 responses to combinations of V1-like stimuli (e.g. local oriented gratings), but have emphasized simple linear combination rules, and used correspondingly simple stimuli lacking naturalistic features [8, 225]. Some of

the most robust documented differences in the responses of the two areas reflect the signaling of border ownership [240], which occurs in a minority of cells, and may reflect attentional modulation [176, 67]. Finding visual features that reliably distinguish the responses of V1 and V2 neurons has therefore proved elusive.

In this Chapter, we characterize response differences between V1 and V2 using stimuli containing complex features found in naturally occurring images. Rather than using actual photographs [237], in which these features are uncontrolled, we focused on homogenous natural textures, and captured their properties by measuring their higher-order image statistics, and then constructing textures containing statistically similar features. Specifically, we first computed the rectified responses of a set of V1 simple and complex cell-like filters tuned to different positions, orientations, and spatial frequencies. We then computed correlations of these responses across different orientations, frequencies, and positions. As described in Chapter 2, these correlations represent the properties of local image features like curvature, sharpness, and periodicity, with a sequence of “canonical” computations – linear filtering, rectifying nonlinearities, and products – that have also been explored in the context of hierarchical models of cortical pattern recognition [180, 123].

Most V2 cells responded more vigorously to these stimuli than to matched control stimuli lacking naturalistic structure, while V1 cells did not. Parallel fMRI measurements in humans revealed differences in V1 and V2 responses to the same textures that were consistent with the neuronal measurements. These results reveal a novel and particular role for V2 in the representation of natural image structure. Finally, neuronal and fMRI responses in V2 depended reliably and similarly on the particular texture types used, a signature which we will exploit in Chapter 4 to link neuronal responses in V2 to perceptual behavior.

3.2 Synthesis of naturalistic stimuli

We generated naturalistic texture stimuli using a synthesis-by-analysis algorithm [98, 175]. We used a two-stage computation to capture a set of higher-order image statistics on an original image, and then used an iterative procedure to transform samples of Gaussian noise into new images with the same statistical properties. Here, we describe the details of the two-stage computation and the synthesis procedure.

3.2.1 Multi-scale multi-orientation decomposition

Images are first partitioned into subbands by convolving with a bank of filters tuned to different orientations and spatial frequencies (Figure 3.1). We use the steerable pyramid, which has several advantages over common alternatives (e.g., Gabor filters, orthogonal wavelets), including direct reconstruction properties (beneficial for synthesis), translation invariance within subbands, and rotation invariance across orientation bands [175]. A matlab implementation is available at <http://www.cns.nyu.edu/~lcv/software.php>. The filters are directional third derivatives of a lowpass kernel, and are spatially localized, oriented, anti-symmetric, and roughly one octave in spatial frequency bandwidth. The pyramid decomposition can be implemented through convolution (in the spatial domain) with these filters, or equivalently, through point-wise multiplication (in the Fourier domain) so as to carve up the two-dimensional Fourier plane into “wedges” corresponding to different ranges of both spatial frequency and orientation (Figure 3.1). We use a set of 16 filters rotated and dilated to cover four orientations and four scales (more than the two scales shown in Figure 3.1). This decomposition also yields unoriented low-pass and high-pass residuals, which we largely ignore for our purposes. We also include a set of even-symmetric filters of identical Fourier amplitude (i.e., Hilbert

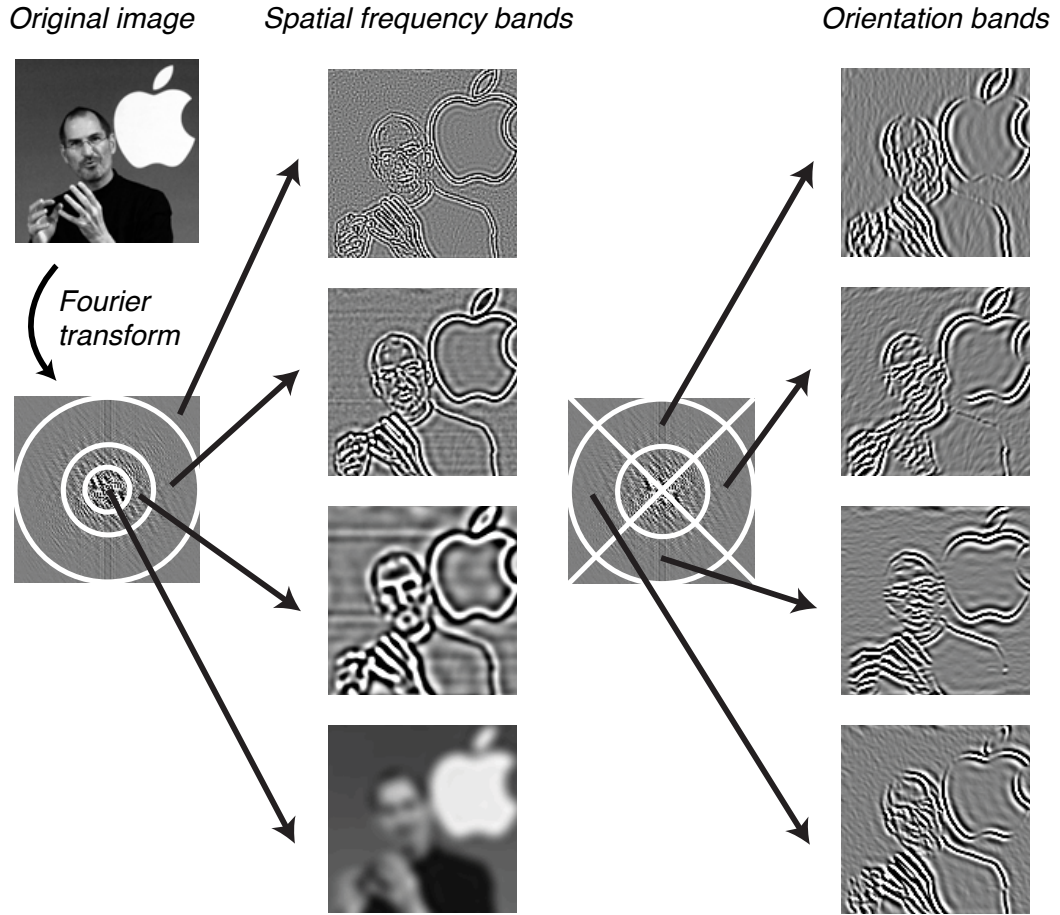


Figure 3.1: The steerable pyramid decomposes an image into subbands at different spatial frequencies (left) and different orientations at a given frequency (right). The decomposition can be expressed as a convolution with bandpassed filters, or as a carving up of the Fourier domain, as shown here.

transforms) [175]. Each subband is subsampled at its associated Nyquist frequency, so that filter spacing is proportional to size. We write the n th subband as $x_n(i, j)$, a two-dimensional array containing the complex-valued responses. We also use vector notation \vec{x}_n . The real part of the subband is denoted $s_n(i, j)$ and represents the responses of V1 simple cells. The square root of the sum of the squared responses of symmetric and anti-symmetric filters yields a phase-invariant measure of local

magnitude, denoted $e_n(i, j)$, and represents responses of V1 complex cells [175, 2].

3.2.2 Second stage computation

In the second stage of computation, we evaluate products of pairs of V1 responses tuned to neighboring orientations, scales, and positions. The particular set of combinations are based directly on those developed in Portilla and Simoncelli [175], and are motivated by statistical dependencies found in natural images [195].

(1) Products of responses at nearby spatial locations (i.e., autocorrelations) for both simple cells (capturing spectral features such as periodicity) and complex cells (capturing spatially displaced occurrences of similarly oriented features). Simple cell autocorrelations are given by,

$$A(n, k, l) = \sum (s_n(i, j) - \mu(\vec{s}_n)) (s_n(i + k, j + l) - \mu(\vec{s}_n)) \quad (3.1)$$

Where (k, l) specifies the spatial displacement (in horizontal and vertical directions), the summation is over (i, j) , and $\mu(\vec{s}_n)$ is the mean,

$$\mu(\vec{s}_n) = \sum s_n(i, j) \quad (3.2)$$

Complex cell autocorrelations are similarly given by,

$$B(n, k, l) = \sum (e_n(i, j) - \mu(\vec{e}_n)) (e_n(i + k, j + l) - \mu(\vec{e}_n)) \quad (3.3)$$

We include spatial displacements in the range $(-3 \leq k \leq 3, -3 \leq l \leq 3)$ for both autocorrelations.

(2) Products of complex cell responses with those at other orientations (capturing structures with mixed orientation content, such as junctions or corners) and with those at adjacent scales (capturing oriented features with spatially sharp transitions such as edges, lines, and contours). These cross-correlations are given by:

$$C(n, m) = \sum (e_n(i, j) - \mu(\vec{e}_n)) (e_m(i, j) - \mu(\vec{e}_m)) \quad (3.4)$$

where indices (n, m) specify two subbands arising from filters at different orientations at the same scale, or at orientations (same or different) at adjacent scales. At each scale, this yields 6 cross-orientation correlations, and 16 cross-scale correlations; across all scales, there are 24 cross-orientation and 48 cross-scale correlations.

(3) Products of the simple cell responses with phase-doubled simple cell responses at the next coarsest scale. Phase relationships at adjacent scales distinguish lines from edges, and can also capture gradients in intensity arising from shading. These correlations are given by,

$$S(n, m) = \sum (x_n(i, j) - \mu(\vec{x}_n)) \left(\frac{x_m^2(i, j)}{|x_m(i, j)|} - \mu \left(\frac{x_m^2(i, j)}{|x_m(i, j)|} \right) \right) \quad (3.5)$$

where indices (n, m) specify two adjacent scales (n is the finer scale). It is worth noting that all of these products may be represented equivalently as differences of squared sums and differences (i.e., $4ab = (a + b)^2 - (a - b)^2$), which might provide a more physiologically plausible form [2], as discussed further in Section 3.5.

(4) We finally include three marginal statistics (variance, skew, kurtosis) of the pixel-domain image, as well as for low-pass images reconstructed at each scale of

the coarse-to-fine process. Higher-order moments of order p are,

$$\mu^{(p)}(\vec{s}_n) = \sum (s_n(i, j) - \mu(\vec{s}_n))^p \quad (3.6)$$

From this, the skew and kurtosis are

$$\gamma(\vec{s}_n) = \frac{\mu^{(3)}(\vec{s}_n)}{(\mu^{(2)}(\vec{s}_n))^{3/2}} \quad (3.7)$$

$$\kappa(\vec{s}_n) = \frac{\mu^{(4)}(\vec{s}_n)}{(\mu^{(2)}(\vec{s}_n))^2} \quad (3.8)$$

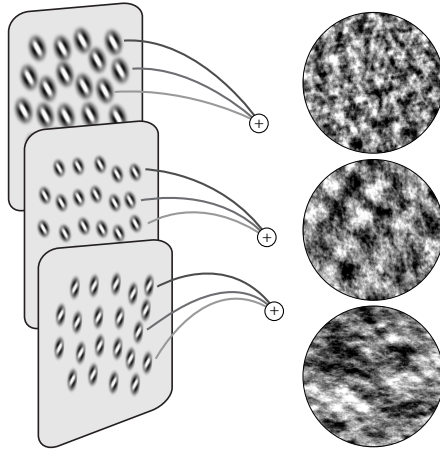
In the pixel domain, these marginal statistics capture the gray-level distribution, e.g. differentiating shades of gray (low kurtosis) from black and white (high kurtosis) [148]. Including the marginal statistics at each scale of the coarse-to-fine process is somewhat redundant with including it for the pixels only, but helps stabilize the synthesis procedure.

3.2.3 Synthesis

After computing responses on an original image, we generate a new image by beginning with an image of Gaussian white noise, and adjusting it using gradient descent until it matches the statistics computed on the original image, as depicted in Figure 3.2. We perform gradient descent in the steerable pyramid basis. Specifically, we perform the pyramid decomposition on the image of noise, compute the gradient of each statistic (or group of statistics) with respect to the pyramid coefficients, and move in the gradient direction so that the statistic is the same as the original. Equations for these gradient adjustments (specifically, projections) can be found in [175]. The adjustments are made separately on each band of the pyramid, from

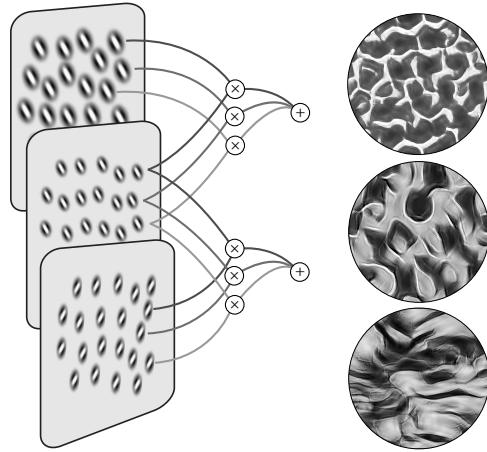
Spectrally-matched “noise”

No correlations matched



“Naturalistic”

Correlations matched



Synthesis algorithm

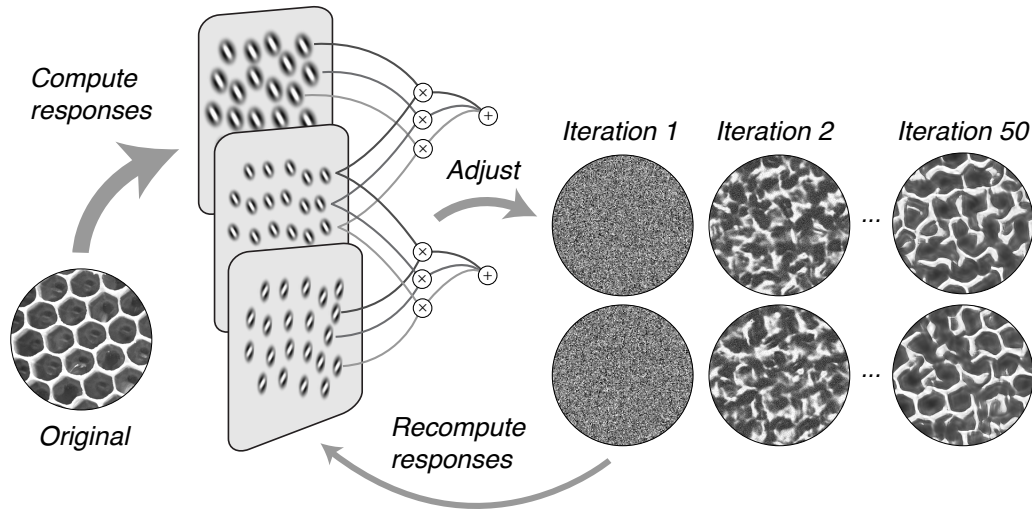


Figure 3.2: *Naturalistic textures were synthesized by iteratively imposing higher-order correlations derived from a two-stage computation (see also Figure 2.12); noise images were synthesized by only imposing the spatially-averaged responses of the initial V1-stage.*

coarse to fine, and the image is then reconstructed from the adjusted pyramid coefficients. Because the pyramid representation is overcomplete, imposing statistics on each band does not guarantee that the resulting reconstructed image will have the correct statistics after one adjustment. Furthermore, because the statistics are not independent, imposing one set of statistics can affect the others. We repeat the adjustments several times, and monitor convergence of all groups of parameters. For all images here, we used 50 iterations, which is typically sufficient for convergence.

Because the dimensionality of the image is larger than the number of parameters, this process yields multiple random high-entropy samples that are statistically identical in terms of the model parameters. In principal, we would like to sample from the maximum entropy distribution subject to the constraints governed by the higher-order statistics computed on the original. In practice, however, this is intractable. Synthesis-through-imposition, starting with a high-entropy distribution (Gaussian), approximates the appropriate sampling and yields high-entropy samples, but we cannot guarantee maximum entropy sampling [175].

For electrophysiological and fMRI experiments, we used as original images 15 diverse natural homogenous black and white photographs of visual texture, drawn from both commercial and personal databases (Figure 3.3). For each, we synthesized 15 distinct “naturalistic” samples. This yields a family of 15 self-similar images, which collectively form what we call a “texture category”.⁸

For each category, we also generated spectrally-matched filtered “noise” images (“noise” for short) (Figure 3.2 and 3.3). These are designed to match the spatially-averaged V1 filter responses of the original, but lack higher-order structure. For consistency in image generation, this should be done by beginning with white Gaussian noise and iteratively matching the spectral power averaged within each band of the multi-scale multi-oriented filter bank. But nearly identical results

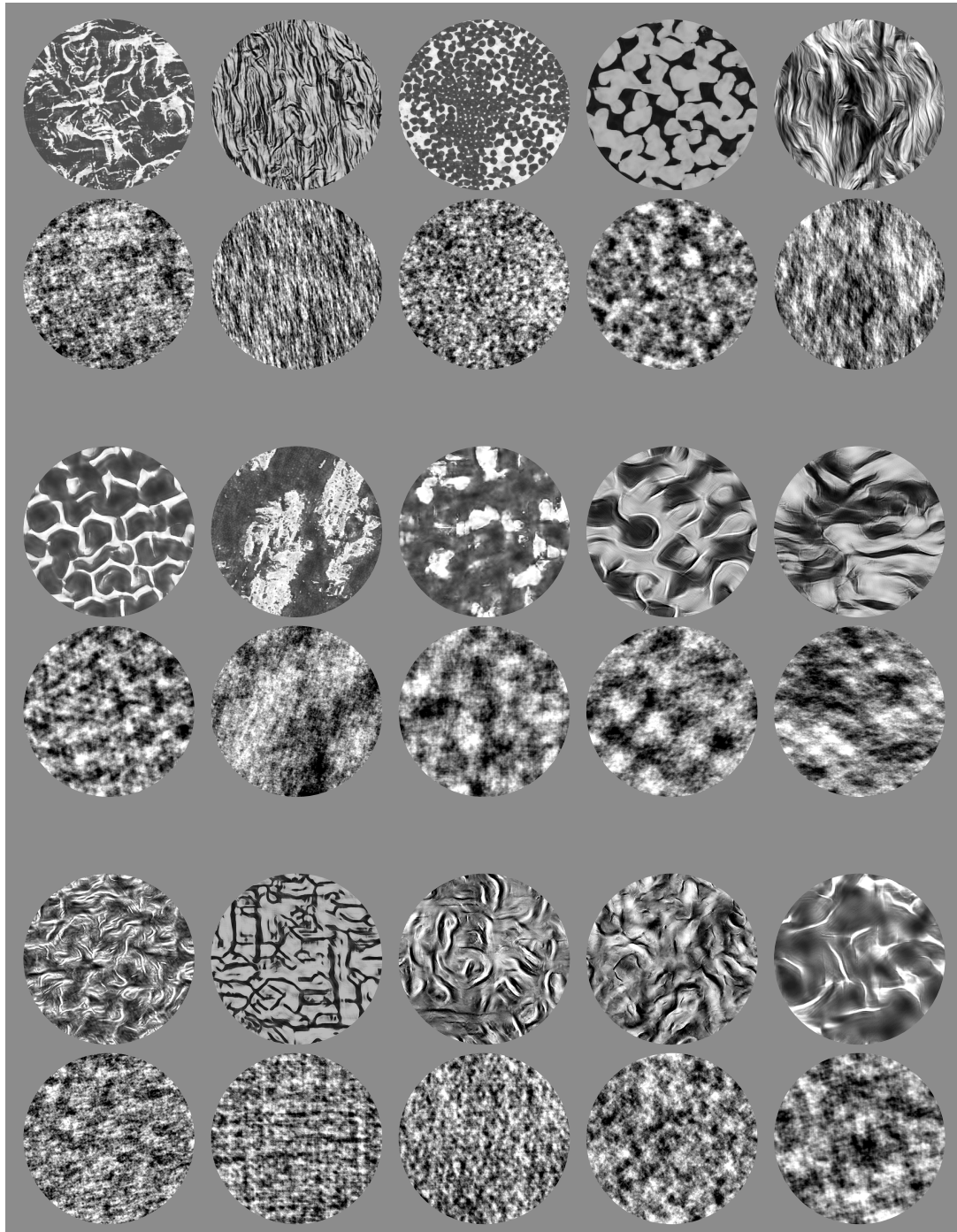


Figure 3.3: *Experimental stimuli for physiology experiments; 15 categories were used, with naturalistic and noise images for each.*

were obtained by matching the complete two-dimensional power spectra of each image and randomizing the phase, so this procedure was used throughout for speed of computation. We applied this procedure to each naturalistic image to generate 15 noise samples for each category.

3.3 Electrophysiology in macaques

3.3.1 Methods

Recording

We recorded from 12 anesthetized, paralyzed, adult macaque monkeys (2 *M. Nemestrina* and 10 *M. Cynomolgus*). Our standard methods for surgical preparation have been documented in detail previously [40]. We maintained anesthesia with infusion of sufentanil citrate ($6\text{--}30\ \mu\text{g kg}^{-1}\text{ hr}^{-1}$) and paralysis with infusion of vecuronium bromide (Norcuron; $0.1\ \text{mg kg}^{-1}\text{ hr}^{-1}$) in isotonic dextrose-Normosol solution. We monitored vital signs (heart rate, lung pressure, EEG, body temperature, urine volume and specific gravity, and end-tidal pCO_2) and maintained them within the appropriate physiological range. The eyes were protected with gas permeable contact lenses and refracted with supplementary lenses chosen through direct ophthalmoscopy. At the conclusion of data collection, the animal was killed with an overdose of sodium pentobarbital. All experimental procedures were conducted in compliance with the NIH *Guide for the Care and Use of Laboratory Animals* and with the approval of the New York University Animal Welfare Committee. We made a craniotomy and durotomy centered approximately 2–4 mm posterior to the lunate sulcus and 10–16 mm lateral and individually advanced several quartz-platinum-tungsten microelectrodes (Thomas Recording) into the brain at an angle

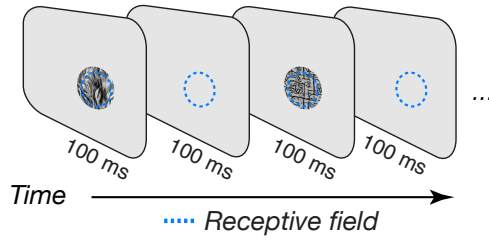


Figure 3.4: *Stimulus sequence in the physiology experiment. Stimuli were suitably vignettted and centered on each neuron's receptive field (dotted blue circle).*

20° from vertical. We distinguished V2 from V1 on the basis of depth from the cortical surface and changes in the receptive field location of recorded units. In an effort to obtain an unbiased sample of single units, we made extracellular recordings in V1 and V2 from every single unit with a spike waveform that rose sufficiently above noise to be isolated, and we fully characterized every unit that demonstrated a measurable visually-evoked response to any class of stimuli (i.e., gratings or naturalistic texture). Data are reported from every unit for which we completed characterization (see below).

Visual stimulation

We presented visual stimuli on a gamma-corrected CRT monitor (Eizo T966; mean luminance, 33 cd/m²) at a resolution of 1280 × 960 with a refresh rate of 120Hz. Stimuli were presented using Expo software on an Apple computer. For each isolated unit, we first determined its ocular dominance and occluded the non-preferred eye. We used drifting sinusoidal gratings to characterize the basic receptive field properties of each unit, including receptive field center, tuning for orientation and direction, spatial and temporal frequency, size, and contrast. We then presented the texture stimuli. We used a set of 15 texture categories, and generated 15 samples for each category for a total of 225 images. 15 spectrally-matched noise samples

of the 15 categories were also presented. The 450 unique images making up our stimulus ensemble were presented in pseudo-random order for 100 ms each, separated by 100 ms of mean luminance (Figure 3.4). Each image was presented 20 times. Images were presented to every unit centered on the classical receptive field, at the same scale and at a size of 4° within a raised cosine aperture. We chose a 4° aperture to be larger than all the receptive fields at the eccentricities from which we typically record. Nearly all recorded units had receptive fields smaller than 4° , and the majority were less than 2° . For a subset of V1 and V2 neurons we additionally presented stimuli in a smaller aperture matched to the receptive field size of that unit. The aperture diameter was set to be the grating summation field as measured with full contrast drifting gratings [40]. We ran the full texture stimulus ensemble within this aperture although typically with only 5-10 repeats per image.

Analysis

The full stimulus ensemble consisted of 450 images presented 20 times each. All analyses discussed in this Chapter were performed after averaging spiking responses across those 20 repeats, and also averaging responses across the 15 samples. Depending on the analysis, responses were further averaged across texture category, neurons, and/or a temporal window, as discussed below. Basic receptive field properties for each neuron were determined offline by using maximum likelihood estimation to fit an appropriate parametric form to each tuning function. These fits were only obtainable for a subset of neurons (78% in V1, 62% in V2) due to incomplete characterization arising from time constraints during the experiment.

3.3.2 Results

We recorded the responses of 103 V1 and 102 V2 neurons in 12 macaque monkeys to the sequence of naturalistic and noise stimuli. As described above, both classes of stimuli were matched to an original in terms of spatially-averaged amplitudes of V1-stage filters, but only the naturalistic stimuli were additionally matched to a set of higher-order correlations, and thus contained more complex features of the original.

Macaque V2 responds differentially to naturalistic stimuli

We first examined responses as a function of time from stimulus onset. We counted spikes within a sliding, non-overlapping 10 ms window, and averaged the resulting time courses across texture categories. V1 neurons responded similarly to the two stimulus classes, while V2 neurons usually responded more vigorously to the naturalistic stimuli (representative example neurons shown in Figure 3.5). This distinction between V1 and V2 was evident in single neurons. To assess it at the population level, we averaged responses across neurons. Before averaging, we first normalized each neuron's response time course by dividing by its maximum response across all texture categories and time points. The distinction between V1 and V2 was clearly evident in the population time courses (Figure 3.5).

To further capture the differential response, we computed a modulation index as the difference in response to naturalistic and noise divided by the sum, within each 10 ms window. The average modulation index of neurons in V1 was near zero for most of the response, except for a weak late positive modulation (discussed more below). Neurons in V2 showed a substantial modulation that was evident

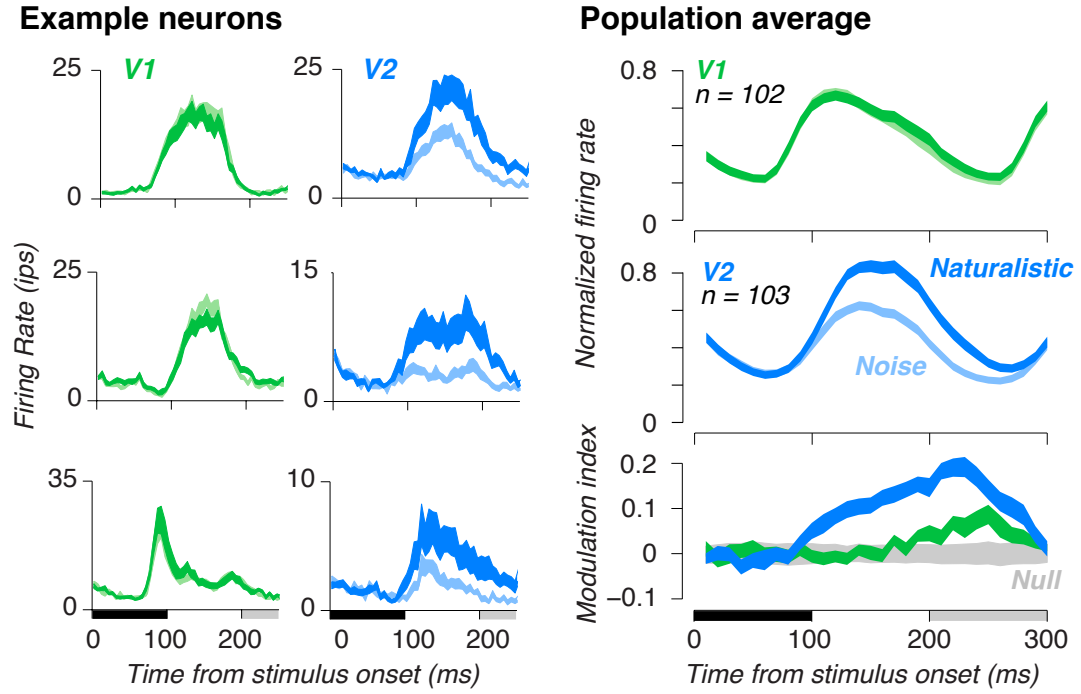


Figure 3.5: Neurons in V2 responded more to naturalistic images (dark) than to noise images (light). Shaded regions for single neurons indicate s.e.m. across texture categories; for the population average, s.e.m. across neurons. Black and gray horizontal bars show periods of stimulus presentation. Modulation index was computed as the difference between the response to naturalistic and noise, divided by the sum.

soon after response onset and maintained throughout the duration of the response (Figure 3.5).

Some neurons were more sensitive overall to naturalistic features than others. We computed a modulation index for each neuron, averaged over the response duration and over all samples of all texture categories. Response duration was defined as an 100 ms window following response onset, and was determined by inspection as the time point eliciting a response above baseline.⁹ Significant positive modulation was observed in 15% of V1 neurons, and 63% of V2 neurons (Figure 3.6, $P < 0.05$, randomization test for each neuron). The difference in modulation

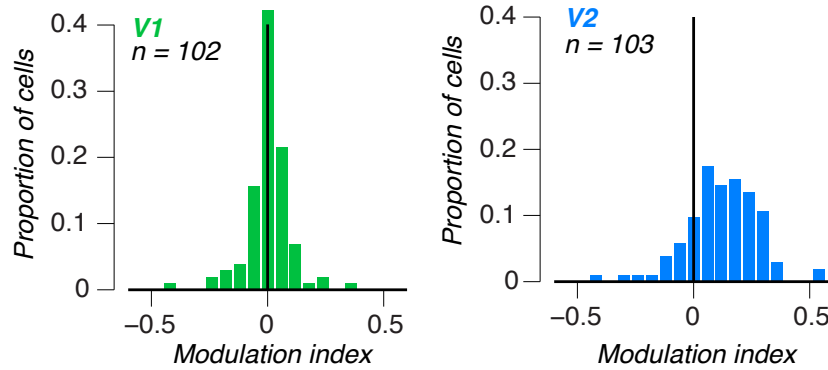


Figure 3.6: *Distribution of modulation indices (difference between response to naturalistic and noise, divided by the sum) across neurons in V2 and V1.*

between V1 and V2 was highly significant ($P < 0.0001$, t -test on signed modulation; $P < 0.0001$, t -test on modulation magnitude, ignoring sign).

V2 neurons were significantly modulated by naturalistic features on average, but the modulation was typically more pronounced for some textures than for others. The effective subset of textures varied from cell to cell, but there was a consistent trend across the population for some categories to be more effective than others (Figure 3.7). The pattern was not simply predicted by variability in firing rate across categories; there was no evidence for a correlation across categories between the average modulation index and the average normalized evoked response ($r = 0.42$, $P = 0.12$). Rather, the consistent pattern suggests that the higher-order correlations of some textures more than others differentially drive V2 neurons, regardless of the baseline response elicited by that category (e.g. due to its spectral properties). This variability will be examined later in this Chapter (when comparing macaque and human), and in Chapter 3 (when relating physiology to perception).

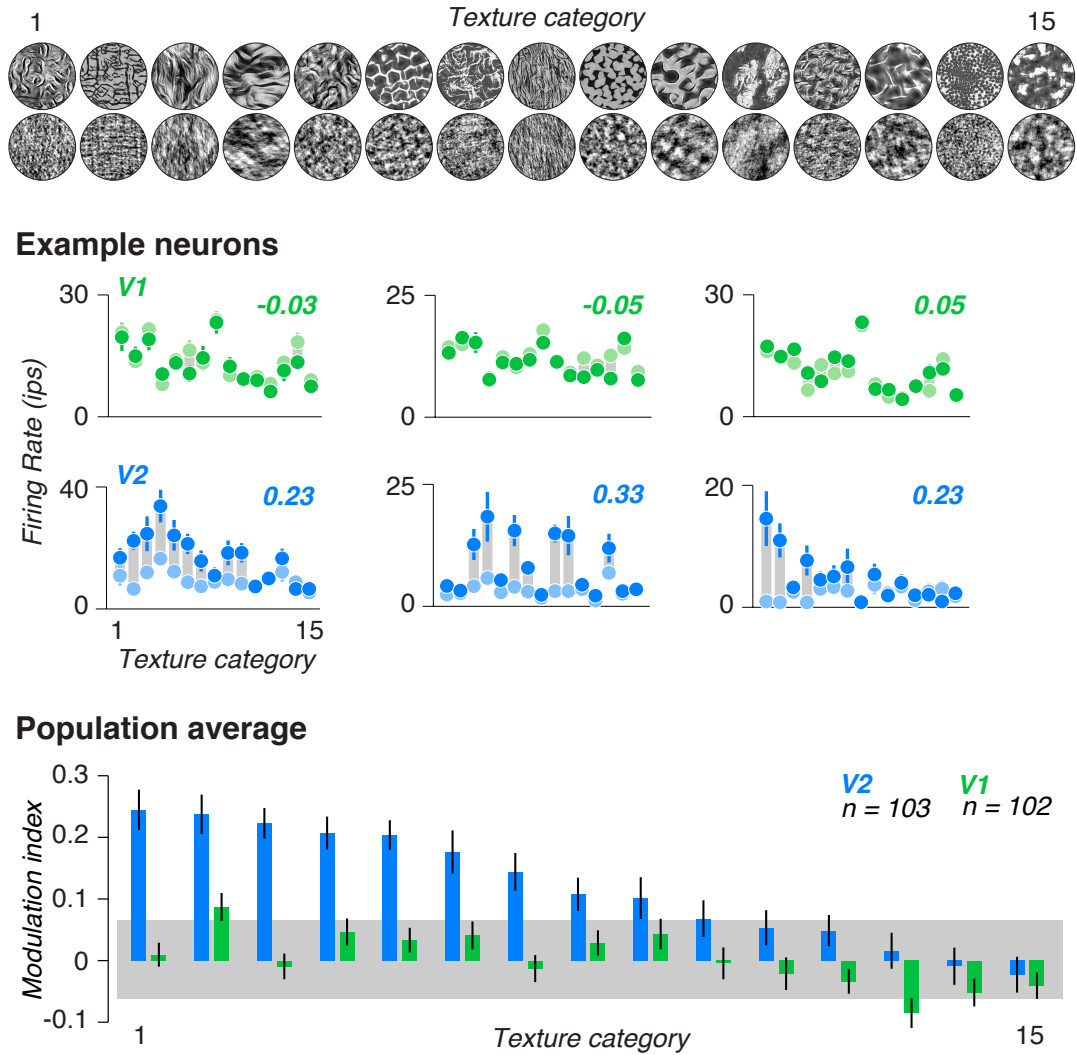


Figure 3.7: In V2 neurons, responses to naturalistic images (dark dots) were larger than to noise images (light dots), but the difference varied across categories (the number in each panel is the average modulation). Averaged across neurons, only some texture categories elicited large modulations. Textures (top) sorted according to modulation in V2. Error bars in population average, s.e.m. across neurons.

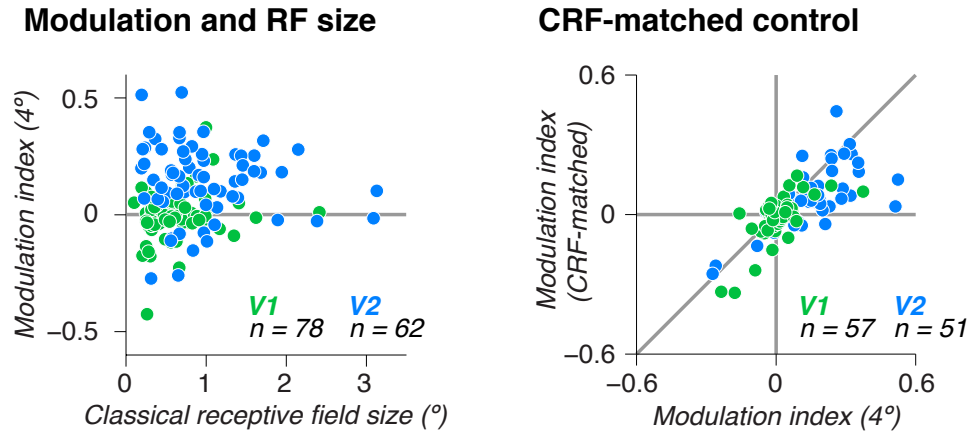


Figure 3.8: *Modulation was not predicted by receptive field size (left) in either V1 or V2. In V2, modulation remained, but was reduced, when shrinking the stimulus aperture to the extent of the classical receptive field of each neuron (right).*

Relationship to basic receptive field properties

The distinction between V2 and V1 was not due to differences in receptive field sizes. The stimuli presented to V1 and V2 cells were of the same size, roughly twice the size of a typical V2 receptive field, and 4 times the size of a typical V1 receptive field. There was, however, no evidence for a correlation between receptive field size and modulation in either area (Figure 3.8, V1: $r = 0.21$, V2: $r = -0.10$, $P > 0.05$), where receptive field size was defined by the standard deviation of the excitatory Gaussian from a center-surround fit to the size-tuning function [39, 40]. When we restricted analysis to subsets of neurons matched for average receptive field size, the difference in modulation index between areas was reduced only by 9% and remained highly significant ($P < 0.0001$, randomization test). We also made measurements on a subset of cells in which the stimulus was confined to each neuron's classical receptive field (CRF). In V1, the modulation was near 0 for both CRF-matched and large stimuli, though there was a small but significant reduction

in modulation for the matched stimuli ($P < 0.05$, paired t -test). In V2, there was a robust but incomplete reduction in modulation for the smaller stimuli ($P < 0.0001$, paired t -test), suggesting that the modulation in V2 depends partly on interactions between receptive field center and surround.

V1 and V2 neurons are generally similar in terms of other commonly measured properties, e.g. orientation and spatial frequency tuning, and contrast sensitivity [133]. Nevertheless, we wondered whether there was a relationship in either area between these properties and the response to naturalistic images. We found no evidence for a correlation in V2 between the modulation and the following properties: orientation tuning bandwidth, preferred spatial frequency, spatial frequency tuning bandwidth, the exponent and c_{50} of the contrast response function, and an index of surround suppression (Figure 3.9, all $P > 0.05$). We thus believe that our measurements have uncovered a new and different dimension of visual processing in V2.

In V1, we similarly found no evidence for relationships between modulation and these properties, except for a weak but significant negative correlation between orientation tuning bandwidth and modulation ($P = 0.033$). This would not be significant, however, were we to correct for multiple comparisons, which seems appropriate given the large number of possible correlations considered. More interestingly in V1, there was a small subset of cells with particularly high positive (or negative) modulation, and this tendency occurred in neurons with high surround suppression and contrast sensitivity (and small receptive field sizes). Generally, we attribute the absence of modulation in V1 to the fact that any mechanism measuring local spectral energy should not distinguish between the two families of stimuli (naturalistic and noise) because they are spectrally matched. But they are only matched over a sufficiently large spatial region, and for particularly small receptive field sizes, the match

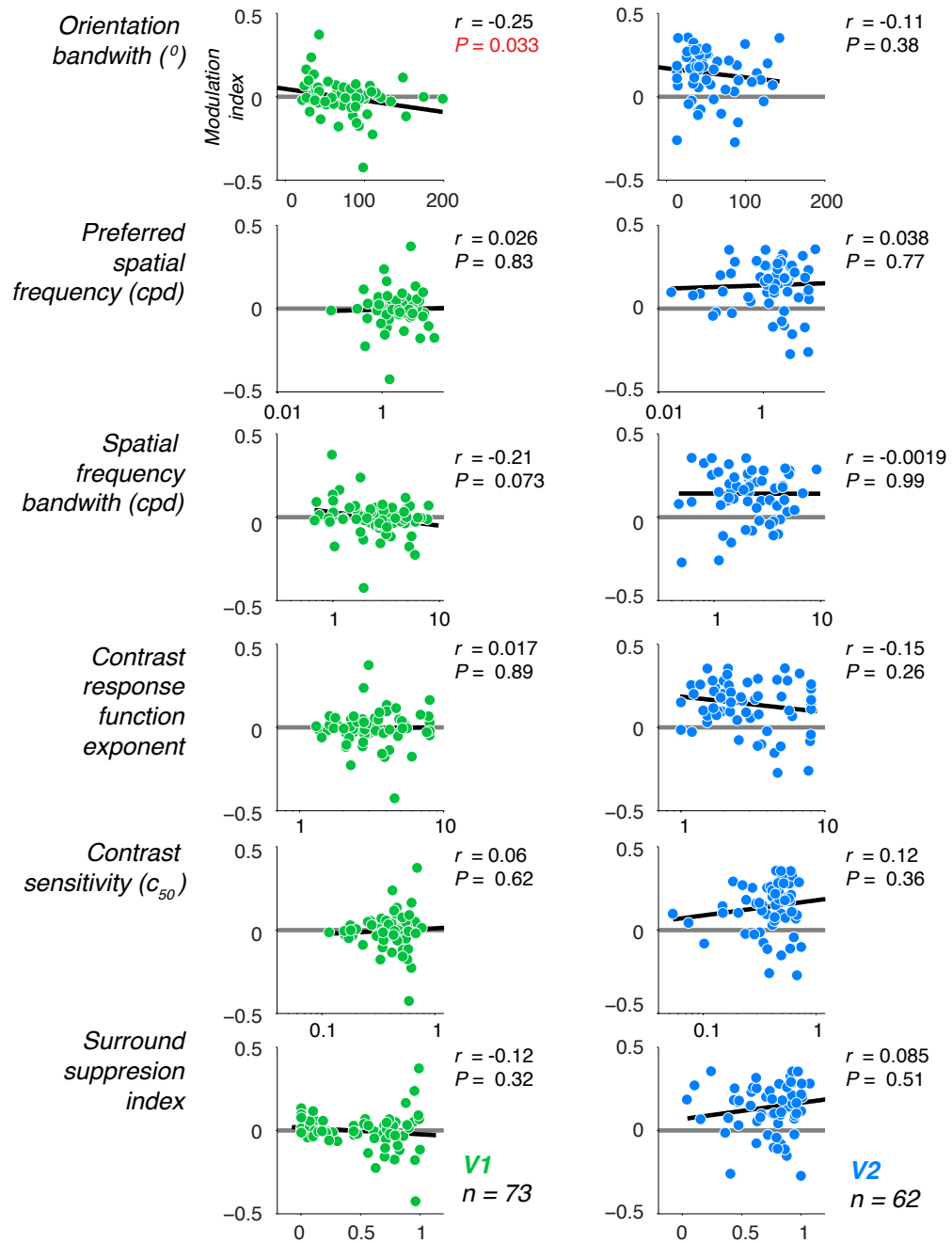


Figure 3.9: There was little evidence for correlations between modulation and six basic properties of neurons in V1 and V2.¹⁰

may be imperfect. This violation, coupled with strong suppression or sensitivity to contrast, could in part explain the small subset of V1 neurons that prefer either naturalistic or noise.

Control for marginal statistics

The robust differential response to naturalistic images in V2 neurons could, in principal, depend on any of the statistical properties imposed during the synthesis of naturalistic images, which include correlations across different positions, orientations, and spatial frequencies, as well as marginal statistics (skew and kurtosis). In principal, it would be possible to identify the importance of each of these groups of parameters by synthesizing images matched to some (but not all) of them, and measuring the responses of V2 neurons (analogous to psychophysical experiments [12]). Here, we describe one such experiment, focused on the importance of the marginal statistics.

For each of the 15 categories, we generated synthetic “marginal” images matched to the original only in terms of marginal statistics, specifically, the mean, variance, skewness, and kurtosis of each band of the steerable pyramid representation. As for the naturalistic images, these statistics were imposed iteratively on the pyramid coefficients, from coarse to fine. But unlike the naturalistic images, no cross-band statistics were imposed. This procedure is nearly identical to the texture model proposed by Heeger and Bergen [101], which generated textures by matching, to an original, the full histogram of each pyramid band. In practice, matching the complete histogram and matching the first four moments of each band yielded comparable results.

We measured the responses of 40 V2 neurons to the “marginal” stimuli (Figure

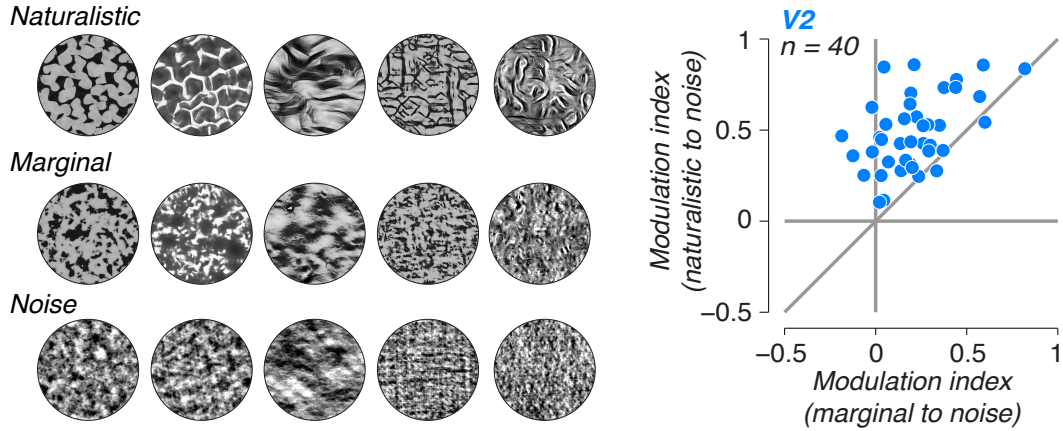


Figure 3.10: *Marginal stimuli (middle row) were matched to the marginal statistics of subbands of the naturalistic images (top row). Modulation for marginals vs noise was reduced, but still present, compared to modulation for naturalistic vs noise (right).*

3.10). We targeted neurons that exhibited a differential response to naturalistic images based on an initial characterization, and we only presented marginal stimuli for the subset of texture categories for which we found clear modulation. We found significantly positive modulation ($P < 0.0001$) for marginal images (compared to noise), but the modulation was significantly lower than it was for the naturalistic images, reduced by 53% ($P < 0.0001$, paired t -test). This suggests that the full magnitude of differential response found for naturalistic images requires the imposition of cross-band correlations, but some of the differential response may reflect only the marginal statistics.

A caveat of this approach, mentioned in Chapter 2, is that the different groups of statistics are not independent, so imposing one set may inadvertently impose another [12]. Imposing highly kurtotic marginals in an image with low spatial frequencies, for example, produces sharp edges that in turn induce strong dependences across spatial frequency bands. It is thus unclear whether the fairly strong residual modulation found for the marginal stimuli was due to the marginals *per se*, or due

to unintentional imposition of cross-band correlations.






How could this approach be improved in future experiments? Maximum differentiation competition (MAD) [232] could be used to isolate the contributions of different groups of statistics. In MAD competition, images are generated that maximally differ with respect to one group of parameters (e.g. cross-scale correlations), but match for another. However, this approach is computationally expensive, would be experimentally taxing to perform for all the groups of parameters in the model, and may be impossible for certain parameter combinations due to their complex dependencies. A more promising alternative would be to reformulate the texture parameters using a small family of independent mechanisms that better capture specific hypothesis about what V2 neurons compute. We return to a discussion of such mechanisms in Section 3.5.

Comparison to other stimuli that differentiate V2 from V1

A number of experiments have measured responses of V2 neurons to specialized artificial stimuli, including angles, curvature, anomalous contours, and second-order patterns [171, 116, 8, 225, 129, 141, 102, 103, 64, 225]. As discussed above, in cases where V2 and V1 were directly compared, the selectivity of V2 neurons for these attributes was qualitatively and quantitatively similar to that of V1, in contrast to the robust differential responses to naturalistic stimuli found here.

In order to compare these results on a common ground, we considered the artificial but rhetorically useful “electrophysiologists’s guessing game” (a variant of the “Turing test” for assessing machine intelligence). Imagine recording from a neuron; you do not know its eccentricity, but you are allowed to measure its response to any stimulus. How well can you determine whether it is in V1 or V2?

Unique selectivity in V2

	Naturalistic texture	77%
	Angles (Ito & Komatsu, 2004; Anzai et al., 2007)	66%
	Curvature (Hegde & Van Essen, 2007)	53%
	Anomalous contours (Peterhans et al., 1989)	61%
	Border ownership (Zhou et al., 2000)	64%

Basic forms of selectivity






	Receptive field size	68%
	Surround suppression	70%
	Contrast sensitivity	52%
	Spatial frequency bandwidth	59%
	Orientation bandwidth	66%

Figure 3.11: Performance discriminating between distributions of neurons in V2 and V1 on the basis of one-dimensional response metrics reported in the literature. Analyses of basic response properties were derived from our data set.

Performance in this game can be estimated by examining, for any one-dimensional response metric, distributions of the metric in V1 and V2. Applying ROC analysis to the distributions assesses how well cells in the two areas can be discriminated based on the response property.

We estimated these performances for a number of properties previously measured in V2, as well as for the response to naturalistic textures, and for the other basic response properties that we measured (Figure 3.11). Compared to properties previously linked to V2, performance was much higher for the differential response to naturalistic textures (77%), approaching that found for differentiating area MT from V1 on the basis of direction selectivity (83%). Performance, however, was also quite high for receptive field size (as expected given the two-fold increase from V1 to V2), as well as for surround suppression, which is consistent with the results of an experiment modeling V2 responses to natural images [237]. But as demon-

strated in Figure 3.9, we found no evidence for correlations between these properties and modulation within V2, suggesting that the response to naturalistic stimuli is a different, unrelated dimension of form processing.

It remains unclear how the other documented properties unique to V2 neurons relate to the effects described here. Some of them may reflect in part the response property identified with our texture stimuli. Sensitivity to angles and curvature, for example, may be related to sensitivity to cross-orientation correlations, but that is just one of several dependencies imposed in our stimuli. To test this and related hypotheses, future experiments could measure our effect alongside sensitivity to these other properties in the same neurons.

3.4 fMRI in humans

Given the reliable effect of higher-order image statistics on the responses of V2 neurons, we wondered if similar effects could be observed in humans using functional magnetic resonance imaging (fMRI), which can be used to reliably distinguish visual areas on the basis of retinotopic organization, and is capable of capturing large-scale differential responses across visual areas and can [230]. In several other domains, including contrast sensitivity [100], pattern direction selectivity [99], and selectivity to faces [217], fMRI responses in cortical areas with ubiquitous neuronal preference have reliably reproduced electrophysiological measures. But thus far, no fMRI studies in humans have reliably distinguished V2 from V1 in terms of basic functional response properties.

3.4.1 Methods

Subjects

Data were acquired from three healthy subjects with normal or corrected-to-normal vision (all male; age range, 26-30 years). One subject was the author of this thesis, and another was a collaborator on the project. The third was naive to the purpose of the experiment. Experiments were conducted with the written consent of each subject and in accordance with the safety guidelines for fMRI research, as approved by the University Committee on Activities Involving Human Subjects at New York University. Each subject participated in three scanning sessions: one session to obtain a set of high-resolution anatomical volumes, one session for standard retinotopic mapping (single wedge angular position, and expanding ring eccentricity), and one session to measure differential responses to naturalistic and spectrally-matched noise stimuli.

Stimuli

Stimuli were presented using Matlab (MathWorks) and MGL (available at <http://justingardner.net/mgl>) on a Macintosh computer. Stimuli were displayed via an LCD projector onto a back-projection screen in the bore of the magnet. Subjects laid supine and viewed the stimuli through an angled mirror. All images were presented within a suitably vignetted annular region (inner radius, 2° ; outer radius, 8°). We used textures that approximately matched in scale the presentation conditions in the electrophysiological and psychophysical experiments.

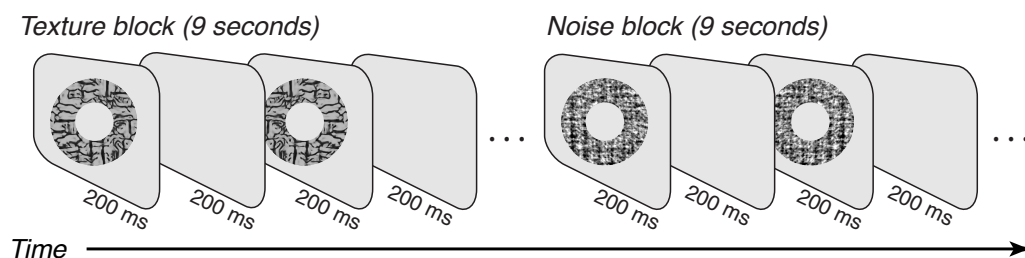


Figure 3.12: *Stimulus sequence in the fMRI experiment. Alternating blocks of naturalistic texture and noise were presented. Subjects performed a demanding task at fixation to divert attention from the peripheral stimulus.*

Protocol

Blocks of naturalistic and spectrally-matched noise stimuli were presented in alternation (Figure 3.12). Within each 9 s block, a random sequence of images from one texture category were presented at 5 Hz. Each run consisted of 20 blocks: 10 naturalistic, 10 noise. Different texture categories were presented in separate runs. Subjects performed two runs for each texture category. A separate localizer run was used to define retinotopic subregions corresponding to the stimulus region. Within each 9 s block of the localizer run, a random sequence of both naturalistic and noise images were presented within the stimulus annulus or the region complementary to the annulus. Each run consisted of 40 blocks: 20 annulus, 20 anti-annulus.

Task

Observers performed a demanding two-back detection task continuously throughout each run to maintain a consistent behavioral state, encourage fixation, and divert attention from the peripheral stimulus. Without any attentional control, or if subjects are attending the peripheral target stimuli, fMRI responses in visual cortex exhibit large and highly variable (trial to trial) attentional effects [177]. Digits (0

to 9) were displayed continuously at fixation, changing every 400 ms. The subject used a button press to indicate whether the current digit matched the digit from two steps before.

Preprocessing

The anatomical volume acquired in each scanning session was aligned to the high-resolution anatomical volume of the same subject's brain, using a robust image registration algorithm [152]. Data from the first half cycle (eight frames) of each functional run were discarded to minimize the effect of transient magnetic saturation and allow the hemodynamic response to reach steady state. Head movement within and across scans was compensated for using standard procedures [152]. The time series from each voxel was high-pass filtered (cutoff, 0.01 Hz) to remove low-frequency noise and drift [205].

Analysis

We performed two complementary analyses to visualize and quantify the fMRI responses to alternating blocks of naturalistic and noise images. First, for each voxel, response time courses were averaged across texture categories, and then fit with a sinusoid with period matched to the block alternation (9 s). The coherence between the best-fitting sinusoid and the average time series is commonly used to quantify the statistical reliability of the fMRI responses modulations, in this case characterizing the differences in cortical activity evoked by naturalistic and noise images. To further quantify responses and compare V1 and V2, time courses were averaged across voxels and across repeated runs, but separately for each texture category.

For each texture category, the time course of each voxel was projected onto a unit-norm sinusoid having period matched to the stimulus alternation and phase given by the responses to the localizer scan (annulus versus anti-annulus). This reference phase provided an estimate of the hemodynamic delay, and the amplitude of this projection isolated the component of the response time course that was positively modulated [99]. This analysis procedure took full advantage of a priori knowledge of the block-alternation experimental design [99], and provided unbiased estimates of the amplitudes of response modulation between the naturalistic and noise images.

MRI acquisition

MRI data were acquired on a Siemens 3T Allegra head-only scanner using a head coil (NM-011; Nova Medical) for transmitting and an eight-channel phased array surface coil (NMSC-071; Nova Medical) for receiving. Functional scans were acquired with gradient recalled echo-planar imaging to measure blood oxygen level dependent changes in image intensity [159]. Functional imaging was conducted with 24 slices oriented perpendicular to the calcarine sulcus and positioned with the most posterior slice at the occipital pole (1500 ms repetition time; 30 ms echo time; 72 flip angle; $2 \times 2 \times 2$ mm voxel size; 104×80 voxel grid). A T1-weighted magnetization-prepared rapid gradient echo anatomical volume (MPRAGE) was acquired in each scanning session with the same slice prescriptions as the functional images (1530 ms repetition time; 3.8 ms echo time; 8 flip angle; $1 \times 1 \times 2.5$ mm voxel size; 256×160 voxel grid). A high-resolution anatomical volume, acquired in a separate session, was the average of three MPRAGE scans that were aligned and averaged (2500 ms repetition time; 3.93 ms echo time; 8 flip angle; $1 \times 1 \times 1$ mm voxel size; 256×256 voxel grid). This high-resolution anatomical scan was used both for

registration across scanning sessions and for gray matter segmentation and cortical flattening.

Defining retinotopic regions of interest

Each subject participated in a standard retinotopic mapping experiment, described in detail elsewhere [128, 72, 80]. The data were analyzed, following standard procedures to identify meridian representations corresponding to the borders between retinotopically organized visual areas V1, V2, V3, and V4. There is some controversy over the exact definition of human V4 and the area just anterior to it; we adopted the conventions proposed by Wandell and colleagues [230]. We used data from an independent localizer scan (see above) to further restrict each visual area to only those voxels responding to the stimulus annulus with coherence of at least 0.25. Qualitatively similar results were obtained using higher or lower thresholds.

3.4.2 Results

Human V2 responds differentially to naturalistic stimuli

For each subject, we visualized responses to naturalistic images on a flattened representation of the occipital cortex. In all three subjects, the naturalistic stimuli produced reliable modulation of the fMRI response throughout V2, compared to weak or absent modulation in V1. The presence of the effect tracked the V1/V2 boundary surprisingly well, providing the first instance of a functional response property that distinguishes human V2 from V1. We further quantified the differential response by computing a measure of response amplitude (see above). Differences in response amplitude averaged over V2 and V1 were highly significant in each subject (Figure 3.13; $P < 0.0001$, paired t -test).

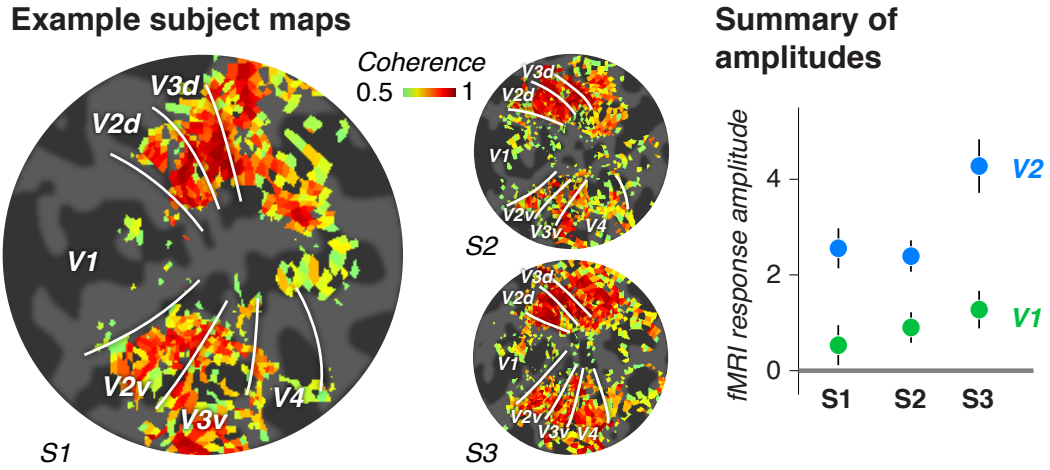


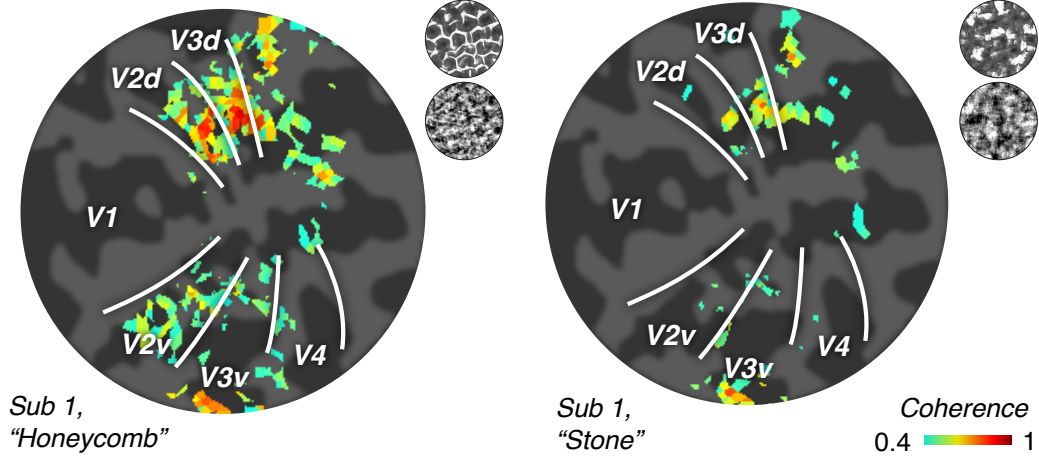
Figure 3.13: Flattened maps (left) of the occipital cortex show differential responses to naturalistic images in V2 but not V1, for three subjects. Estimates of response amplitude (far right) were averaged across voxels in each area.

Just as in the single neurons, there was variability in the magnitude of differential response across texture categories (Figure 3.14 shows two examples). In each subject, these magnitudes were highly correlated across independent runs (Figure 3.14; S1: $r = 0.95$; S2: $r = 0.87$; S3: $r = 0.99$; all $P < 0.0001$). Magnitudes were also highly correlated across subjects (average pairwise correlation, $r = 0.86$).

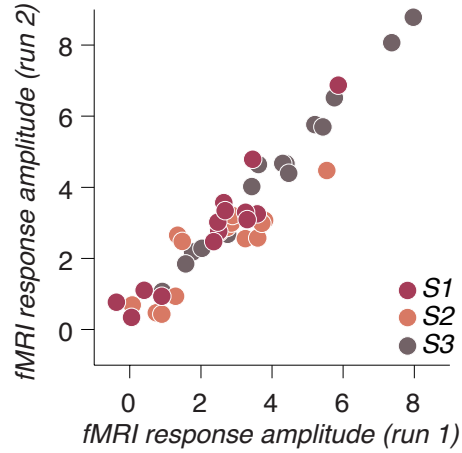
Event-related experiment

The fMRI experiment described thus far used a block design. This protocol is optimal with respect to signal-to-noise ratio because the power in the response time course is concentrated at a single temporal frequency, selected to obtain a reasonable trade-off between the signal attenuation at high frequencies due to the sluggishness of the hemodynamics [23] and the noise and drift that dominate fMRI signals at low frequencies [205]. However, a block design only measures response

Example textures



Texture-to-texture reliability across runs



Texture-to-texture reliability across subjects

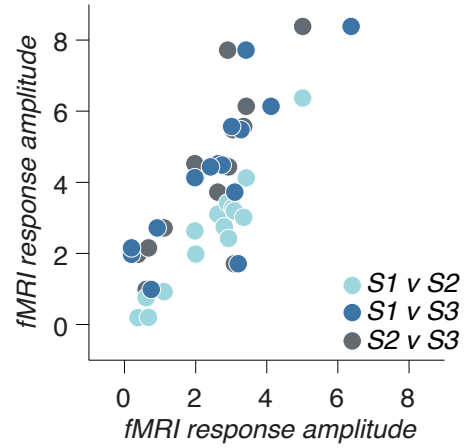


Figure 3.14: *fMRI responses varied across texture categories (two examples, top). For three subjects, this variability was consistent across runs within a subject (lower left), and also across subjects (lower right).*

modulation, rather than measure the response to each stimulus class relative to a baseline, making it less comparable to the electrophysiological measures.

In one subject, we performed an event-related experiment. The procedure and analyses were similar to those described in [28, 73]. Briefly, on each trial, we

presented a burst of 5 samples from one texture category, either naturalistic or noise, for 1 s (100 ms on, 100 ms off). The presentation was followed by 2 s of mean luminance, and an inter-stimulus interval that was jittered randomly across trials. On a small fraction of trials, no stimulus was present. In total, there were 30 trials types (15 naturalistic, and 15 noise). The subject completed 24 runs of the experiment. Data collection and stimulus presentation was otherwise similar to that described for the block design above. To analyze the data, we used deconvolution to estimate a hemodynamic response function for each visual area (after averaging response time courses across all 30 stimulus categories) [50]. We then used a general linear model to estimate the response amplitude for each voxel and stimulus category. We included as regressors in the linear model the convolution of each stimulus sequence with the estimated HRF, as well as the convolution of each stimulus sequence with the derivative of the HRF. Including the derivative absorbs voxel-by-voxel variability in the sluggishness of the hemodynamics, rather than absorbing it into variability in the estimates of response amplitude. We estimated response amplitudes to each category of texture and noise, and computed, for each voxel, a modulation index similar to that computed for the single-unit data, taking the difference in the response to naturalistic and noise divided by the sum of the absolute value of the response to naturalistic and the absolute value of the response to noise.

Figure 3.15 shows modulation indices averaged across voxels in V1 and V2 for each of the 15 texture categories, sorted as in Figure 3.7. Modulation indices were significantly higher in V2 than in V1 ($P < 0.0001$, paired t -test across categories), corroborating both the single-unit results as well as the results of the block design experiment. However, when examining variability in modulation across categories, which was reliable in the block design across subjects and runs, we did not find evidence for a significant correlation between the modulation measured in the event-

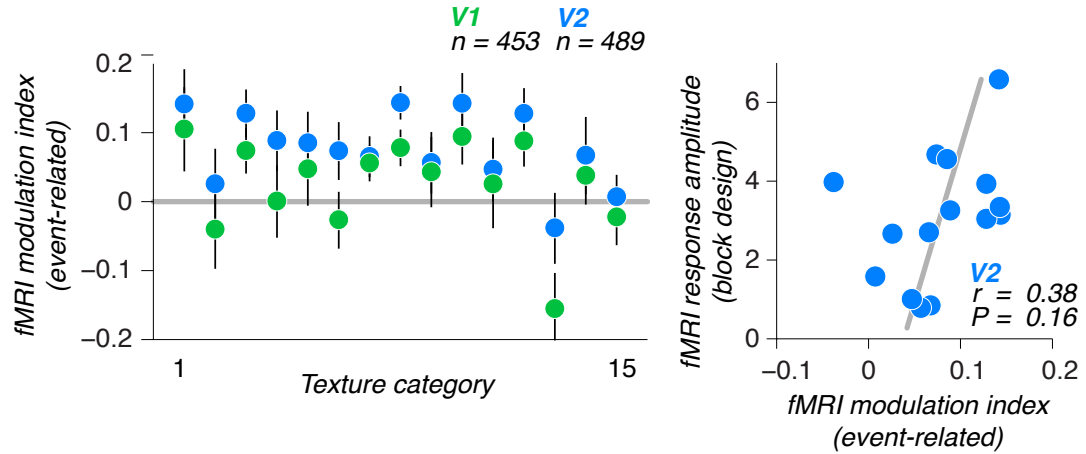


Figure 3.15: Modulation indices for 15 texture categories obtained from an event-related fMRI experiment. (Left) Modulation was higher in V2 than V1, error bars indicate s.e.m. across 23 runs. (Right) Modulation was not significantly correlated across categories with response amplitudes measured using a block design.

related and block-design experiments (Figure 3.15, $r = 0.38$, $P = 0.16$). We attribute this failure to unstable estimates of response amplitudes in the event-related design. Unlike the block design experiment, estimates of response amplitude for each texture in the event-related experiment were highly variable across runs, and we found no evidence for reliable accuracy when training a classifier to discriminate among the 30 image categories based on the multivariate pattern of responses in either V1 or V2 (not shown). Thus, although the event-related design yields a measure of modulation more comparable to that obtained from single-units, and broadly reproduced the difference between V2 and V1, the number of stimuli would likely need to be reduced (or the number of runs and subjects greatly increased) to compensate for the design's lower signal-to-noise ratio.

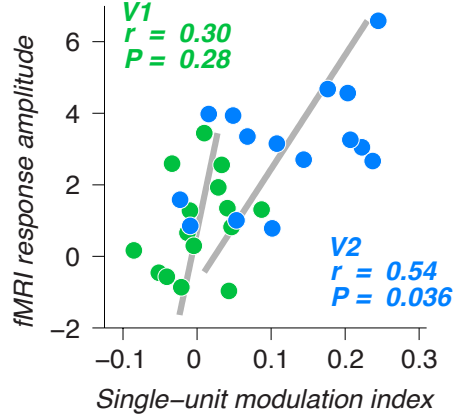


Figure 3.16: The diversity of modulation across texture categories was similar between fMRI responses (averaged across 3 subjects) and single-unit modulation (averaged across neurons, 102 in V1, 103 in V2).

3.4.3 Comparing human to macaque

V2 responses vary similarly across categories

In both single-unit and fMRI responses, we found that differential responses to naturalistic images varied across texture categories (Figures 3.7 and 3.14). We compared the amplitude of modulation of the fMRI responses with the average single-unit modulation index across texture categories. The fMRI and electrophysiological measures of response modulation were reliably correlated in V2 (Figure 3.16; $r = 0.54$, $P < 0.05$), but this was not evident in V1 ($r = 0.30$, $P = 0.28$). We also correlated the modulation indices from each individual neuron with the fMRI response modulations, and found that correlations were significantly higher in V2 than in V1 ($P < 0.005$, t -test on Fisher Z-transformed correlations).¹¹ The presence and diversity of the differential responses to naturalistic images in V2 were thus similar in anesthetized macaque neuronal populations and awake human fMRI. The

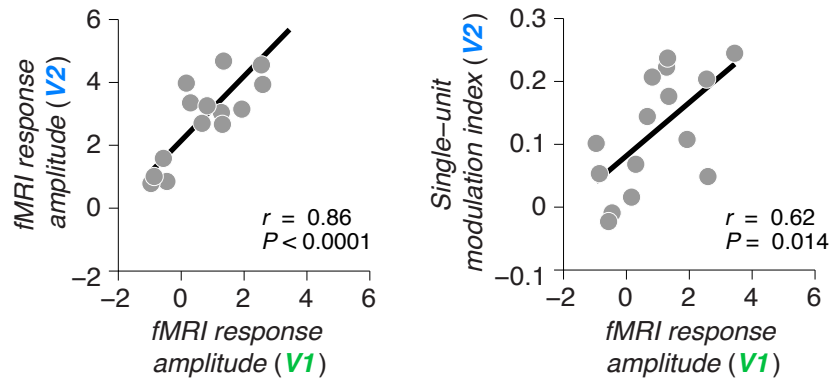


Figure 3.17: Although fMRI response amplitudes to naturalistic textures in V1 were small, across texture categories they were reliably correlated with fMRI response amplitudes in V2 (left), and single-unit response modulations in V2 (right)

comparison is complicated somewhat by the nature of the block design experiment, which only measures the difference between naturalistic and noise, whereas in single neurons we measured the difference divided by the sum. An event-related experiment would have enabled a more appropriate comparison, but as discussed above, the lower signal-to-noise ratio of that approach makes it unsuitable for measuring differential responses across a large variety of categories.

Dynamics and possible evidence for feedback in V1

Despite the robust differences between V2 and V1 emphasized thus far, both single-unit and fMRI measures revealed weak residual responses in V1 to naturalistic stimuli. In the single-unit data, V1 responses showed a late component of modulation, approximately 100 ms after response onset (Figure 3.5), though this was only reliably present in a small subset of neurons. In the fMRI experiment, responses were significantly larger in V2 than in V1, but responses in V1 were larger than 0 (Figure 3.13), significantly so in two of three subjects ($P < 0.05$, t -test).

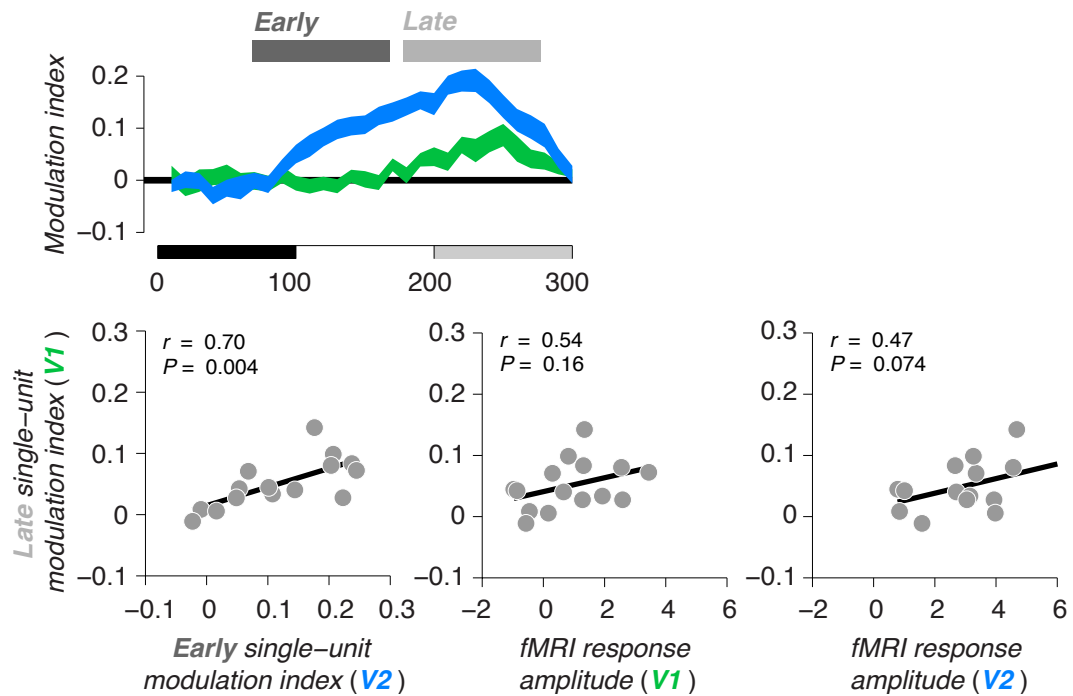


Figure 3.18: Response modulation was computed separately within two temporal windows, early and late in the response. The early period followed response onset and was adjusted for each neuron; the late period was fixed across neurons, chosen as the window showing modulation in the V1 population average. Across texture categories, there was some evidence for a correlation between this late V1 component and other measures of response amplitude.

We wondered whether these responses in V1 might reflect feedback from V2 to specific subpopulations of V1 neurons. Rigorously demonstrating feedback would require measuring responses in V1 while inactivating V2, e.g. through cooling [174]. However, two pieces of additional evidence suggest a feedback explanation. First, although fMRI responses in V1 were not correlated with single-unit responses in V1 (Figure 3.16), fMRI responses in V1 were reliably correlated with fMRI responses in V2 (Figure 3.17). If the variability across textures within V2 is reliable and functionally relevant (as we will argue in Chapter 4), the residual signal in V1 carries similar information. We also analyzed the late component of single-unit responses in

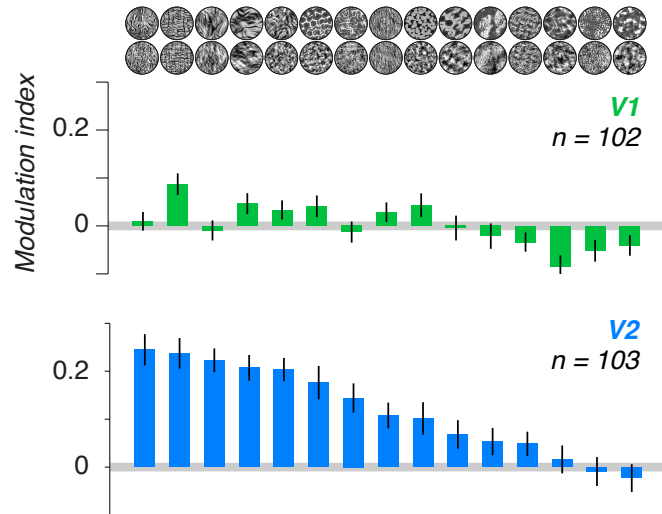
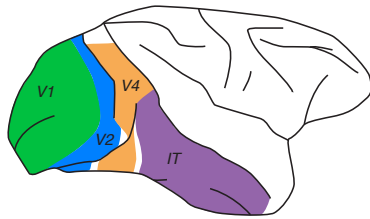
V1, by computing a modulation index for each V1 neuron within an 100 ms window in which weak modulation was present in the population response (Figure 3.18). The late V1 modulation, across texture categories, was reliably correlated with the V2 modulation earlier in the response duration ($r = 0.70$, $P < 0.004$). Furthermore, the late V1 modulation was weakly, though not significantly, correlated with fMRI responses in both V1 ($r = 0.54$, $P = 0.16$) and V2 ($r = 0.47$, $P = 0.074$). Each of these results are anecdotal, but an intriguing and parsimonious summary is that the late component of the V1 response reflects feedback from V2, and due to the sluggishness of the hemodynamics, this is manifest in a weak but reliable fMRI response in V1.

3.5 Possible mechanisms

V2 neurons responded more to images containing naturalistic structure than images that do not. Our unusual approach – generating stimuli to test hypotheses about V2 neurons, rather than directly modeling their responses – helped identify this response property, but also leaves mysterious the functional mechanisms in V2 that might give rise to the observed modulation.

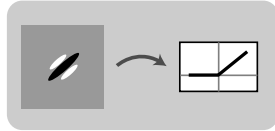
We can begin by identifying mechanisms that *do not* show modulation. First, we return to our simulation assessing the basis of selectivity to shape and curvature from Chapter 2 in which we applied simple LN mechanisms to images (Figure 2.5). Specifically, we compute the outputs of model neurons with either oriented filters, or filters with mixtures of two orientations, both followed by rectification (Figure 3.19). We randomize the orientation of the filters because the stimuli were not adapted to the orientation preference of each neuron. We compute the response of each model neuron to each naturalistic and noise image, and compute modulation

Sensitivity to naturalistic features



Simulations

Oriented filter



Mixture of oriented filters

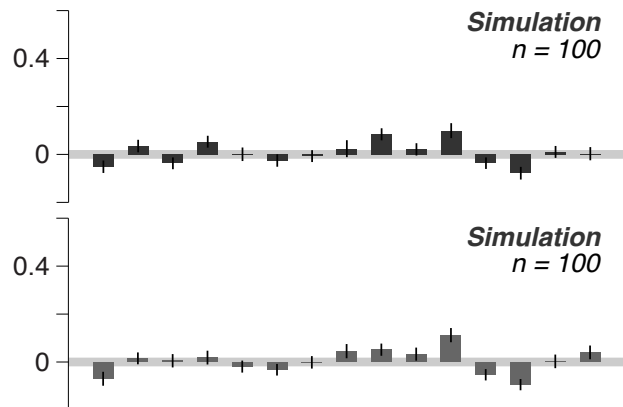
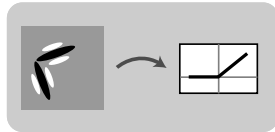


Figure 3.19: Simple LN mechanisms fail to respond differentially to naturalistic stimuli. For each simulation responses were averaged across 100 model neurons tuned to different orientations and spatial frequencies (or mixtures of orientation and frequency).

indices just as we did for the physiological data. For all texture categories, these simple LN mechanisms fail to yield significant modulation, and thus fail to account for the modulation found in V2 (Figure 3.19).

To complicate the model slightly, we use complex cell oriented filters, each of which computes a sum of squared responses of two phase-shifted filters, yielding a local measure of spectral energy (or magnitude) (Figure 3.20). We consider both single complex cell filters, as well as linear combinations of two complex cell filters at nearby locations, but with different orientation preferences. Again, we find little or no modulation. The latter case is particularly interesting because it is the model considered by many efforts to characterize unique properties of V2 neurons [8, 225, 237]. Just as simple V1 neurons are modeled as computing linear combinations of their afferents, and complex V1 neurons as computing sums of squared linear combinations (Figure 2.3), it might be useful to think of the sum of V1 complex cells as a “simple” V2 neuron.

Why do these mechanisms fail to distinguish naturalistic from noise? A complex cell signals local spectral energy. In so far as the two kinds of stimuli are matched for spectral energy over a suitable spatial window, and are relatively homogenous, so long as the receptive field is reasonably large, any spectral mechanism, or linear combination of spectral mechanisms, should yield similar outputs to the two stimulus ensembles. Averaging responses across samples will also help eliminate preferences arising due to inhomogeneity in the spatial structure of the receptive field. Given the nature of the modeling exercise, we certainly cannot claim that *no* linear combination of complex cell responses would produce the observed modulation – it might be possible to achieve by including normalization [97, 195, 36], for example – but at least simple instantiations do not appear to suffice.

Recall that the distinguishing feature of the naturalistic stimuli is the presence of

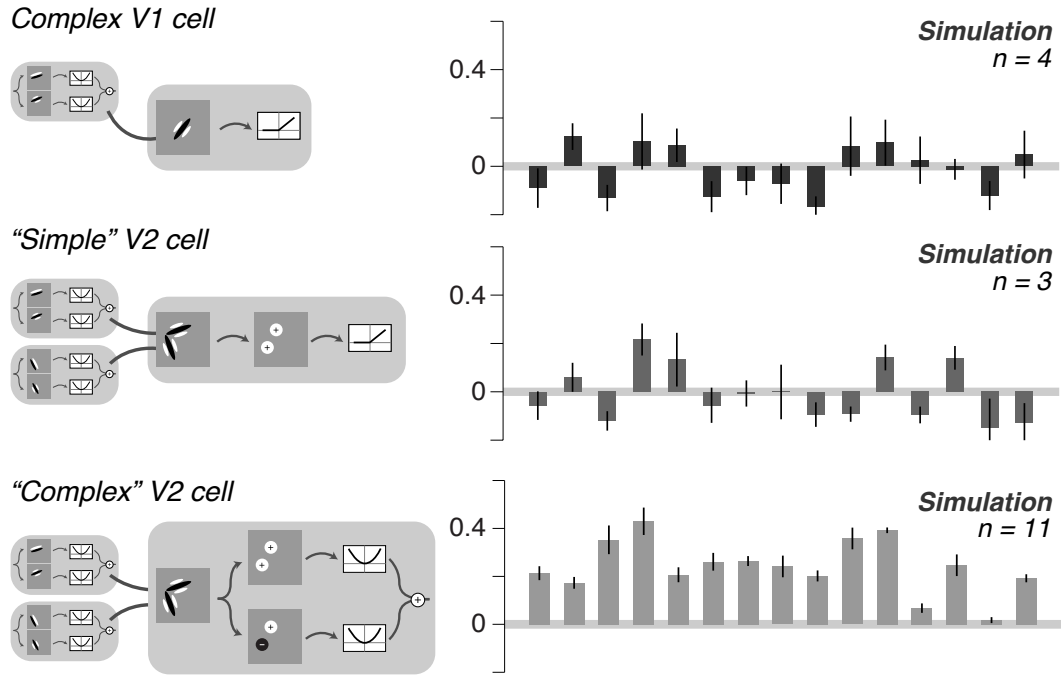


Figure 3.20: “Complex” V2 cells signal the difference between naturalistic and noise. Each simulation began with the responses of V1-like orientation-tuned complex cells (small insets). Taking its response directly yields a complex V1 cell; adding the responses of nearby filters yields a simple V2 cell; combining squared linear combinations of filter yields a complex V2 cell. In each case, the responses of a small set of mechanisms were averaged, each sensitive to different orientations and, for the complex V2 cell, different spatial frequencies.

magnitude correlations, which occur because we imposed spatially-averaged products of filter magnitudes. In relating two models of motion representation – the Reichardt correlational model, and the energy model – Adelson and Bergen [2] showed how products can be computed through a cascade of linear filtering, squaring, and linear filtering. A similar equivalence can be used to construct a mechanism sensitive to magnitude correlations in the domains of space, orientation, and spatial frequency. Specifically, consider two complex cell-like oriented filters, each of which compute a phase-invariant measure of oriented energy, at all locations in an image.

Denote the filter outputs as C_{xy1} and C_{xy2} . At any location, if we take the sum of the filter outputs and square it, take the difference and square it, and then take the difference, it is equivalent to computing a product (up to scale factor). Adding these outputs across spatial locations is equivalent to spatially averaging products,

$$r = \sum_{xy} ((C_{xy1} + C_{xy2})^2 - (C_{xy1} - C_{xy2})^2) = 4 \sum_{xy} C_{xy1} C_{xy2} \quad (3.9)$$

This mechanism is a form of hierarchical convolutional filtering, where the filters (or “subunits”) are simple differencing operators, $[1, 1]$ and $[1, -1]$ applied to complex cell outputs. The filters are applied at each location and, after squaring, are again linearly combined and pooled across space. This example was constructed to explicitly compute the product between the two filter outputs (and thus capture their correlation); the pair of complex cell filters determines the form of correlation (e.g. across scale, position, or orientation). Figure 3.20 shows the result of applying mechanisms like these – 11 in total, capturing different pairwise correlations – to naturalistic and noise images. The mechanisms exhibit modulation. The ranking of the average modulation across texture categories does not match that found in the physiology. However, each of the 11 individual mechanisms exhibits a different ranking (not shown), and if V2 neurons implement mechanisms like these ones, the ranking in the population average would depend on the relative frequency of the different mechanisms.

The above computation is deliberately artificial, but can readily be generalized by considering arbitrary linear filters acting on complex cells tuned to multiple orientations and spatial frequencies. Let C_{xyb} be a three-dimensional representation of complex cell outputs at different locations (x, y) , as well as different bands (b) of a multi-scale multi-oriented pyramid. We can spatially convolve this output with

a filter that computes an arbitrary linear combination of magnitudes, then run the output through a point-wise nonlinearity $g(\cdot)$ (e.g. squaring), and pool the result across space,

$$r = \sum_{xy} g \left(\sum_b C_{xyb} *_{space} k_{xyb} \right) \quad (3.10)$$

We can also allow for arbitrary spatial weights, and multiple filters,

$$r = \sum_{xy} w_{xy} \left(g \left(\sum_b C_{xyb} *_{space} k_{xyb}^1 \right) + g \left(\sum_b C_{xyb} *_{space} k_{xyb}^2 \right) + \dots \right) \quad (3.11)$$

We can further allow for a convolution across subbands rather than space (with appropriate boundary handling),

$$r = \sum_b w_b \left(g \left(\sum_{xy} C_{xyb} *_{bands} k_{xyb}^1 \right) + g \left(\sum_{xy} C_{xyb} *_{bands} k_{xyb}^2 \right) + \dots \right) \quad (3.12)$$

Eqs. 3.11 and 3.12 describe convolution along two separable dimensions. A complete model would allow for a “generalized convolution” and weighting across both subbands and space,

$$r = \sum_{xyb} w_{xyb} \left(g \left(C_{xyb} *_{general} k_{xyb}^1 \right) + g \left(C_{xyb} *_{general} k_{xyb}^2 \right) + \dots \right) \quad (3.13)$$

In such a model, the filter and the form of convolution (and spatial weighting) determine the selectivity and invariance. For example, a filter that combines neighboring orientations coupled with spatial convolution captures cross-orientation dependencies but is invariant to spatial location. A filter that combines neighboring scales coupled with “convolution” in orientation could capture scale dependencies invariant to orientation.

We propose the above family of mechanisms as a description of “complex” V2 neurons, and imagine that V2 neurons implement the computation through the combination of their V1 afferents, possibly across multiple stages of processing within V2. But the proposal relies on a description of V1 neurons as only computing local spectral energy. Many V1 neurons themselves exhibit complex contextual effects, e.g. cross-orientation and surround suppression, as captured by normalization models [37, 97, 101, 195, 36]. As discussed in Chapter 2, some of these computations can be described by sums and differences (and possibly division) of squared outputs of filters tuned to multiple orientations and positions. Such inputs from V1 could take the place of the initial linear combinations of complex cell outputs in the above mechanisms, and these inputs, alongside a simpler combination rule in V2, could give rise to the same sensitivity. An interesting hypothesis, then, is that a single computation is distributed across the two areas, each of which contain neurons with varying degrees of complexity in how they pool their inputs; but the most complex computations still only arise in V2. Also interesting is how normalization in the second stage – subtractive or divisive interactions across the proposed subunit outputs – could contribute to the selectivity and invariance properties of the proposed mechanism. Recent efforts to incorporate normalization in the filter-rectify-filter model [231, 95] could prove useful in guiding this exploration.

An important avenue for future work is developing methods for directly fitting such a nonlinear, hierarchical model to responses of individual V2 neurons [203, 187, 186, 226]. Crucial to that effort will be generating stimuli that probe the dependencies potentially captured by the model. Although the Portilla and Simoncelli textures exhibit many of the relevant magnitude dependencies, a more elegant approach would be to develop a synthesis-by-analysis algorithm for this particular model, thereby yoking the experimental stimuli directly to the model fitting. Com-

puting gradients of these nonlinear subunit mechanisms, with respect to either the filters or the complex cell coefficients, is relatively straightforward, and could be used both to fit filters required to predict the response of a given neuron to a large ensemble of stimuli, or to “fit” the stimulus that predicts a population of neuronal responses (e.g. as computed on an original image). In the latter case, the responses of neuronal populations act effectively as image statistics, and could be used online during experiments to synthesize targeted experimental stimuli [134, 38, 108]

3.6 Discussion

We have discovered a family of image features that modulate the responses of neurons in area V2, while having only a minimal effect on neurons in area V1. This modulation of activity in V2 was strong, and similar, in both anesthetized macaques and awake humans.

Previous studies have identified specialized response properties in subpopulations of V2 neurons, but the differences between V2 and V1 are usually small [171, 129, 141, 103, 64]. One attribute that has robustly distinguished V2 from V1 is border ownership [240], which seems, like our effect, to depend on signals from the receptive field surround in V2 [48, 86]. Border ownership signaling, however, may also rely on attentional feedback [176, 67], whereas the response pattern we have discovered probably does not, as it is evident in both awake humans with diverted attention and anesthetized macaques.

We compared responses to naturalistic texture stimuli with responses to spectrally-matched noise images, similar to the globally phase-randomized images that have been used previously in fMRI [142], psychophysics [210], and physiology [70] experiments. Comparing fMRI responses for intact and phase-randomized

objects, for example, reveals differential responses throughout the human lateral occipital cortex [142]. But none of these studies distinguished responses in V1 and V2. This may be due to the use of uncontrolled images of natural objects or scenes [69, 187], which obscures the influence of higher order correlations. The spatial homogeneity of our stimuli, coupled with a synthesis method that enabled the generation of an unlimited number of images from each category, facilitated the comparison between neurophysiology (averaging across neurons with different receptive field locations), and human fMRI [99]. Synthetic naturalistic stimuli like ours thus offer a balance between natural and artificial that may prove useful in physiological characterization in other sensory domains.

The naturalistic stimuli contained correlations among V1-like filter outputs, which were imposed during synthesis of the naturalistic images [175], but which were absent from the otherwise matched noise images. It is tempting to hypothesize that V2 neurons directly encode these correlation parameters. However, as discussed above, a family of nonlinear computations on V1-like outputs, similar in function but differing in detail, can effectively capture the same correlations, and would create the sensitivity to naturalistic stimuli that we found in V2. Specifically, selectivity for correlations could be achieved by combining squared and spatially pooled linear combinations of appropriate V1 inputs, analogous to those used to compute contrast or motion energy [106, 2, 187, 24]. Such complex cells in V2 would give enhanced responses to stimuli containing higher-order correlations, unlike V2 simple neurons which linearly combine the output of orientation-tuned filters [8, 237, 225]. This hypothesis is conceptually satisfying because it suggests that nonlinear computations of identical form reappear at multiple stages of the cortical hierarchy [101, 36].

In the language of Marr, this Chapter has focused largely on implementation

and algorithm. As information ascends the cortical hierarchy, the computational goal is presumably not representing local texture patches, but enabling the perception of scenes and objects. A common view, as discussed in Chapter 1, is that early computations encode the primitive elements of which scenes are made, and that subsequent stages of processing assemble these elements into larger and more complex combinations, capturing the structural relationships that determine the visual world. This constructionist view has stumbled on the problem of V2, whose neurons have stubbornly refused to reveal the form of their elementary feature combinations [171, 116, 8, 225, 129, 141, 102, 103, 64, 225], perhaps because the set of potential local feature combinations is vast, and particular images contain only sparse samples from this set. Texture stimuli, such as those used here, can facilitate the search by enabling the presentation of dense arrays of features under experimental control. But they also suggest that the representation of two fundamental constituents of visual scenes [1] – the specific feature combinations that comprise objects, and the statistics of local features that characterize textures – may both reside in V2.

This thesis will not answer the question of how V2 responses subserve object recognition, but in the next two Chapters, we will provide evidence linking the novel V2 responses described here to perceptual capabilities for recognizing naturalistic stimuli, including the discrimination of simple homogenous texture patterns (Chapter 4) and the perception of complex natural scenes (Chapter 5).

Notes

⁸We selected these 15 categories somewhat arbitrarily, but emphasized categories for which naturalistic images differed from spectrally-matched noise images, while ensuring there was vari-

ability in the extent of that difference. We return to this point when describing a much larger distribution of naturalistic stimuli in Chapter 4.

⁹Results were nearly identical when using a quantitative criterion based on the standard deviation of the response.

¹⁰In this, and all other plots, that relate two dependent measures, lines show fits using total least squares, which is more appropriate than plotting the best fitting regression line, which assumes no variability along the independent dimension; but the line is shown only for visualization.

¹¹When performing statistical tests that assume normality on correlation coefficients, we applied Fisher's normalizing Z-transform, $r' = (1/2) \log((1 + r)/(1 - r))$.

Chapter 4

Linking perception and physiology through V2

4.1 Introduction

The responses of visual neurons support and constrain perceptual capabilities. In the primate, physiological investigations into the early visual pathways were complemented, and in some cases motivated, by psychophysical studies of sensitivity to contrast, spatial frequency, and orientation (reviewed in [58, 87, 89]). Our understanding of the functions of extrastriate cortex has similarly been guided and influenced by finding links between neuronal responses and behavior, notably in the cases of disparity-selectivity in V2 [155, 156, 157], direction-selectivity in MT [153, 198, 27, 26, 96], selectivity for self-motion in MSTd [93], attentional modulation in V4 [45], and object invariances in inferotemporal cortex [115, 135]. In humans, fMRI has been used to relate surround suppression, pattern detection, and contrast sensitivity to responses in early visual areas [239, 178, 22], and to relate effects of attention to response modulation throughout visual cortex [170].

Establishing such links is of particular interest in areas that encode complex and behaviorally-relevant image features, of which V2 may be the earliest example. Nevertheless, other than disparity, there have been few perceptual properties attributed specifically to V2 neurons. There is an extensive literature on the perception of second-order patterns, which consist of distinct regions containing different texture patterns, the boundaries of which cannot be detected with linear mechanisms [18, 88, 124, 222, 43, 126, 127]. Perception of these stimuli has been examined alongside a two-stage modeling framework – the filter-rectify-filter model – that can extract second-order texture boundaries (Figure 2.15). The two stages of the model suggest processing in V1 and V2. But electrophysiological evidence in macaque suggests that sensitivity to second-order form is comparable between V2 and V1, and is present in a minority of neurons in each area [63]. fMRI studies have similarly shown that sensitivity to these patterns, while present (and comparable) in V1 and V2, is more pronounced in higher areas (but see [95]). Thus, while the extraction of texture-defined boundaries may be a crucial component of scene segmentation, it does not appear to describe the transformation from V1 and V2. Furthermore, it is difficult to link this particular class of laboratory stimuli to the statistical structure of natural images.

The synthesis-by-analysis procedure discussed in Chapters 2 and 3 provides a tool for generating controlled families of experimental stimuli with complex naturalistic statistical structure. Previous studies have explored the perception of these texture stimuli; for example, by assessing the importance of different subsets of parameters for yielding synthetic textures that are visually similar to an original [12]. But perception of these stimuli has not been linked to any underlying physiology. In Chapter 3, we used synthetic naturalistic textures to study neuronal and fMRI responses in V1 and V2. We found that V2 neurons responded consistently

more vigorously to these stimuli than to spectrally-matched control stimuli lacking naturalistic structure.

Here, we describe the results of behavioral experiments that link the physiological responses from Chapter 3 to perception. First, we found that variability in differential response in V2 across texture categories predicts variability in perceptual sensitivity. We accomplished this by titrating the inclusion of higher-order statistics so as to measure fine-grained differences in sensitivity to naturalistic features. We then used a crowdsourcing technique to measure perceptual sensitivity to a much larger ensemble of textures and identified image properties contributing to perceptual sensitivity and, we infer, physiological responses in V2. Finally, we analyzed population responses in V2 to show that the neuronal representation in V2, more so than in V1, supports the perceptual similarity of statistically-matched textures. This provides a neuronal substrate for a novel form of statistical invariance that will motivate the perceptual experiments described in Chapter 5.

4.2 Perceptual discrimination

In Chapter 3, we showed that V2 neurons responded consistently more to synthetic stimuli than to spectrally-matched control stimuli (“noise”) lacking naturalistic structure. But the degree of that differential response varied reliably with the category of texture. If perception depends on neuronal signals in V2, then texture categories that evoke a larger differential response should be those for which the naturalistic and noise images are more perceptually distinct.

For all of our texture categories, the naturalistic and noise images are reliably distinguishable with near perfect accuracy. Nevertheless, some pairs may appear

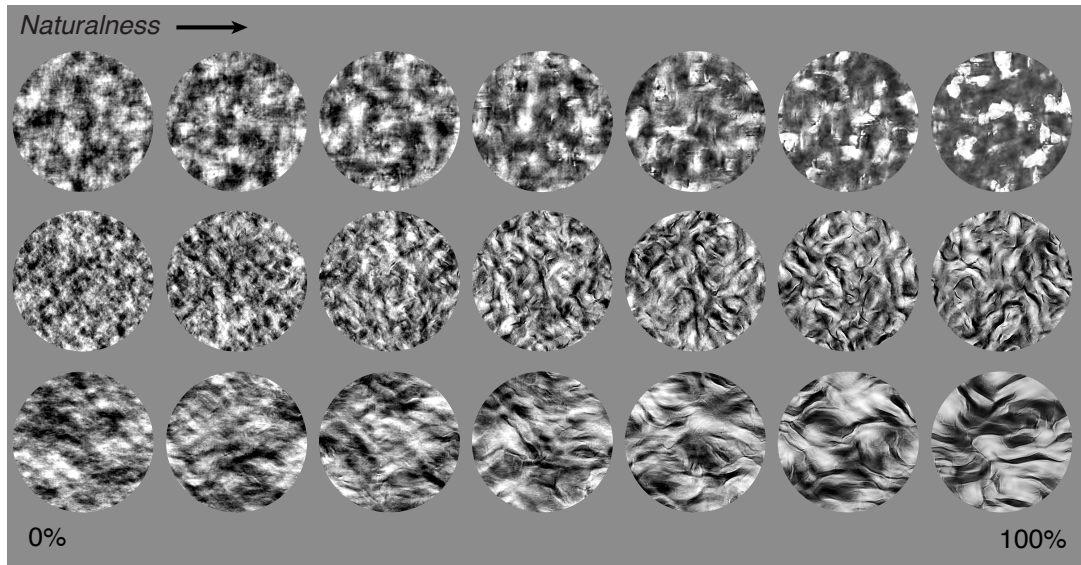


Figure 4.1: *Titration of the inclusion of higher order correlations during image synthesis yields images spanning a “naturalness” axis between noise and naturalistic; examples are shown for three texture categories, with high, medium, and low naturalness thresholds (from top to bottom).*

more distinct than others. Several approaches exist for measuring perceptual distinctiveness [143, 211]. We chose to generate a family of stimuli that interpolated between the two extremes, by titrating the inclusion of the higher-order statistics. This yielded images spanning an axis which we call “naturalness” (Figure 4.1). We measured how well observers could discriminate naturalistic from noise for different levels of naturalness, and identified the threshold required to attain a criterion level of performance. Under the assumption that just noticeable differences (JNDs) are approximately constant for any pair of discriminations along the naturalness axis, threshold will be monotonically related to the perceptual distance between the end-points. Below, we describe the details of the image generation and the perceptual task.

4.2.1 Psychophysical methods

Observers

Three observers with normal or corrected-to-normal vision participated in the experiments (all male; age range, 26-30 years). Protocols for selection of observers and experimental procedures were approved by the Human Subjects Committee of New York University. One observer was the author of this thesis, and another was a collaborator on the project. The third was naive to the purpose of the experiment.

Stimuli

To generate synthetic stimuli spanning a “naturalness” axis between naturalistic and spectrally-matched “noise”, for each texture category we computed the model parameters \vec{p}_{nat} on the original natural photograph and parameters \vec{p}_{noise} on a spectrally-matched noise image, and then linearly interpolated the model parameters between the two endpoints, $\vec{p}_{interp} = \delta \vec{p}_{nat} + (1 - \delta) \vec{p}_{noise}$. For each linear interpolation – each value of δ – we used the synthesis procedure described in Chapters 2 and 3 to generate 15 image samples. Pilot experiments suggested that the distribution of thresholds across texture categories was approximately normally distributed in the log domain, so we sampled the naturalness axis with points δ equally spaced on a logarithmic scale. We refer to naturalness as a percentage – e.g., 10% naturalness means $\delta = 0.1$.

Stimuli were presented on a 41×30 cm flat screen CRT monitor at a distance of 46 cm. Texture images were presented within vignetted 4° circular patches at three locations equidistant from fixation, each 4° eccentricity (one above fixation, one to the lower left, and one to the lower right). A 0.25° fixation square was shown throughout the experiment.

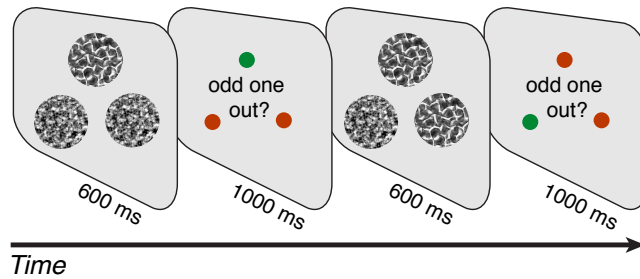


Figure 4.2: *In each trial of the 3AFC “oddbity” task, observers saw three images – two noise and one naturalistic, or vice versa – and indicated the odd one out.*

Task

Every trial of the 3AFC “oddbity” task presented three different images in the three patches: two images were noise and one was naturalistic, or one was noise and two were naturalistic (Figure 4.2). The observer judged which was the odd one out. All three images were distinct synthetic samples, so observers needed to make their judgements based on the similarity or difference in the statistical structure of the images, not by comparing images pixel-by-pixel. Oddity tasks are useful in cases like this, where individual stimuli reflect stochastic samples from statistically-defined categories [104, 12].

The naturalness of the naturalistic image(s) varied across trials, between 4% and 80%. If two naturalistic images were presented on a trial, they had the same level of naturalness. Images were presented for 600 ms, after which observers had 1 sec to indicate their response with a keypress. There was no feedback. Before the experiment, each observer performed a small number of practice trials (≈ 10) with feedback to become familiar with the task. Different texture categories were run in separate blocks. Each observer performed 480 trials in a block; the order of conditions and location of the target were appropriately randomized and counterbalanced. Blocks were performed in random order for each observer.

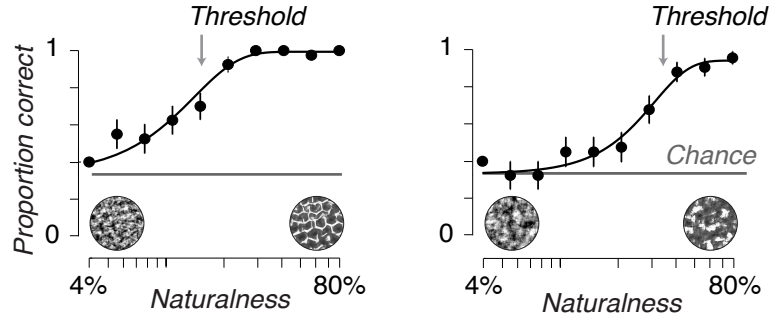


Figure 4.3: Performance in the 3AFC task as a function of naturalness was fit with a cumulative Weibull (black line) by maximizing likelihood. Chance performance was $1/3$. Functions are shown for a single human observer, for two texture categories with high sensitivity (left) and low (right), defined as $1/\text{threshold}$.

Analysis

For each texture category, we fit the parameters of a cumulative Weibull function that maximized the likelihood of the psychometric data. The function was parameterized with a threshold, slope, and lapse rate [235]. Estimated lapse rates were typically very small (mean 1%, maximum 6%). Threshold was converted to its reciprocal (sensitivity) for all subsequent analyses, and statistics, e.g. correlations, were computed in the log domain.

4.2.2 Relating physiology and perception

We used the above psychophysical procedure to measure perceptual threshold for each of our 15 texture categories (examples in Figure 4.3). For each texture category, we also computed the modulation index for each neuron – the difference between the response to naturalistic and noise, divided by the sum (from Figure 3.5).

Across the categories, perceptual sensitivity – the reciprocal of threshold – was significantly correlated with modulation, averaged across neurons, in V2 (Figure 4.4,

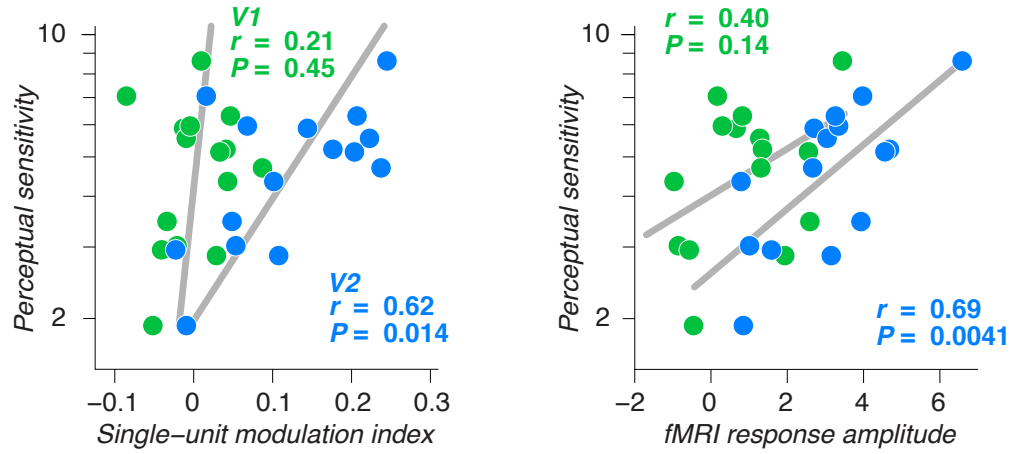


Figure 4.4: (Left) Across 15 texture categories, perceptual sensitivity ($1/\text{threshold}$, averaged across three human subjects) was significantly correlated with the single-unit modulation index in V2 (blue) but not in V1 (green). (Right) Similar results were found for fMRI response amplitudes.

$r = 0.62$, $P < 0.05$) but not in V1 ($r = 0.21$, $P = 0.45$). We also computed these correlations separately for each neuron, and found that correlations for V2 neurons were significantly larger than correlations for V1 neurons ($P < 0.0001$, t -test on Fisher Z-transformed correlations).

Ideally, we would have measured each neuron's response modulation for different levels of naturalness and used an ROC analysis to construct a neurometric function and derive thresholds [153, 27]. The following argument, however, suggests that modulation at 100% naturalness should vary monotonically with threshold. In our existing data, recorded at 100% naturalness, ROC analyses (not shown) showed that single V2 neurons discriminated between naturalistic and noise with less than 100% accuracy; even for the neurons and categories with the strongest modulation, accuracy was typically 60 – 70%. Thus, without having measured them, we know neurometric functions would fail to saturate at 100% naturalness. If we further assume that the slopes of all neurons' neurometric functions are comparable, then

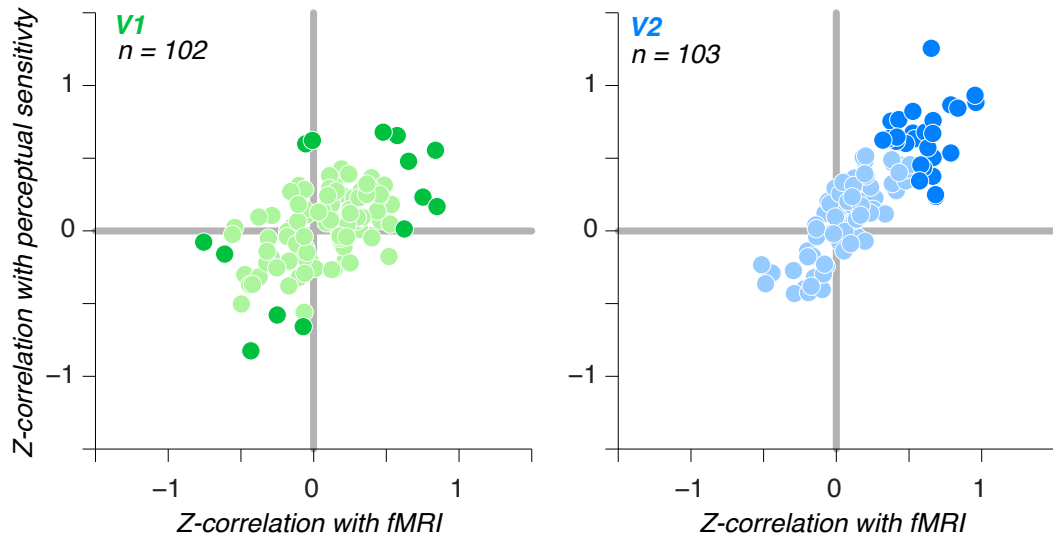


Figure 4.5: The response of each neuron in V1 (left) and V2 (right) to the 15 categories was correlated separately with fMRI response amplitudes and perceptual sensitivity. Dark shading indicates neurons for which either correlation was significant ($P < 0.05$).

larger modulations at 100% should correspond to lower thresholds, justifying the above comparison.

We similarly compared perceptual sensitivity and fMRI response amplitude. Sensitivity was significantly correlated with fMRI responses in V2 (Figure 4.4; $r = 0.69$; $P < 0.005$) but not in V1 ($r = 0.40$; $P = 0.14$). We also computed, separately for each subject, the correlation between fMRI responses in each area and average perceptual sensitivity, and found that correlations were significantly higher in V2 than V1 ($P < 0.005$, paired t -test on Fisher Z-transformed correlations). The hint of correlation in V1 is consistent with the suggestion from Chapter 3 that V1 fMRI responses reflect feedback from V2.

Finally, for each neuron, we examined its correlation with both perceptual sensitivity and fMRI response amplitude. These relationships are depicted in Figure 4.5; each dot is a neuron, and neurons for which either correlation was significant

($P < 0.05$) are shown in a darker color. In V2, more so than in V1, correlations with both variables tended to occur together. Furthermore, there is a clear cluster of points in V2 in the upper right quadrant – neurons with variability in modulation across texture categories that were tightly correlated with both perceptual sensitivity and fMRI responses. Given the functional heterogeneity and diversity of projections from V2 (e.g. signals related to motion and disparity rather than textural form) [204, 154], only a subset of V2 neurons are likely directly involved in perceptual decisions concerning these stimuli.

4.3 Crowdsourcing psychophysics

Although robust, the relationship between perception and physiological response described in the previous section was restricted to 15 image categories. To examine the distribution of sensitivities across a much wider range of natural texture images, we measured perceptual sensitivity for 494 categories using Amazon.com’s “Mechanical Turk”, an internet-based crowd-sourcing method, obtaining approximately 300 hours of behavioral data from thousands of humans subjects (called “turkers”). We then used this ensemble to pick stimuli for a targeted experiment validating the relationships described above, as well as to explore image features contributing to the variability in sensitivity – and presumably neuronal responses in V2.

4.3.1 Methods

Observers

Several hundred turkers were recruited for experiments through Amazon.com’s Mechanical Turk website. Each was paid \$0.40 for approximately 5 minutes of their

time. Payment was made so long as turkers completed the task. Demographic data were not collected. Participation was restricted to those turkers achieving 95% approval rating on other Mechanical Turk tasks.¹² Protocols for selection of turkers and experimental procedures were approved by the human subjects committee of New York University. All turkers signed an electronic consent form at the beginning of the experiment. We ensured that 10 unique turkers completed the task for each texture category, but we did not prevent the same turker from completing the task for multiple texture categories.

Stimuli and Task

We developed a version of our 3AFC task for display in a web browser (see example at <http://www.jeremyfreeman.net/public/turk/code/?csv=tex-018-files.csv>), using Javascript and CSS. Each trial began with 700 ms blank period, followed by a 600 ms stimulus presentation, and a second 700 ms blank period. As in the laboratory version of the experiment, images were presented in three patches equidistant from fixation. A small red fixation dot was shown throughout the experiment. After the second blank, three arrows were presented near fixation pointing towards the three possible target locations. Turkers were instructed that “One image will look different from the other two – your task is to identify it by clicking the black arrow that points to it.” There was no other explanation of the nature of the stimuli or the conditions.

Trial types were similar to those in the laboratory experiment, except naturalness was varied across ten points equally spaced on a logarithmic scale between between 10% and 100%. This range was chosen because pilot experiments suggested moderately higher thresholds compared to the laboratory data. Each turker performed 60

trials, and different texture categories were run separately. There was no feedback during the experiment, but turkers performed 6 trials at the beginning with 100% naturalness, and were told that these initial trials would be easier than the rest.

Given the nature of crowd-sourced experiments, we were unable to control viewing distance, size, or eccentricity. However, we demonstrate below that data obtained from the crowd, and from the lab, were comparable, suggesting that such variations were unimportant, at least with respect to this stimulus and task.

Assessing observer quality

Each turker and texture category yielded a psychometric function, based on six trials for each of ten levels of naturalness. Typically, for each texture category, a small number of turkers performed at or near chance at all naturalness levels, suggesting that they may not have been performing the task appropriately. If data from all turkers were averaged, the influence of these lazy turkers would have yielded fitted psychometric functions with very high lapse rates. As an alternative, we developed an analysis procedure to estimate the quality of turkers and appropriately weight their contribution to estimates of threshold.

For each texture category, we fit data from all turkers with a mixture model.¹³ The model consists of a psychometric function common to all turkers, parameterized with a slope and threshold, and a parameter that controls the quality of each turker. Specifically, for each category, we assume that N turkers perform the psychophysical task. The task contains C conditions (the different levels of naturalness), and there are T trials for each condition. On every trial, the turker provides a response x_{nct} that is either correct ($x_{nct} = 1$) or incorrect ($x_{nct} = 0$). The probability of a response being correct is governed by an turker-independent function $p_c = F(c, \theta)$

which relates the conditions to a probability of correct response via the parameters θ of a cumulative Weibull function. The probability of correct response is also determined by an observer-dependent lapse parameter λ_n which gives the probability that an observer will lapse on any trial, that is, respond randomly rather than according to p_c . Let Θ represent all parameters (those governing the psychometric function, and the lapse rates for all observers). We introduce the latent variable z_{nct} to represent whether or not an observer lapsed on a particular condition/trial combination. We will make use of the indicator variable z_{nctk} : If $z_{nct} = 1$, then $z_{nct1} = 1$ and $z_{nct0} = 0$; If $z_{nct} = 0$, then $z_{nct1} = 0$ and $z_{nct0} = 1$.

Consider a particular observer, trial, and condition. If the observer lapses, she will respond correctly at chance, so the distribution of her response is given by a Bernoulli random variable

$$P(x_{nct}|z_{nct} = 1, \theta) = \gamma^{x_{nct}}(1 - \gamma)^{1-x_{nct}} \quad (4.1)$$

where γ is $1/3$ for the 3AFC task. And if she does not lapse, her response will be governed by the psychometric function,

$$P(x_{nct}|z_{nct} = 0, \theta) = p_c^{x_{nct}}(1 - p_c)^{1-x_{nct}} \quad (4.2)$$

We can use the indicator variables to write the joint distribution over the data and the latent variables as

$$P(x_{nct}, z_{nct}|\theta, \lambda_i) = [\gamma^{x_{nct}}(1 - \gamma)^{1-x_{nct}} \lambda_i]^{z_{nct1}} [p_c^{x_{nct}}(1 - p_c)^{1-x_{nct}} (1 - \lambda_i)]^{z_{nct0}} \quad (4.3)$$

Note the dependence on the marginal probability of a lapse, λ_i . When $z_{nct} = 1$, the above reduces to the first term alone, which is

$$P(x_{nct}|z_{nct} = 1)P(z_{nct} = 1) \quad (4.4)$$

and likewise for the second term. Thus, the expression for the joint uses the indicator variable to capture what is essentially a piecewise combination of Bernoulli distributions.

The complete log likelihood of the data under this model is,

$$\ln P(X|\Theta) = \ln \prod_{nct} (P(x_{nct}|\theta, \lambda_i)) \quad (4.5)$$

$$= \ln \prod_{nct} \left(\sum_z P(x_{nct}, z_{nct}|\theta, \lambda_i) \right) \quad (4.6)$$

Directly maximizing this function with respect to θ and λ_n would be intractable (it is a mixture of Bernoulli distributions). However, note that if the true values of the latent variables were known, maximizing the log likelihood of the data would become linear in the parameters (by taking the log of Eq. 4.3). Thus, this problem is naturally suited to the EM (expectation-maximization) algorithm. Given a current setting of the parameters $\lambda_n^{(t)}$ and $\theta^{(t)}$, we can write the expected log likelihood of the data with respect to the conditional distribution of the latent variables,

$$Q(\Theta|\Theta^{(t)}) = \mathbb{E}_{Z|X, \Theta^{(t)}} [\ln P(X, Z|\Theta)] \quad (4.7)$$

We alternate between computing this expected value, and then estimating the parameters that maximize Eq. 4.7. By the linearity of expectation, it will suffice to compute a point estimate of the expected value of each z_{nct} . To compute that

expected value, we need the probability of the latent variables given a known set of parameter values, which we obtain using Bayes rule,

$$P(Z|X, \Theta^{(t)}) = \frac{P(X|Z, \Theta^{(t)})P(Z|\Theta^{(t)})}{\sum_{Z'} P(X|Z', \Theta^{(t)})P(Z'|\Theta^{(t)})} \quad (4.8)$$

Because $\tilde{z}_{nct} = \mathbb{E}(z_{nct}|x_{nct}, \Theta^{(t)}) = P(z_{nct} = 1|x_{nct}, \Theta^{(t)})$, we need only compute,

$$P(z_{nct} = 1|x_{nct}, \Theta^{(t)}) = \frac{(\gamma^{x_{nct}}(1-\gamma)^{1-x_{nct}})\lambda_n}{(p_c^{x_{nct}}(1-p_c)^{1-x_{nct}})(1-\lambda_n) + (\gamma^{x_{nct}}(1-\gamma)^{1-x_{nct}})\lambda_n} \quad (4.9)$$

Where p_c and λ_n depend on $\Theta^{(t)}$. We now consider the quantity to be maximized,

$$\begin{aligned} \mathbb{E}_{Z|X, \Theta^{(t)}}[\ln P(X, Z|\Theta)] &= \sum_{nct} \tilde{z}_{nct} [x_{nct} \ln(\gamma) + (1-x_{nct}) \ln(1-\gamma) \\ &\quad + \ln(\lambda_n)] + (1-\tilde{z}_{nct}) [x_{nct} \ln(p_c) \\ &\quad + (1-x_{nct}) \ln(1-p_c) + \ln(1-\lambda_n)] \quad (4.10) \end{aligned}$$

where we have used the linearity of expectation to replace z_{nct} with \tilde{z}_{nct} from Eq. 4.9. Differentiating with respect to the parameters of interest yields maximum likelihood estimators. In practice, we want to differentiate with respect to the parameters θ that control p_c , or find maximum likelihood estimates through numerical optimization if the derivatives are non-trivial. But for simplicity and intuition, here we differentiate with respect to p_c directly to obtain a non-parametric estimate of the fraction of correct responses

$$\hat{p}_c = \frac{\sum_{nt} (1-\tilde{z}_{nct})x_{nct}}{\sum_{nt} (1-\tilde{z}_{nct})} \quad (4.11)$$

$$\hat{\lambda}_n = \frac{\sum_{ct} \tilde{z}_{nct}}{CT} \quad (4.12)$$

The expression for \hat{p}_c is simply the fraction of correct responses weighted by the lapse occurrence; if there were no lapses, the denominator would contain only 1s, and Eq. 4.11 would reduce to the number of correct responses divided by the total number of trials. The expression for $\hat{\lambda}_n$ is similarly intuitive: the number of lapse trials divided by the total number of trials (across all conditions). Having obtained these estimates on the M step, they are used on the E step to compute the expectation in Eq. 4.9.

We confirmed that the parameter estimates obtained from this algorithm reliably converged from multiple random initializations. As expected, the analysis estimated high lapse rates for turkers with outlier behavior (e.g. near chance performance in all conditions), and the analysis ensured that these turkers contributed minimally to estimates of slope and threshold.

4.3.2 Perceptual sensitivity in the crowd

We measured perceptual sensitivity in the crowd for 494 categories, obtaining data from 10 subjects per category. This corresponded to nearly 300 hours of psychophysical data collection (300,000 trials!), which would have been exhausting to collect in the laboratory using traditional methods. But this increase in yield came at the expense of experimental control, and we wondered whether thresholds measured from the crowd were consistent with thresholds estimated in the laboratory. Figure 4.6 shows psychometric functions for two example texture categories with different thresholds, measured in the crowd and in the lab. For the crowd-sourced psychophysics, each colored line indicates a psychometric function from a different turker, and the thickness of each line indicates the quality assigned to that observer (1 minus the inferred lapse rate, λ_n). In each case, the dark line indicates the

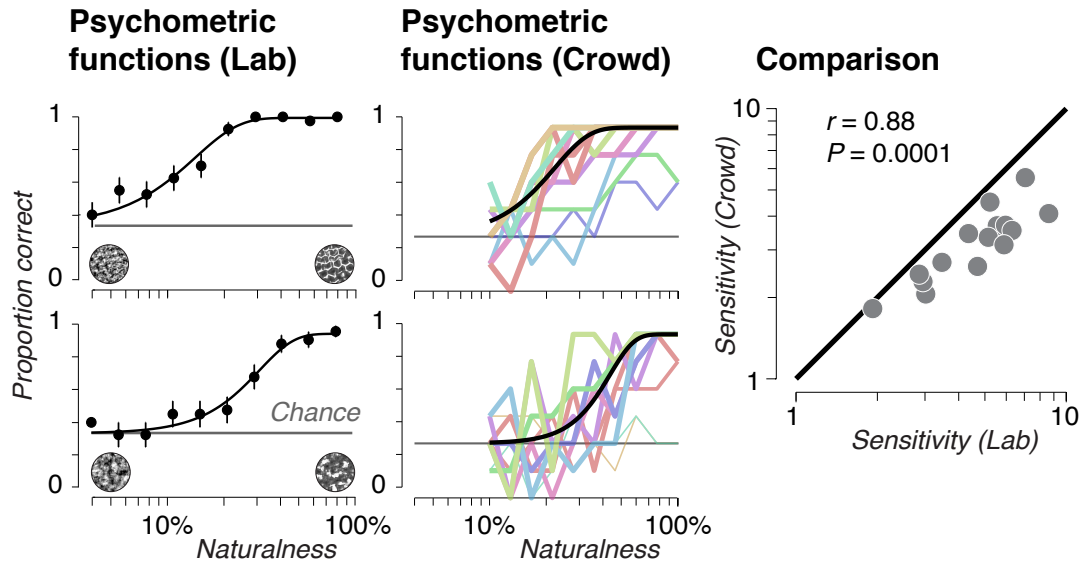


Figure 4.6: Psychometric functions for an example human subject (left) and for 10 turkers (middle). Across 15 categories, sensitivities measured in the laboratory and in the crowd were highly correlated (right).

best fitting Weibull function. Figure 4.6 also shows the correlation between perceptual sensitivity ($1/\text{threshold}$) measured in the crowd and in the lab, for 15 texture categories. The two sensitivities were reliably correlated, albeit systematically lower for the crowd. The correlation of $r = 0.88$ was comparable to the average pairwise correlation obtained among our three human observers ($r = 0.88$). Furthermore, the lower performance in the crowd may have reflected unusual properties of the lab. Two of the lab observers were the author of this thesis and his primary collaborator, both of whom had extensive experience looking at and discriminating these stimuli, and likely had unusually high sensitivities.

Having established the validity of crowd-sourced psychophysics, we considered the distribution of sensitivity across the 494 categories, grouping them into groups of low (0-25th percentile), medium (25th-75th), and high sensitivity (75th-100th) (Figure 4.7). The 15 categories used for the experiments described in Chapter 3

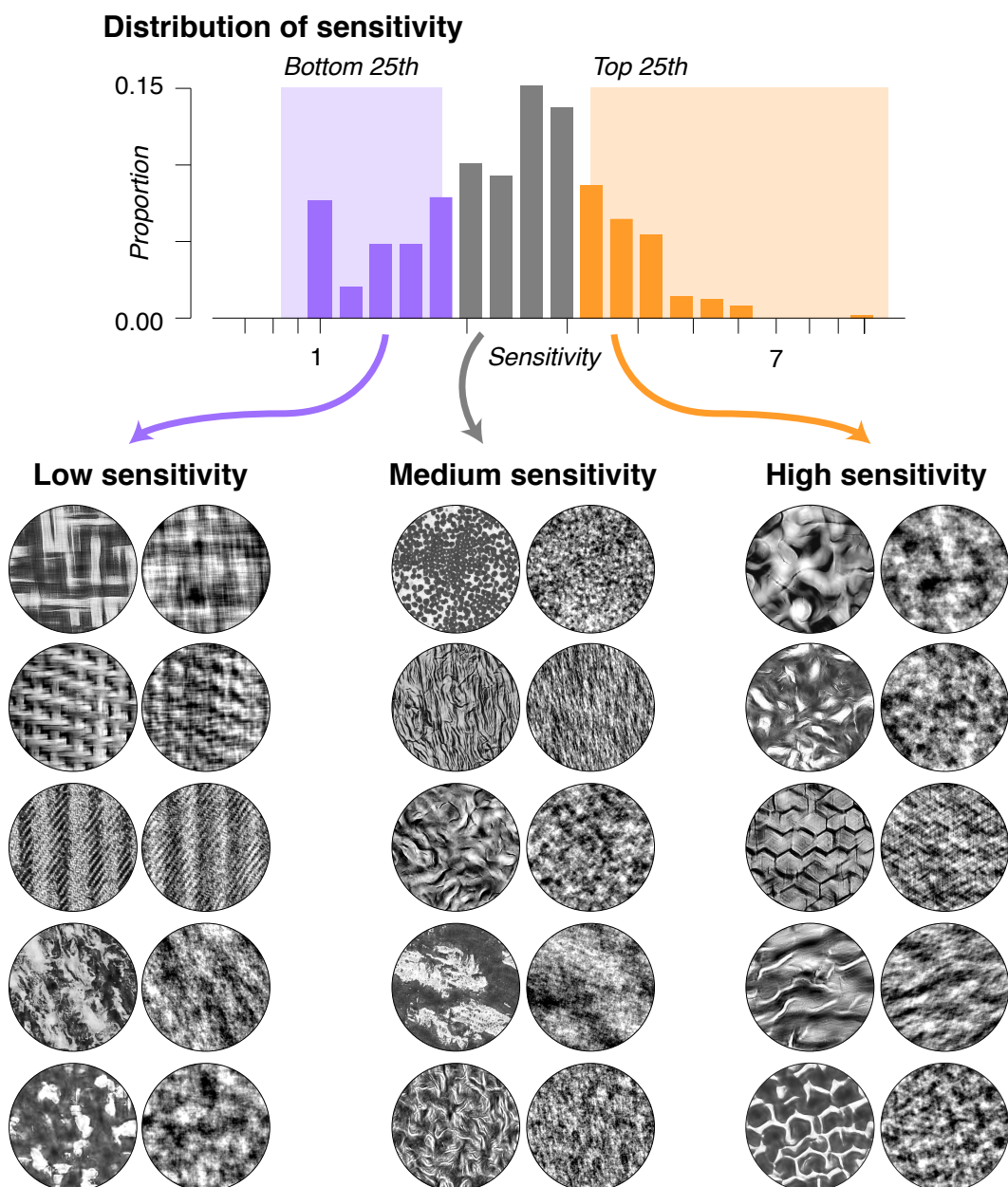


Figure 4.7: *Distribution of perceptual sensitivity, and example images from groups with low, medium, or high sensitivity.*

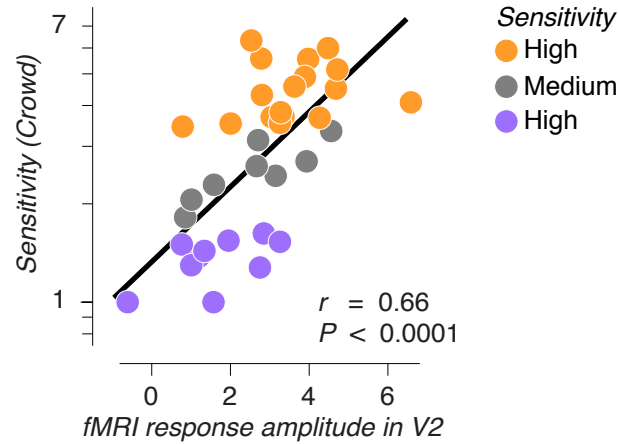


Figure 4.8: *fMRI response amplitudes in V2 were measured for 20 stimuli selected from the groups in Figure 4.7 to span a range of sensitivities. Responses were averaged across two human subjects. Data from Figure 4.4 for 10 categories are replotted here.*

spanned a range comparable to that of the full distribution, but tending towards higher sensitivities; of the original 15, 8 fell within the medium sensitivity group, and 7 in the high. This is not surprising given how we selected those 15 categories; as noted in Chapter 3, we informally emphasized textures for which the naturalistic and spectrally-matched noise images looked at least somewhat different.

We leveraged the full ensemble of sensitivities to generate a targeted set of experimental stimuli, and validate the previously described relationship between sensitivity and physiological response (Figure 4.4). We selected 10 images from the high sensitivity group, and 10 from the low. Because the original relationship was robust for both single-neuron responses and fMRI, we used fMRI for the validation. In two subjects, using the same methods described in Chapter 3, we measured response amplitudes for the 20 texture categories in V1 and V2. Once again, response amplitudes were much larger in V2 than V1 ($P < 0.0001$, paired t -test). Figure 4.8 reports the correlation between sensitivity and fMRI response amplitude in V2 for

the 20 new categories. We also include in the plot responses from the same two subjects for the original 15 images (colored according to their sensitivity group, and with sensitivity estimated from the crowd rather than the laboratory). The correlation for these 35 categories was robust and significant ($r = 0.66$, $P < 0.0001$), confirming and extending the relationship. Interestingly, the correlation between sensitivity and fMRI response amplitude was also significant in V1, albeit half the magnitude ($r = 0.34$, $P = 0.034$). As discussed above, some residual correlation in V1 is expected if the V1 responses reflect feedback from V2.

4.3.3 Predicting diversity of sensitivity

Figure 4.7 shows example images from the different sensitivity groups. No obvious properties distinguish the groups, except perhaps that the high sensitivity images have richer, more complex, and edgier patterns. An artist looking at the images had a more poetic take: “I couldn’t agree more with the preferences. The bad textures made me think of rough cloth on the skin, worms wriggling, vomit, allergies. The good textures were all solar flares, heavens, magic, steam, lava, movement and liquid” (Meredith Leich, personal communication).

We leveraged the large ensemble of sensitivities to provide a quantitative account of what makes the high sensitivity images unique. Specifically, we considered whether the parameters of the Portilla and Simoncelli model, which govern the synthesis of the naturalistic images, could predict the observed variability in sensitivity. This is not straightforward because of the dimensionality of the model. It contains hundreds of parameters, and there are redundancies both within and across parameter groups [175, 12]. We simplified the problem, however, by considering the natural “groups” into which the parameters are organized. That reduction in

dimensionality, combined with the constraints of perceptual sensitivity from nearly 500 images, made the problem tractable.

We began by computing all the parameters for each of the naturalistic texture images. It sufficed to compute the parameters for only one sample in each category, because multiple samples had identical parameters, by design. Many of the parameters have scales that depend, often trivially, on the scales of the underlying quantities; parameters controlling low frequencies, for example, are larger because of the arbitrary normalization of the steerable pyramid. To control for these differences, parameters were Z-scored so that, for each parameter, the mean of its value across the images was 0, and the standard deviation was 1. We then grouped the parameters as follows: (1) pixel statistics (mean, variance, skew, and kurtosis in the pixel domain), (2) low-pass marginal statistics (skew and kurtosis of multiple spatial frequency bands), (3) products of simple cell responses at neighboring locations, (4) products of complex cell responses at neighboring locations, (5) average energy in each subband (spectral properties), (6) correlations of complex cell responses at neighboring orientations, (7) correlations of complex cells at neighboring scales, and (8) correlations of simple cells at neighboring scales. For each group of parameters g , we constructed the $n \times p_g$ matrix P_g containing the p_g texture parameters in that group for the n texture categories. We then reduced the dimensionality of each group of parameters separately using principal components analysis, projecting each parameter matrix into the space spanned by the first k components, yielding a $n \times k_g$ matrix \hat{P}_g (Figure 4.9). We used the k components required to capture 70% of the parameter variance (typically between 2 and 6).

Having reduced the dimensionality of each parameter group, we obtained a combined predictor matrix X , with n rows and a number of columns that depended on the number of components per group. We added a column of ones to the matrix

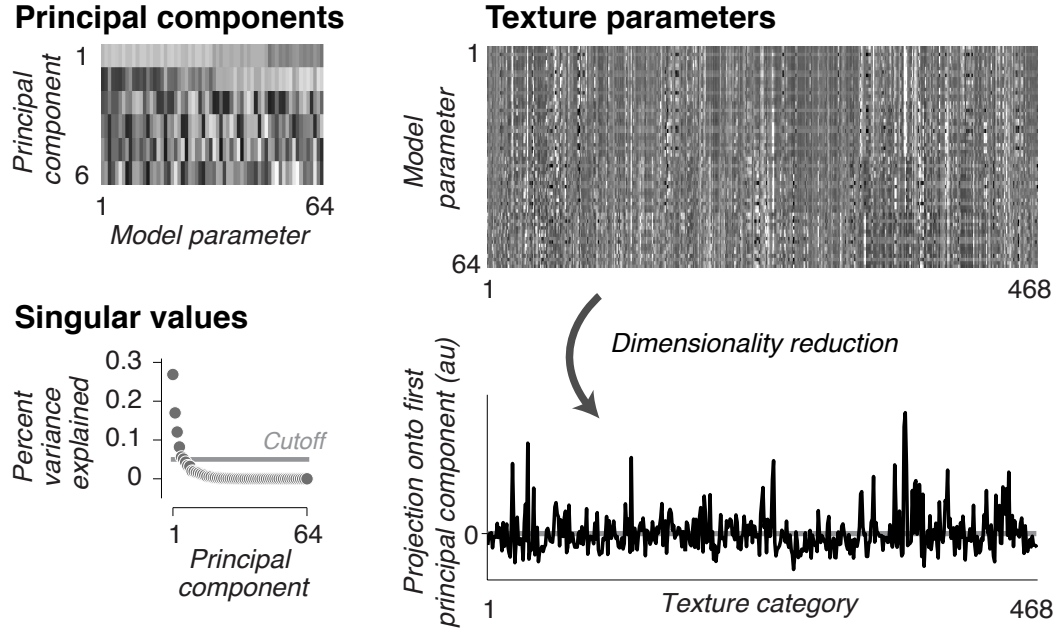


Figure 4.9: Across an ensemble of 468 images, each group of texture parameters was computed (e.g. the 64 cross-orientation correlations), and principal components analysis was used to reduce dimensionality, yielding a small number of predictors for each texture (one predictor shown here, bottom right).

(to account for a constant offset), and used simple least squares regression to solve for the weights $\hat{\vec{b}}$ that minimized the squared error,

$$\epsilon = ||X\vec{b} - \vec{y}||^2 \quad (4.13)$$

where \vec{y} is a vector of log sensitivities for each of the texture categories (as mentioned above, we worked in the log domain because log sensitivities were approximately normally distributed). We removed from analysis any categories where thresholds were estimated as greater than 100% or less than 0% naturalness, to avoid the influence of outliers due to unstable threshold estimates (only 5% of categories).

With 35 predictors from all parameter groups, the linear model predicted 47%

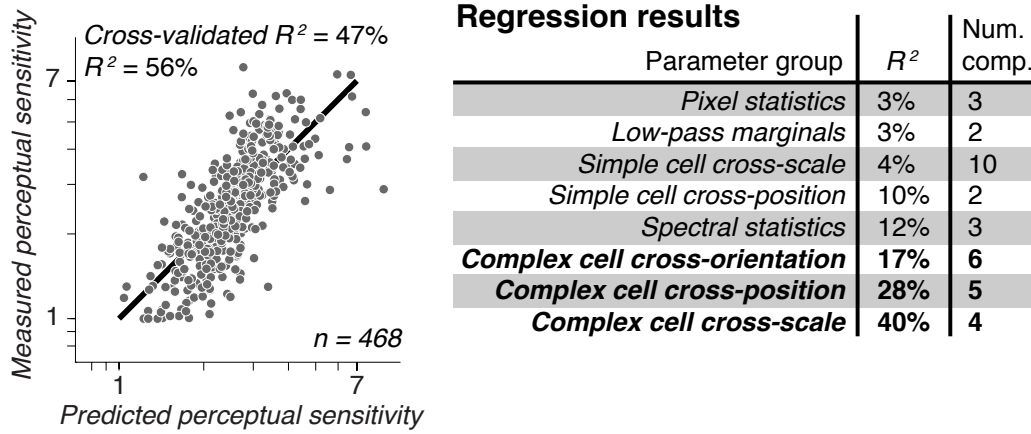


Figure 4.10: (Left) Linear regression was used to predict perceptual sensitivity across 468 images using 35 components derived from the texture parameters, as shown in Figure 4.9. (Right) Regression was performed separately on each group of parameters to identify its explanatory power.

of the variance in perceptual sensitivity (Figure 4.10); 10-fold cross-validation was used to ensure no overfitting. Performance was nearly identical when using only $k = 1$ component per parameter group (only 8 predictors total), but this made it difficult to compare the different parameters (see below), because $k = 1$ component captured more parameter variance for some groups than others.

To compare the relative importance of the different parameter groups, we repeated the regression restricted to subsets of predictors. As reported in Figure 4.10, accuracy was particularly high when using the complex cell cross-scale (40%) and cross-position (28%) statistics. Accuracies were also relatively high for the complex cell cross-orientation (17%) and simple cell cross-position (10%) statistics. The spectral statistics alone yielded an accuracy of 12%. Why are spectral statistics at all relevant to discriminating images that are *matched* for their spectra? Spectral properties can be thought of broadly as controlling visibility. It may be difficult to discern structure in a texture with primarily high spatial frequencies, for example, at

the parafoveal eccentricities used in our experiments, due to the spatial frequency bandwidth of the relevant neuronal populations. As a result, it would also be difficult to discriminate between naturalistic and noise. In this way, spectral properties could predict some of the variability in sensitivity. Finally, the marginal statistics, and the simple cell cross-scale statistics, contributed little (less than 5%). Qualitatively similar estimates of parameter importance were obtained when only using one principal component per parameter group. An important caveat to this analysis is that there is redundancy across parameter groups (as discussed in Chapters 2 and 3). Thus, high accuracy achieved for any given set of parameters may reflect, in part, other parameter groups. However, low accuracy for any group implies that variability in that parameter alone has little predictive power.

As an independent control for the importance of the higher-order statistics, we repeated the entire procedure – measuring statistics, performing principal components analysis, and predicting sensitivity – using the spectrally-matched noise images. This analysis should only reflect the extent to which spectral properties of the image categories predict sensitivity, and indeed, despite including all parameter groups in the analysis, we found an accuracy of only 12%. This accuracy is consistent with the contribution of spectral statistics estimated in the analysis above.

In so far as perceptual sensitivity is related to the overall physiological response difference between naturalistic and noise (Figures 4.4 and 4.8), the above analyses provide evidence that a combination of cross-scale, cross-position, and cross-orientation complex cell dependencies, in that order, may be important for eliciting both high perceptual sensitivity and, we infer, a differential physiological response. The predictive power of spectral properties additionally suggests that interactions between image power spectra and, presumably, the underlying tuning of V2 neurons contributes to the physiological response. As discussed above, this could arise

trivially in so far as spectral properties control visibility. But there could also be non-trivial interactions between spectral properties and the modulation depending, for example, on the orientation and spatial-frequency tuning of V1 neurons providing input to a particular V2 neuron. All of these properties could be assessed further by fitting rich hierarchical functional models to V2 neurons alongside careful measurements of specific functional connectivity between neurons in the two areas.

The poor predictive accuracy of marginal statistics is notable given the control experiment described in Chapter 3, in which synthetic images matched only for the marginal statistics of an original were found to yield reduced, but significant, response modulation in V2 neurons. However, as discussed in Chapter 3, merely matching marginals, especially in images with low spatial frequencies, can induce sharp discontinuities which yield magnitude dependencies across scales. The poor predictive accuracy of the marginals, compared to the cross-scale and cross-orientation statistics (Figure 4.10), suggests that in the electrophysiological experiments, inadvertent imposition of those statistics, and not the marginals *per se*, may have been driving the differential responses. Further physiological control experiments, in either human or macaque, using stimuli matched only for the cross-scale or cross-position statistics, could be used to further explore this hypothesis [12].

The clear importance of magnitude correlations across scale is interesting because it has been largely ignored by previous efforts in V2, which instead focused primarily on interactions among local orientations [8, 225, 236]. As noted in Chapter 3, those approaches also emphasized *linear* combinations of orientation; capturing magnitude dependencies requires additional nonlinearities, even restricted to the domain of orientation (Figure 3.20). But our results further suggest that additional investigations must use stimuli containing rich cross-scale dependencies. That there are such dependencies in natural images is not immediately obvious when focus-

ing on the elementary components that constitute shape, especially compared to the intuitive way in which groups of local orientation appear to form curves and contours. But cross-scale dependencies are readily apparent when analyzing the statistical structure of images in terms of V1-like outputs (Figure 2.8), and appear to play a role in the distinctive functional properties of V2 neurons.

4.4 Perceptual invariance

We have focused thus far on the conditions under which naturalistic images appear *distinct* from spectrally-matched images. An orthogonal, but equally intriguing, perceptual property of these stimuli is the fact that multiple stimuli matched in their statistical properties appear similar. Figure 4.11 shows five images synthesized for each of three texture categories. The five “samples” from each category are matched for the same set of higher-order statistics, but are physically different images. Comparing any two pixel-by-pixel reveals large differences. Despite these differences, samples from each category are perceptually similar, and form clear perceptual groups. Do neuronal responses in V2 support this perceptual property?

In the physiological experiments described in Chapter 3, we measured responses to multiple image samples from each texture category. All analyses described thus far averaged across responses to samples. Below, we describe three analyses that assess the consistency of response across samples, at both the single-neuron and neuronal-population level. In all of these analyses, we focused on responses to naturalistic images, and ignored responses to spectrally-matched noise.

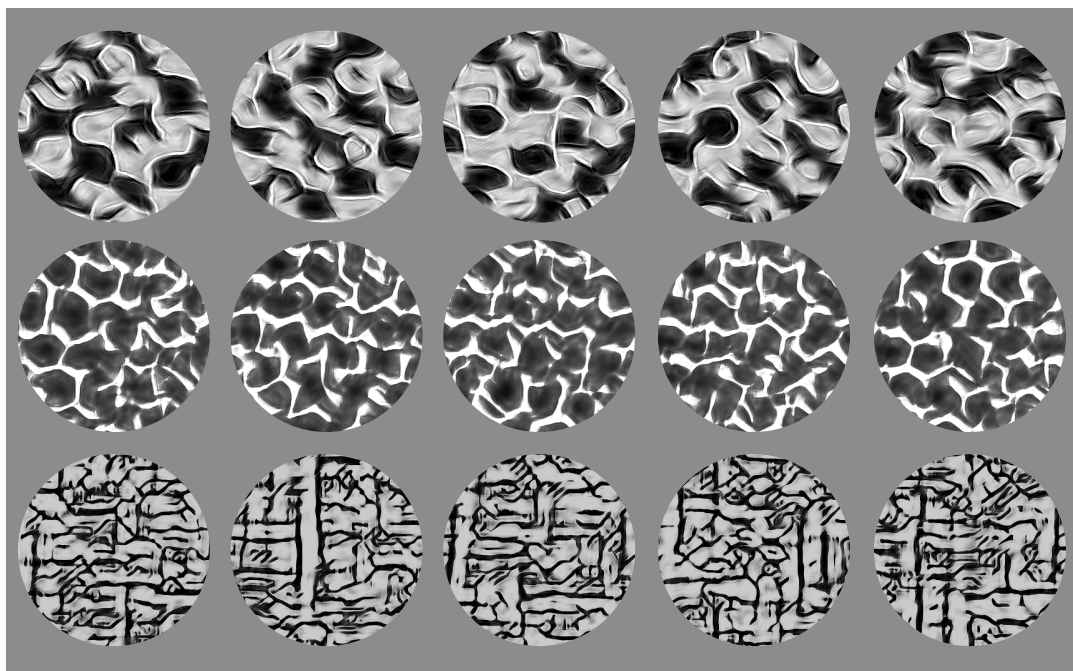


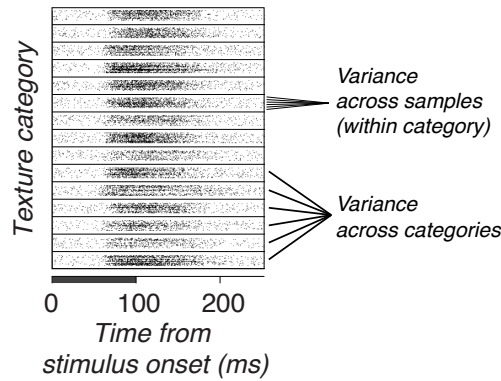
Figure 4.11: *For each texture category, multiple stochastic samples synthesized from different noise images have the same higher-order statistics, but are physically distinct.*

4.4.1 Invariance in single neurons

Do single neurons in V2 or V1 respond similarly to samples from the same texture category? Even to repeated identical stimuli neuronal responses are variable, so different samples are unlikely to elicit identical responses. However, if a neuron's response distinguishes different categories but is to some extent tolerant to sample-to-sample variation, its response should vary less across samples within a category than it does across the different categories.

For each neuron, we measured responses – specifically, spike counts – to 15 samples of each of 15 texture categories, each repeated 20 times. The raster in Figure 4.12 depicts these different sources of variability; the 15 big rows correspond to the 15 categories; within each big row, the 15 small rows are PSTHs, each showing the response to a sample, averaged across the 20 repeats.

Responses of an example V2 neuron



Population summary

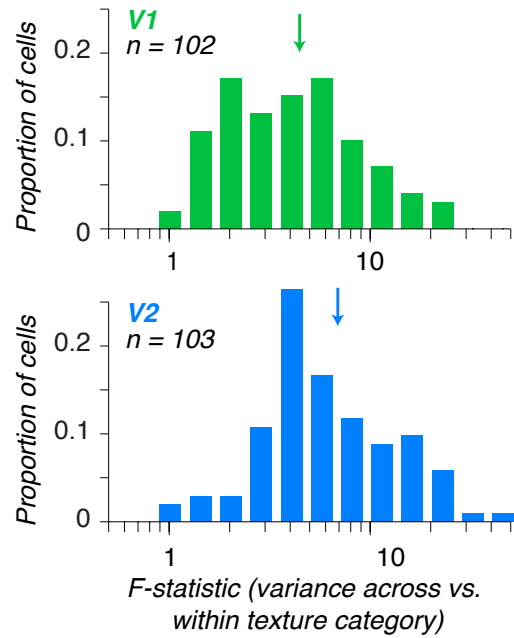


Figure 4.12: For each neuron, responses were measured to multiple repeats of different samples of each of several texture categories (left), and a nested ANOVA was used to partition the sum of squares. An F -statistic (right) captured the ratio of variance across categories to variance across samples within a category.

To assess sample-to-sample tolerance, we partitioned the total variance of the firing rates into components that captured variability across texture categories, across the samples within each category, and across the repeats for each sample. This partition was obtained using a “nested ANOVA”, which estimates the sum of squares at each level of such a hierarchy, and yields F -statistics that capture ratios of variance at each pair of levels [206] (similar to a repeated measures ANOVA, but with an extra level). We focused on the F -statistic that captured the ratio of variances across categories to the variance across samples within a category. This statistic will only be large when variability across samples is small relative to the variability

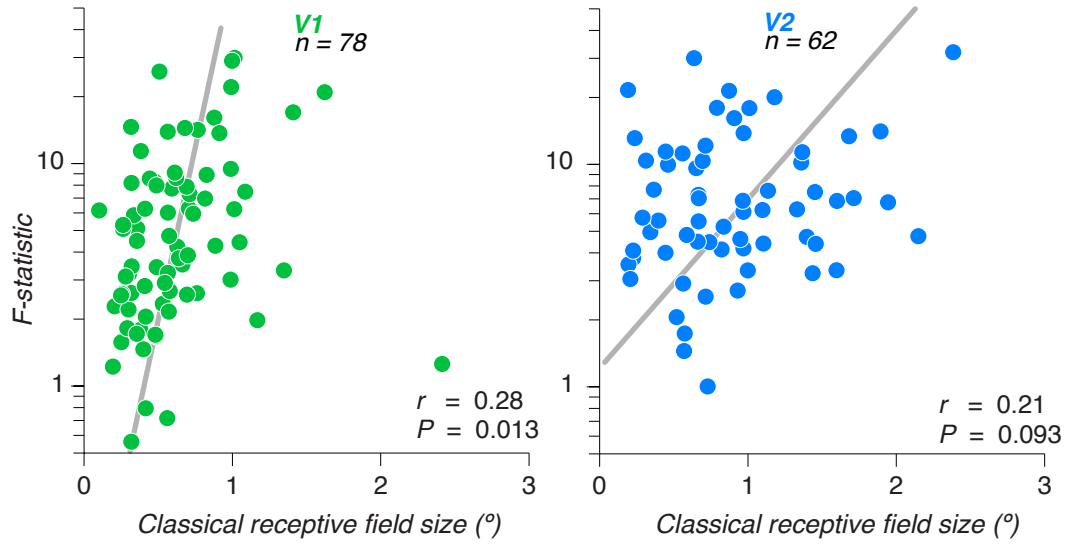


Figure 4.13: In V1, and to some extent in V2, the F -statistic from Figure 4.12 was larger for neurons with larger receptive field sizes.

across categories, which would be a single-neuron correlate of invariant or “tolerant” responses to images of the same category.

Figure 4.12 shows distributions of this F -statistic across our neurons in V1 and V2. We performed the analysis after first applying a Freeman-Tukey variance-stabilizing transformation on the raw firing rates f ,

$$f' = \sqrt{f+1} + \sqrt{f} \quad (4.14)$$

We found a highly significant difference in the F -statistic between V2 and V1, whether computed in the log domain ($P < 0.0001$, t -test) or the linear domain ($P < 0.005$); the log domain seems more appropriate because the F -distribution has a long tail. This suggests that single neurons in V2, more so than in V1, respond invariantly to multiple samples of a texture category. Unlike the modulation, however, this difference between the two areas was due in part to receptive field sizes.

In V1, we found a significant correlation between the F -statistic and receptive field size (Figure 4.13, $r = 0.28$, $P < 0.05$), though we did not find evidence for a relationship in V2 ($r = 0.21$, $P = 0.09$) (both correlations were computed after taking the log of the F -statistic). In the next section, we consider the consequences of this increased invariance at the level of neuronal populations.

4.4.2 Visualizing and quantifying population responses

Single neurons in V2 responded more consistently across samples than neurons in V1. But there was variability in the degree of consistency, and no neurons in either area responded identically to all samples. More relevant to perception is the extent to which the responses of neuronal *populations* in V1 and V2 to images from the same category are similar. Here, we describe two methods for assessing invariance at the population level. One is a visualization, the other a quantification.

Visualizing the responses of large populations is a common problem in sensory neuroscience. Many methods aim to embed high-dimensional data in low-dimensional maps that preserve as much data structure as possible. Common methods include both linear mappings (e.g. principal components analysis and multidimensional scaling), as well as nonlinear mappings (e.g. Sammon Mapping, Isomap, and Local Linear Embedding). A common problem with these methods, when applied to real data sets, is that they cannot simultaneously preserve both global and local structure in the data. We used a recently developed method called t -SNE, which largely overcomes this difficulty [56]. t -SNE is a modification of the Stochastic Neighbor Embedding technique developed by Hinton and Roweis [105]. Briefly, SNE (and t -SNE) describe the relationship among data points in a high-dimensional space, and in a corresponding low-dimensional space, probabilistically.

Adapting the notation from Hinton and Roweis [105], we express our data points as vectors \vec{x} (e.g. containing the spike rate of multiple neurons to a single image). The conditional probability that data point \vec{x}_i would pick \vec{x}_j as its neighbor is given by a Gaussian,

$$p_{j|i} = \frac{\exp(-\|\vec{x}_i - \vec{x}_j\|^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|\vec{x}_i - \vec{x}_k\|^2/2\sigma_i^2)} \quad (4.15)$$

where σ_i^2 is the variance of the Gaussian. A similar conditional distribution is constructed for each of the data points \vec{y} in the two-dimensional space,

$$q_{j|i} = \frac{\exp(-\|\vec{y}_i - \vec{y}_j\|^2/2)}{\sum_{k \neq i} \exp(-\|\vec{y}_i - \vec{y}_k\|^2/2)} \quad (4.16)$$

SNE uses gradient descent to find a set of data points \vec{y} that minimizes the KL divergence between these two conditional distributions. *t*-SNE uses a *t*-distribution in the low dimensional space, rather than a Gaussian; the heavy tails of the *t*-distribution better compensates for the “crowding” that arises when high-dimensional data are packed into a low-dimensional space, and thus better preserves structure at multiple scales [56].

We applied *t*-SNE to our population responses from V1 and V2.¹⁴ The input to the algorithm was a set of 225 data vectors \vec{x}_i , each of which collected the firing rates of all neurons in an area to a stimulus. We also normalized the data so that, for each neuron, responses to the 225 images had mean 0 and standard deviation 1. We ran the algorithm multiple times to ensure convergence and stability of map estimates. Figures 4.14 and 4.15 show *t*-SNE-derived maps for V1 and V2, using pictures to show the different images; Figure 4.16 shows the same results using colored dots to indicate the different images, where each color is a texture category and multiple dots of the same color are samples from the same category.

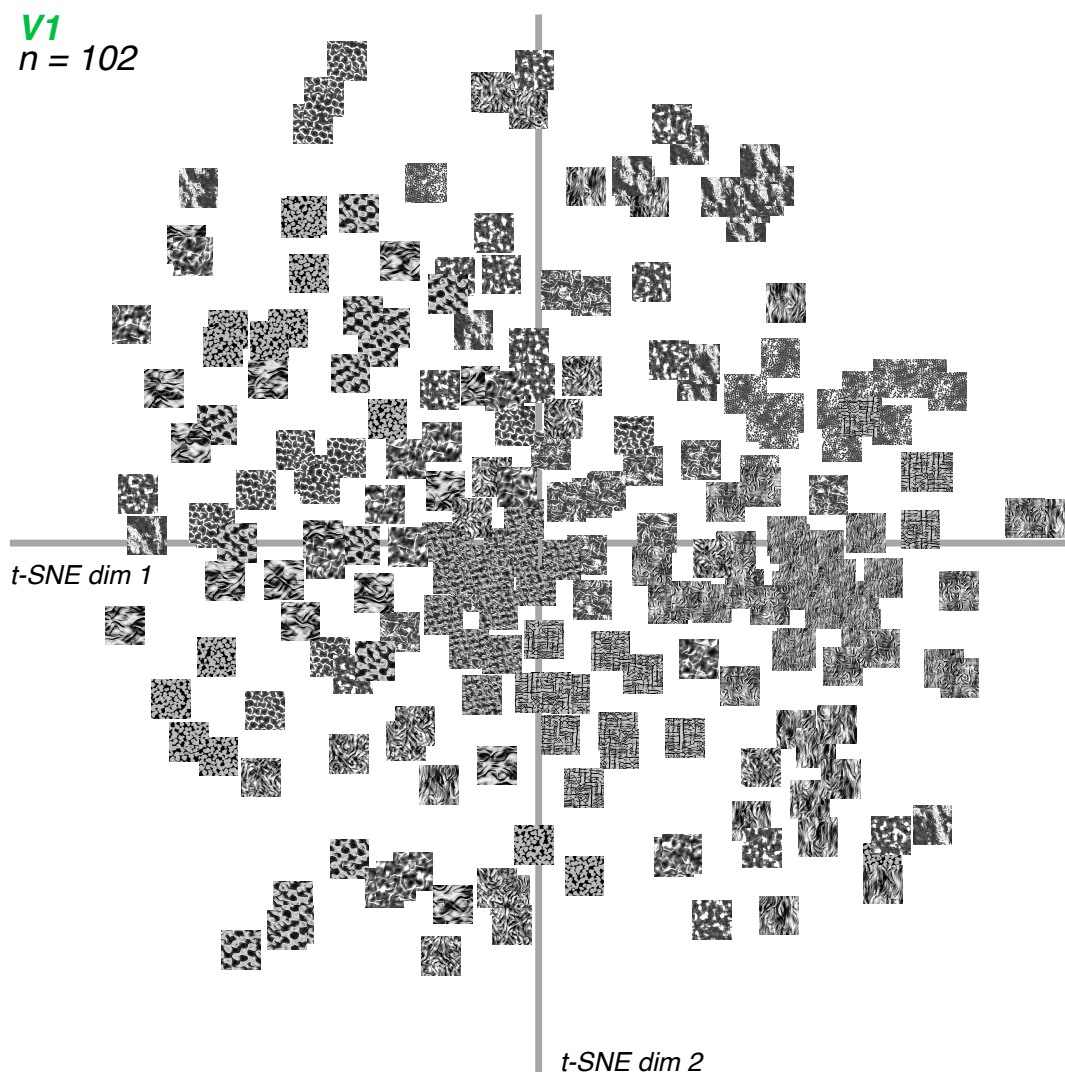


Figure 4.14: *t-SNE map for the V1 population response to 225 images.*

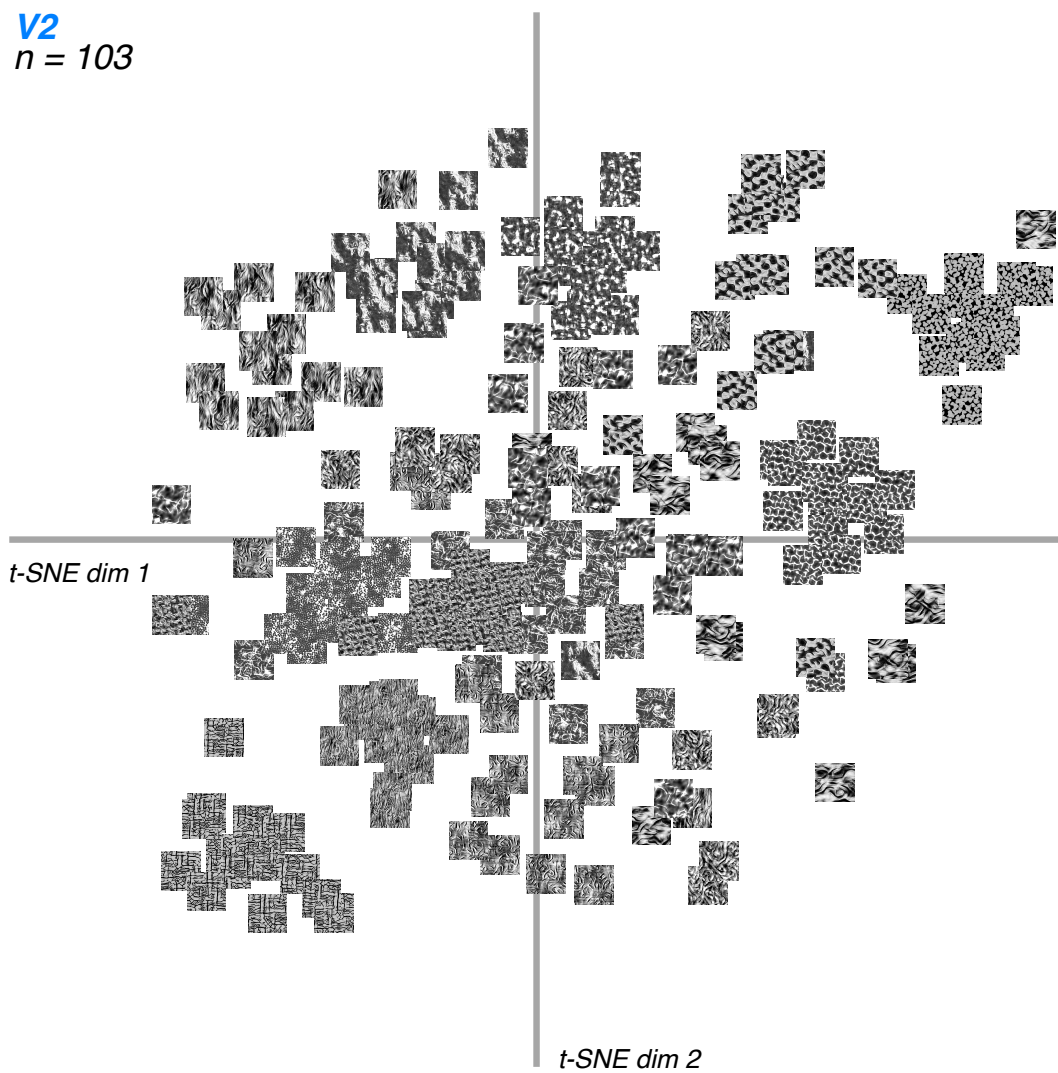


Figure 4.15: *t-SNE map for the V2 population response to 225 images.*

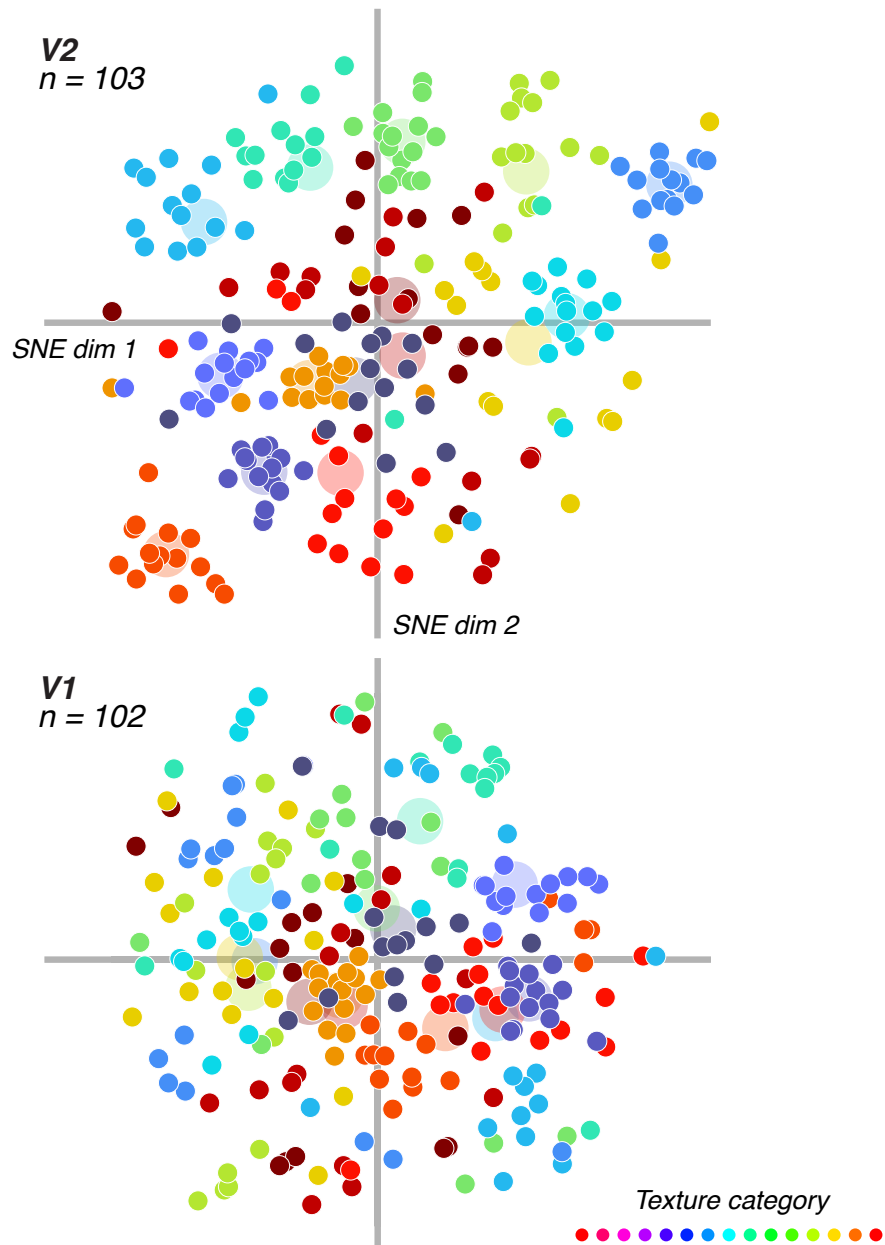


Figure 4.16: Low-dimensional maps in V1 and V2 recovered by *t*-SNE. Each dot is an image; dots of the same color are from the same texture category. Transparent circles show category centroids.

When dots of the same color are closer together, it suggests that the neuronal responses to multiple samples of the same category are more similar. Although there is some evidence of such clustering in V1, the visualizations suggests that clustering substantially increases from V1 to V2.

To quantify this increase in clustering, we directly computed neuronal distances within and across categories in each area (in the full-dimensional space of neuronal response). For each pair of images, we computed within category squared distances,

$$d_{within} = ||\vec{x}_i - \vec{x}_j||^2 \quad (4.17)$$

for all (i, j) such that \vec{x}_i and \vec{x}_j are from the same texture categories. We similarly computed

$$d_{across} = ||\vec{x}_i - \vec{x}_j||^2 \quad (4.18)$$

for all (i, j) such that \vec{x}_i and \vec{x}_j were from different texture categories. We then computed histograms of these distances, separately for each area. If images of the same category yield neuronal responses that are similar, the distribution of within-category distances should be separated from the distribution of across-category distances. We captured the separation between distributions by computing the difference in means divided by the standard deviation, which we refer to as d' .

Figure 4.17 reports these distributions, showing that there was separation – evidence for clustering – in each population, but it was significantly higher in V2 than V1, with a d' 1.7 times larger in V2 ($P < 0.01$, bootstrap test resampling repeated presentations, all neurons included). However, we again wondered whether this increased invariance was due to larger receptive field sizes in V2. For the full population, receptive fields were larger in V2 ($1.93 \pm 1.27^\circ$) than in V1 ($1.16 \pm 0.77^\circ$)

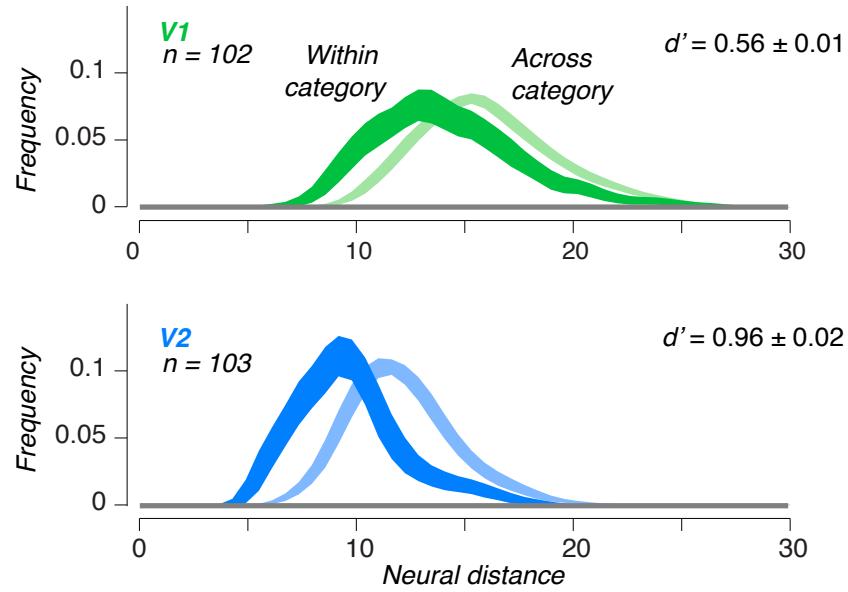


Figure 4.17: Distributions of distances in neuronal response space were computed separately for images from the same category (dark) and different categories (light). Shaded region indicates 2 standard deviations of a bootstrapped distribution obtained by resampling repeated stimulus presentations. For each area we computed d' as the difference between the distribution means divided by the standard deviation.

(mean \pm sd). Following the approach of Rust and DiCarlo [189], we repeated the population analysis on subpopulations with matched receptive fields. First, we confirmed the differences between V2 and V1 when restricting analysis to random subset of 39 neurons (ignoring receptive field size); differences were significant ($P < 0.01$, bootstrap test resampling neurons) and of similar magnitude, with d' 1.8 times larger in V2 than in V1. We then analyzed subsets of neurons from each area with distributions of receptive field sizes matched for mean and variance. d' increased slightly in V1, and there was a pronounced decrease in V2, eliminating the difference between areas ($P > 0.05$). Thus, both single neuron and population analyses show that neurons in V2 support the perceptual similarity of statistically matched images, but that capacity may in part reflect the larger receptive fields in V2.

4.5 Discussion

We have identified two ways in which perceptual capabilities involving naturalistic stimuli depend on neuronal signals in V2, but not V1. These results provide strong evidence that V2 plays a direct functional role in representing properties of these naturalistic stimuli. Alongside studies of disparity tuning [155, 156, 157], these results provide some of the first specific links between perceptual capabilities and V2 neurons.

First, we showed that perceptual performance discriminating naturalistic and spectrally-matched noise stimuli was better for texture categories that modulated V2 responses than for those that did not. We conjecture that these textures contain more or stronger features that drive the elements of a functional mechanism sensitive to local magnitude dependencies like that postulated in Chapter 3, thereby evoking larger differential responses in many V2 neurons, which in turn support the perception of these textures. Our analysis of a large ensemble of textures suggests that, indeed, the degree of cross-scale, cross-position, and cross-orientation magnitude dependencies predicts sensitivity. We also suspect, however, that neurons in V2 do not blindly signal the presence of these dependencies, but rather exhibit selectivity or tuning to particular forms of dependency. The distribution of tuning in V2 may result, at the population level, in the tendency for certain textures to be more effective than others. Future work can directly test this hypotheses by measuring response modulation for textures matched to different subsets of statistics [11], or by fitting nonlinear, hierarchical model to responses of individual V2 neurons [203, 187, 186, 226].

We established the above relationship indirectly, by comparing psychophysical performance (in humans), single unit responses (in anesthetized macaque), and

fMRI responses (in humans performing an attentionally demanding fixation task). To establish an even more direct relationship, we could measure V2 responses to our stimuli in awake macaques performing a suitable version of our psychophysical task; we could then relate neurometric and psychometric functions, and use choice probability, alongside measures of local neuronal correlations, to infer direct neuronal contributions to behavior [94, 153, 27, 26, 198]. Measuring perceptual performance in macaques would also allow us to assess any species differences in the relative efficacy of different texture categories, perhaps arising from different visual experience during development. More generally, these stimuli could prove useful for studying the development of selectivity in area V2; macaques could be exposed, from birth, to particular texture categories, to test the hypothesis that exposure is partly responsible for the efficacy of some texture categories over others. Because differential responses to these stimuli are present throughout V2 but not in V1, coarse, noninvasive measures, like fMRI or even EEG, could be use to track V2-specific signals as a function of development [65, 4].

Second, we found that responses in V2, at both the single-neuron and population level, were capable of supporting the perceptual similarity of physically-distinct but statistically-matched texture samples. We can think of this property as an “invariance”, whereby neuronal responses are tolerant to differences in the precise physical structure of a texture. Together with a presumed increase in selectivity to these texture stimuli, our results suggest how both selectivity and invariance may increase from V1 to V2, mirroring a similar effort by Rust and DiCarlo (2011) to describe representational changes between V4 and IT [189, 76]. That study also found that differences in receptive field sizes were in part responsible for increases from V4 to IT in tolerance to position, size, and context [189]. As those authors note, however, increases in receptive field size are complicated by concomitant increases

in selectivity. They found that IT neurons, along with showing more tolerance to physical transformations, were also more sensitive than V4 neurons to scrambling object features, analogous to our finding that V2 neurons were more sensitive to the difference between naturalistic and noise (recall that our noise images were generated by scrambling phase, a lower-level version of the object scrambling used by Rust and DiCarlo, which *preserved* textural features while removing global object properties). These increases in conjunction selectivity show that the transformations from V1 to V2 and from V4 to IT cannot merely reflect linear magnifications of receptive fields. We should thus interpret the concomitant increase in invariance as an *interaction* between the increased selectivity for textural features and an increase in receptive field size. Put another way, increasing receptive field size without increasing selectivity should not produce clustering into different textural categories; indeed, we found that matching receptive field sizes made V1 and V2 comparable more because it reduced clustering in V2 more than because it increased clustering in V1. This hypothesis could be further explored by simulating the responses of model V1 and V2 neurons to these stimuli and assessing the effect of receptive field sizes on clustering.

The invariance described here may differ from the invariance commonly described in the context of object recognition. Those studies have emphasized tolerance to changes in the size or position of an object [14, 117, 115, 241, 189], rather than tolerance to random structural variability. A series of elegant experiments by DiCarlo and colleagues have provided compelling evidence that neurons in IT achieve invariance to physical transformations because an animal ordinarily experiences a stable world in which objects maintain their identity across changes, for example, in an observer's viewpoint [135, 136, 137, 47]. How might an observer learn invariances for textural statistics? Exposure to physical transformations seems insufficient –

there is no simple transformation that turns one random sample of texture into another. Rather, we may learn textural statistics as we *sample*, through fixations, our visual environment. With each glance at a complex pattern, we acquire a new sample of texture, and if we assume that the images we are seeing arise from a common material, we may learn the set of higher-order statistics associated with that material. We could test this hypothesis by performing, in the domain of texture, experiments analogous to those performed by DiCarlo and colleagues, in which ordinary experience is disrupted to “teach” neurons novel forms of invariance.

The invariances described in this chapter have focused on homogenous patches of naturalistic texture. Perception of complex scenes depends on local computations like those described here, operating at multiple locations across a heterogeneous image. In the next Chapter, we describe the perceptual consequences of such a representation.

Notes

¹²We tried additionally restricting our experiment to the United States, but found that participation was higher and performance more stable when allowing turkers from all countries. Anecdotal evidence suggests that most of our participants were from India.

¹³The procedure described here was inspired by a mixture model used in a clinical setting, to pool multiple, potentially-unreliable first-year clinicians’ ratings of patients’ symptoms [53]. To psychophysicists, this model may seem either unfamiliar, or familiar and troubling in so far as it resembles high-threshold theory. But it is well motivated in this case. Lapse parameters are commonly used when modeling psychophysical data, but lapse rates greater than 5% indicate that subjects should be removed[235]. Our mixture model accepts and appropriately models a wider variety of behaviors: some likely pay attention to every trial, some simultaneously play online poker and perform the task carefully only every other trial, and some ignore the task entirely.

¹⁴We used publicly available code: <http://homepage.tudelft.nl/19j49/t-SNE.html>

Chapter 5

Population representations in V2 predict visual metamers

5.1 Introduction

The concept of invariance has been emphasized throughout the study of the ventral stream – the sequence of cortical areas from V1 to IT thought to be involved in representing and recognizing visual objects. A key fact about functional organization in the ventral stream is that receptive field sizes increase across successive areas. Many models of pattern recognition in the ventral stream [90, 78, 49, 179, 197, 183, 173] have proposed that such increases in spatial pooling provide invariance to geometric transformations (e.g., changes in position or size). Physiological experiments in inferotemporal cortex have identified neural correlates of invariant representations [14, 117, 115, 241, 189], along with mechanisms by which the brain may learn invariances by exploiting spatiotemporal dependencies in the visual input [135, 136, 137, 47].

In the previous two Chapters, we showed that V2 neurons respond uniquely

to a class of naturalistic stimuli, and we showed that their responses reflect both perceptual sensitivity and invariance to the features of the stimuli. In particular, statistically-matched images, physically different but belonging to the same category, yielded similar population-level responses, more so in V2 than in V1. This result implies that V2, along with explicitly encoding features of naturalistic stimuli, also discards information from V1 so as to achieve a more invariant representation of statistically-similar textures.

Compared to invariance to position or size, the invariance described in Chapter 3 is more stochastic in nature; images from the same texture category are related not through a fixed set of physical transformations, but because they have the same higher-order statistics. This invariance may, however, similarly depend on increases in receptive field sizes, specifically from V1 to V2. Furthermore, although the experiments described in Chapters 3 and 4 were restricted to a fairly narrow range of receptive field sizes, we would expect the consequences of this “invariance” to be even more pronounced for larger receptive fields. It is well established that within individual ventral stream areas, receptive field sizes scale linearly with eccentricity, and that this rate of scaling is larger in each successive area along the ventral stream, providing a signature that distinguishes different areas (Figure 5.1) [81, 82, 62].

We hypothesize that the increase in spatial pooling, in successive ventral stream areas, and with eccentricity, induces an irretrievable loss of information at each stage of processing, and this loss of information should have perceptual consequences. Stimuli that differ only in terms of this lost information will yield identical population-level responses. If a human observer is unable to access the discarded information, such stimuli will be perceptually indistinguishable; thus, we refer to them as metamers. Visual metamers were crucial to one of the earliest and most successful endeavors in vision science – the elucidation of human trichromacy. Behavioral

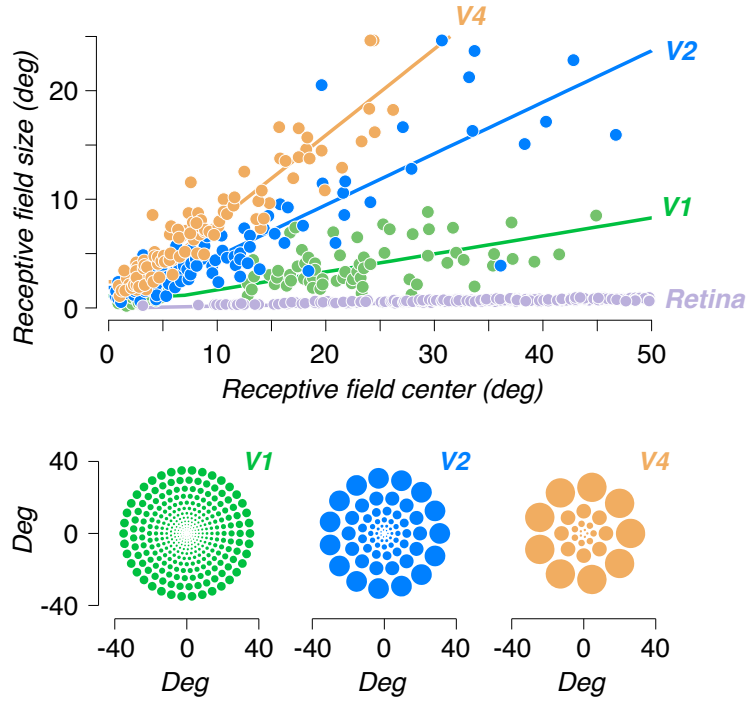


Figure 5.1: *Receptive field sizes increase along the ventral pathway, and as a function of eccentricity. Estimates of receptive field size adapted from [81, 82], and fit with a “hinged” line. “Flower” plots (below) depict receptive fields as circles, with sizes governed by the slope fit for each area.*

experiments predicted the loss of spectral information in cone photoreceptors 100 years before the physiological mechanisms were confirmed [229]. The concept of metamerism is not limited to trichromacy, however, and a number of authors have used it to understand aspects of pattern or texture vision [104, 119, 125, 175].

We developed a population-level functional model for ventral stream computation in and beyond V1 that allowed us to synthesize, and examine the perception of, a novel type of visual metamer. The model was based on the same two-stage computation used in Chapters 2 and 3 to generate experimental stimuli and identify novel differential responses in V2 neurons. But we implemented the computation within localized receptive fields, so as to process and synthesize images of complex

inhomogenous natural scenes. The first stage of the model decomposes an image with a population of oriented V1-like receptive fields. The second stage computes local averages of nonlinear combinations of these responses over regions that scale in size linearly with eccentricity, according to a scaling constant that we vary parametrically. Motivated by the results of Chapters 3 and 4, we refer to the two stages as the “V1 model” and “V2 model”, but it is important to emphasize that the parameters of each model, especially the V2 model, may not correspond to the responses of individual V2 neurons. Rather, we interpret the parameters as collectively reflecting the population-level representation in V2.

Given a photographic image, we synthesized distinct images with identical model responses, and asked whether human observers can discriminate them. From these data we estimated the scaling constant that yielded metameric images, and found that it was consistent with receptive field sizes in area V2. We also used our model to explain the phenomenon of visual crowding [169, 132], in which humans fail to recognize peripherally presented objects surrounded by clutter. Crowding has been hypothesized to arise from compulsory pooling of peripheral information [131, 164, 168, 91], and the development of our model was partly inspired by evidence that crowding is consistent with a representation based on local texture statistics [11]. Our model offers an instantiation of this hypothesis, providing a quantitative explanation for the spacing and eccentricity dependence of crowding effects, generalizing them to arbitrary photographic images, and linking them to the underlying physiology of the ventral stream.

5.2 Model structure and image synthesis

Our “V2 model” was motivated by known facts about cortical computation, human pattern vision, and the functional organization of ventral stream receptive fields. The underlying V1 representation uses a bank of oriented filters covering the visual field, at all orientations and spatial frequencies. Simple cells encode a single phase at each position; complex cells combine pairs of filters with the same preferred position, orientation, and scale, but different phase [2]. The second stage of the model was based directly on the algorithm for analyzing homogenous texture patterns described in Chapters 3 and 4. It achieves selectivity for compound image features by computing products between particular pairs of V1 responses (both simple and complex) and averaging these products over local regions, yielding local correlations. These correlations have been shown to capture key features of naturalistic texture images, and have been used to explain some aspects of texture perception [87, 175, 10]. In Chapter 3, we showed that stimuli containing these correlations differentially drive V2 neurons. And as discussed in Chapter 3, local correlations are compatible with models of cortical computation that propose hierarchical cascades of linear filtering, point non-linearities, and pooling [90, 78, 2, 49, 201, 179, 52, 197] (Figure 3.20).

To complete the model we must specify the pooling regions over which pairwise products of V1 responses are averaged. Receptive field sizes in the ventral stream grow approximately linearly with eccentricity, and the slope of this relationship (i.e. the ratio of receptive field diameter to eccentricity) increases in successive areas. In our model, pooling is performed by weighted averaging, with smoothly overlapping positive-valued weighting functions that grow in size linearly with eccentricity, parameterized with a single scaling constant. Below, we present notation for the

weighting functions, and then provide weighted versions of all the model parameters.

5.2.1 Pooling regions

The weighting functions, generically denoted $w(i, j)$, are smooth and overlapping, and arranged so as to tile the image (i.e., they sum to a constant). These functions are separable with respect to polar angle and log eccentricity, ensuring that they grow linearly in size with eccentricity (Figure 5.2). Weighting in each direction is defined in terms of a generic mother window, with a flat top and squared cosine edges:

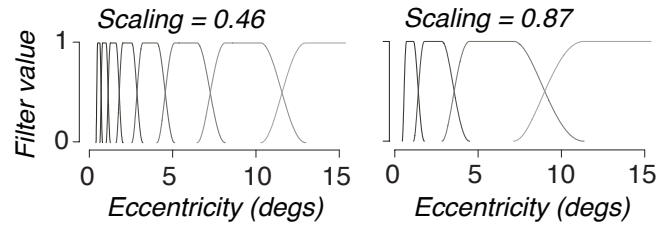
$$f(x) = \begin{cases} \cos^2 \left(\frac{\pi}{2} \left(\frac{x-(t-1)/2}{t} \right) \right), & -(1+t)/2 < x \leq (t-1)/2 \\ 1, & (t-1)/2 < x \leq (1-t)/2 \\ -\cos^2 \left(\frac{\pi}{2} \left(\frac{x-(1+t)/2}{t} \right) \right) + 1, & (1-t)/2 < x \leq (1+t)/2 \end{cases} \quad (5.1)$$

These window functions sum to a constant when spaced on the unit lattice. The parameter t specifies transition region width, and is set to $1/2$ for our experiments. For polar angle, we require an integer number N_θ of windows between 0 and π . The full set is:

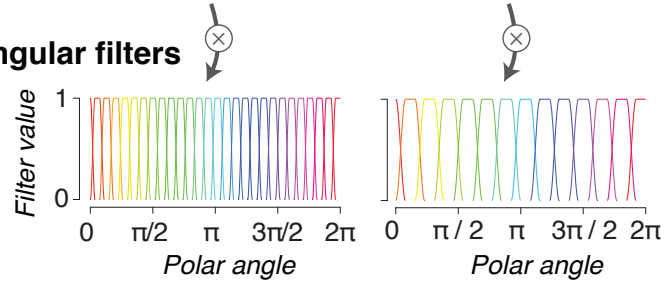
$$h_n(\theta) = f \left(\frac{\theta - \left(w_\theta n + \frac{w_\theta(1-t)}{2} \right)}{w_\theta} \right), w_\theta = \frac{2\pi}{N_\theta}, n = 0 \dots N_\theta - 1 \quad (5.2)$$

where n indexes the windows, w_θ is width. For log eccentricity, an integer number of windows is not required. However, to equate boundary conditions across scaling conditions in our experiments, we center the outermost window on the radius of the image (e_r). And for computational efficiency, we also do not include windows below a minimum eccentricity $-e_0$, approximately half a degree of visual angle in our

Eccentricity filters



Angular filters



2D Filters

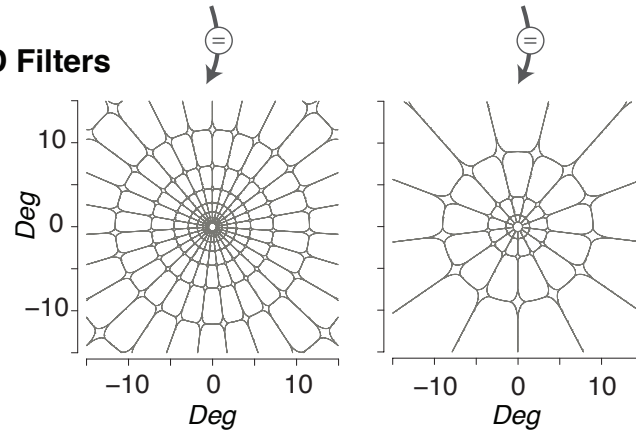


Figure 5.2: The pooling regions in the model are two-dimensional weighting functions (bottom) defined separately in log eccentricity (top) and polar angle (middle). Pooling regions shown for two different scalings (ratio of radial extent to eccentricity).

stimuli. For lower eccentricities, pooling regions are extremely small and constrain the model to reproduce the original image. Between the minimum and maximum eccentricities, we construct N_e windows:

$$g_n(e) = f\left(\frac{\log(e) - [\log(e_0) + w_e(n+1)]}{w_e}\right), w_e = \frac{\log(e_r) - \log(e_0)}{N_e},$$

$$n = 0 \dots N_e - 1 \quad (5.3)$$

n indexes the windows, w_e is the width. The number of windows, N_e determines the ratio of radial full-width at half-maximum to eccentricity, which is reported as the scaling. We can achieve an arbitrary scaling (i.e., a non-integer number of windows) by releasing the constraint on the endpoint location. For each choice of scaling, we choose an integer number of polar-angle windows (N_θ) that yields an aspect ratio of radial width to circumferential width of approximately 2. There are few studies on peripheral receptive field shape in the ventral stream, but our choice was motivated by reports of radially elongated receptive fields and radial biases throughout the visual system [192, 182, 72]. Future work could explore effects of both the scaling and the aspect ratio on metamerism.

To use each window at different scales of the pyramid, we create an original window in the pixel domain, and then blur and downsample the window so that it has the correct spatial extent when applied to the blurred and downsampled versions of the image at each level of the pyramid decomposition (i.e., we construct a “Gaussian pyramid” for the window). The information captured by averages computed with this full set of two dimensional windows is approximately invariant to global rotation or dilation: shifting the origin of the log-polar coordinate system in which they are defined would reparameterize the model without significantly changing the class of metameric stimuli corresponding to a particular original image.

5.2.2 Weighted statistics

Having specified the pooling regions (i.e. weighting functions), we can formulate weighted versions of all the parameters described in Chapter 3. As in 3, we write the n th subband as $x_n(i, j)$ or \vec{x}_n , we denote the simple cell responses (real part) as $s_n(i, j)$ and the complex cell responses (square root of the sum of the squared responses of symmetric and anti-symmetric filters) as denoted $e_n(i, j)$. The statistics are:

(1) Weighted products of responses at nearby locations (i.e. weighted autocorrelations) for simple cells are given by,

$$A_w(n, k, l) = \sum \sqrt{w(i, j)} (s_n(i, j) - \mu_w(\vec{s}_n)) \times \sqrt{w(i + k, j + l)} (s_n(i + k, j + l) - \mu_w(\vec{s}_n)) \quad (5.4)$$

Where (k, l) specifies spatial displacement, the summation is over (i, j) , and $\mu_w(\vec{s}_n)$ is the weighted mean,

$$\mu_w(\vec{s}_n) = \sum w(i, j) s_n(i, j) \quad (5.5)$$

And weighted complex cell autocorrelations are similarly given by,

$$B_w(n, k, l) = \sum \sqrt{w(i, j)} (e_n(i, j) - \mu_w(\vec{e}_n)) \times \sqrt{w(i + k, j + l)} (e_n(i + k, j + l) - \mu_w(\vec{e}_n)) \quad (5.6)$$

For both autocorrelations, we use spatial displacements in the range $(-3 \leq k \leq 3, -3 \leq l \leq 3)$.

(2) Weighted products of complex cell responses with those at other orientations and scales are given by

$$C_w(n, m) = \sum w(i, j) (e_n(i, j) - \mu_w(\vec{e}_n)) (e_m(i, j) - \mu_w(\vec{e}_m)) \quad (5.7)$$

where indices (n, m) specify two subbands arising from filters at different orientations at the same scale, or at different orientations and adjacent scales. We include 6 cross-orientation correlations at each scale, and 16 cross-scale correlations.

(3) Weighted products of simple cell responses with phase-doubled responses at the next coarsest scale are given by,

$$S_w(n, m) = \sum w(i, j) (x_n(i, j) - \mu_w(\vec{x}_n)) \left(\frac{x_m^2(i, j)}{|x_m(i, j)|} - \mu_w \left(\frac{x_m^2(i, j)}{|x_m(i, j)|} \right) \right) \quad (5.8)$$

where indices (n, m) specify two adjacent scales (n is the finer scale).

(4) Weighted marginal statistics of order p are given by,

$$\mu_w^{(p)}(\vec{s}_n) = \sum w(i, j) (s_n(i, j) - \mu_w(\vec{s}_n))^p \quad (5.9)$$

yielding weighted skew and kurtosis as,

$$\gamma_w(\vec{s}_n) = \frac{\mu_w^{(3)}(\vec{s}_n)}{\left(\mu_w^{(2)}(\vec{s}_n) \right)^{3/2}} \quad (5.10)$$

$$\kappa_w(\vec{s}_n) = \frac{\mu_w^{(4)}(\vec{s}_n)}{\left(\mu_w^{(2)}(\vec{s}_n) \right)^2} \quad (5.11)$$

5.2.3 Synthesis

Metameric images were synthesized to match a set of measurements made on an original image. An image of Gaussian white noise was iteratively adjusted until it matched the model responses of the original. Synthesizing from different white noise samples yielded distinct images. This procedure approximates sampling from the maximum entropy distribution over images matched to a set of model responses [175]. We used gradient descent to perform the iterative image adjustments. For each set of responses, we computed gradients, following the derivations in Portilla and Simoncelli (2000) but including the effects of the window functions. Descent steps were taken in the direction of these gradients, starting with the low-frequency subbands (i.e., coarse-to-fine). For autocorrelations, gradients for each pooling region were combined to give a global image gradient on each step. Gradient step sizes for each group of parameters were chosen to stabilize convergence. For the cross correlations, single-step gradient projections were applied to each pooling region iteratively.

We used 50 iterations for all images generated for the experiments. Parameter convergence was verified by measuring one minus the mean squared error normalized by the parameter variance. For samples synthesized from the same original image, this metric was 0.99 ± 0.015 (mean \pm standard deviation) across all images and scalings used in our experiments. As an indication of computational cost, synthesis of a 512×512 pixel image for a scaling of $s = 0.5$ took approximately 6 to 8 hours on a linux workstation with 2.6 GHz dual Opteron 64-bit processor and 32 GB RAM. Smaller scaling values require more windows, and thus more parameters and more time. The entire set of experimental stimuli took approximately one month to generate.

5.2.4 Experimental stimuli

Stimuli were derived from four naturalistic photographs, three from the authors' personal collection, and one courtesy of Rob Miner. One image depicts a natural scene (trees and shrubbery), and the other three depict people and man-made objects. For each photograph, we synthesized three images for each of six values of the scaling parameter s . Pilot data showed that performance was at chance for the smallest value tested, so we did not generate stimuli at smaller scalings.

5.3 Perceptual determination of critical scaling

5.3.1 Psychophysical methods

Task and subjects

Four observers (ages 24-32, three male, one female) with normal or corrected-to-normal vision participated. Protocols for selection of observers and experimental procedures were approved by the human subjects committee of New York University and all subjects signed an approved consent form. One observer was the author of this thesis; all others were naive to the purposes of the experiment.

Two observers (S3 and S4) were tested with eye tracking (see below), with stimuli presented on a 22 flat screen CRT monitor at a distance of 57 cm. Two observers (S1 and S2) were tested without eye tracking, with stimuli presented on a 13 flat screen LCD monitor at a distance of 38 cm. In both displays, all images were presented in a circular window subtending 26° of visual angle and blended into the background with a 0.75° wide raised cosine. A 0.25° fixation square was shown throughout the experiment.

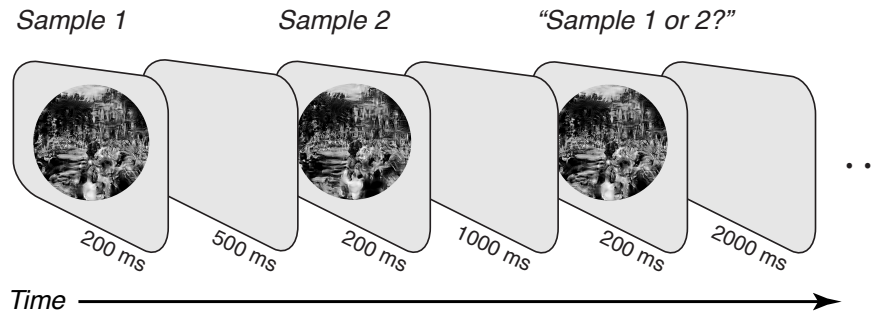


Figure 5.3: In the ABX task, two distinct image samples were presented across two intervals; after a pause, one of the two was repeated, and the observer indicated which one.

Each trial of the “ABX” task (Figure 5.3) presented two different synthesized image samples, matched to the model responses of a corresponding original image. At the start of each trial, the observer saw one image for 200 ms. After a 500 ms delay, the observer saw the second image for 200 ms. After a 1000 ms delay, the observer saw one of the two images, repeated, for 200 ms. The observer indicated with a key press whether the third image looked more like the first (“1”) or the second (“2”). There was no feedback during the experiment. Before the experiment, each observer performed a small number of practice trials (≈ 5) with feedback to become familiar with the task.

In the “V2 model” experiment, we used four original images and six scaling conditions, and created three synthetic images for each original / scaling combination. This yielded 12 unique ABX sequences per condition.¹⁵ In each block of the experiment, observers performed 288 trials, one for each combination of image (4), scaling (6), and trial type (12). Observers performed four blocks (1152 trials). Blocks were performed on different days, so the observer never saw the same stimulus sequence twice in the same session. Psychometric functions and parameter estimates were similar across blocks, suggesting that observers did not learn particular features of

any individual images. Results were also similar across the four original images, and were thus combined.

Eye tracking

Two observers (S3 and S4) were tested while their gaze positions were measured (500 Hz, monocular) with an Eyelink 1000 (SR Research) eye tracker, for all four metamer experiments. A 9-point calibration was performed at the start of each block. We analyzed the eye position data to discard trials with broken fixation. We first computed a fixation location for each block by averaging eye positions over all trials. This was used as fixation, rather than the physical screen center, to account for systematic offset due to calibration error. We then computed, on each trial, the distance of each gaze position from fixation; a trial was discarded if any gaze position exceeded 2° . We discarded 5% (S3) and 17% (S4) of trials across all experiments. Using a more conservative (1°) threshold discarded more trials, but did not substantially change psychometric functions or critical scaling estimates. By only including trials with stable fixation, we ruled out the possibility that systematic differences in fixation among scaling conditions, presentation conditions, or models, could account for our results.

5.3.2 Generation of metameric stimuli

If our model accurately describes the information captured (and discarded) in V2, and human observers cannot access the discarded information, then any two images that produce matching model responses should appear identical. To directly test this assertion, we examined perceptual discriminability of synthetic images that were as random as possible while producing identical model responses [175]. First, model

responses were computed for a full-field photograph. Then synthetic images were generated by starting from Gaussian white noise and iteratively adjusting them until they matched the model responses of the original.

Figure 5.4 shows two such synthetic images, generated with a scaling constant (derived from the experiments described below) that yields nearly indiscriminable samples. The synthetic images are identical to the original near the intended fixation point (red circle), where pooling regions are small, but features in the periphery are scrambled, and objects are grossly distorted and generally unrecognizable. When viewed with proper fixation, however, the two images appear nearly identical to the original and to each other.

5.3.3 Perceptual determination of critical scaling

To test the model more formally, and to establish a specific link to area V2, we measured the perceptual discriminability of synthetic images as a function of the scaling constant used in their generation. If the model, with a particular choice of scaling constant, captures the information represented in any visual area, then model-generated stimuli will appear metameric. If the scaling constant is made larger, the model will discard more information than the associated visual area, and model-generated images will be readily distinguishable. If the model scaling is made smaller, the model discards less information, and the images will remain metameric. Thus, we seek the largest value of the scaling constant such that stimuli appear metameric. This critical scaling should correspond to the scaling of receptive field sizes in the area where the information is lost.

As a separate control for the validity of this paradigm, we also examined stimuli generated from a “V1-only model” that only computes pooled V1 complex cell

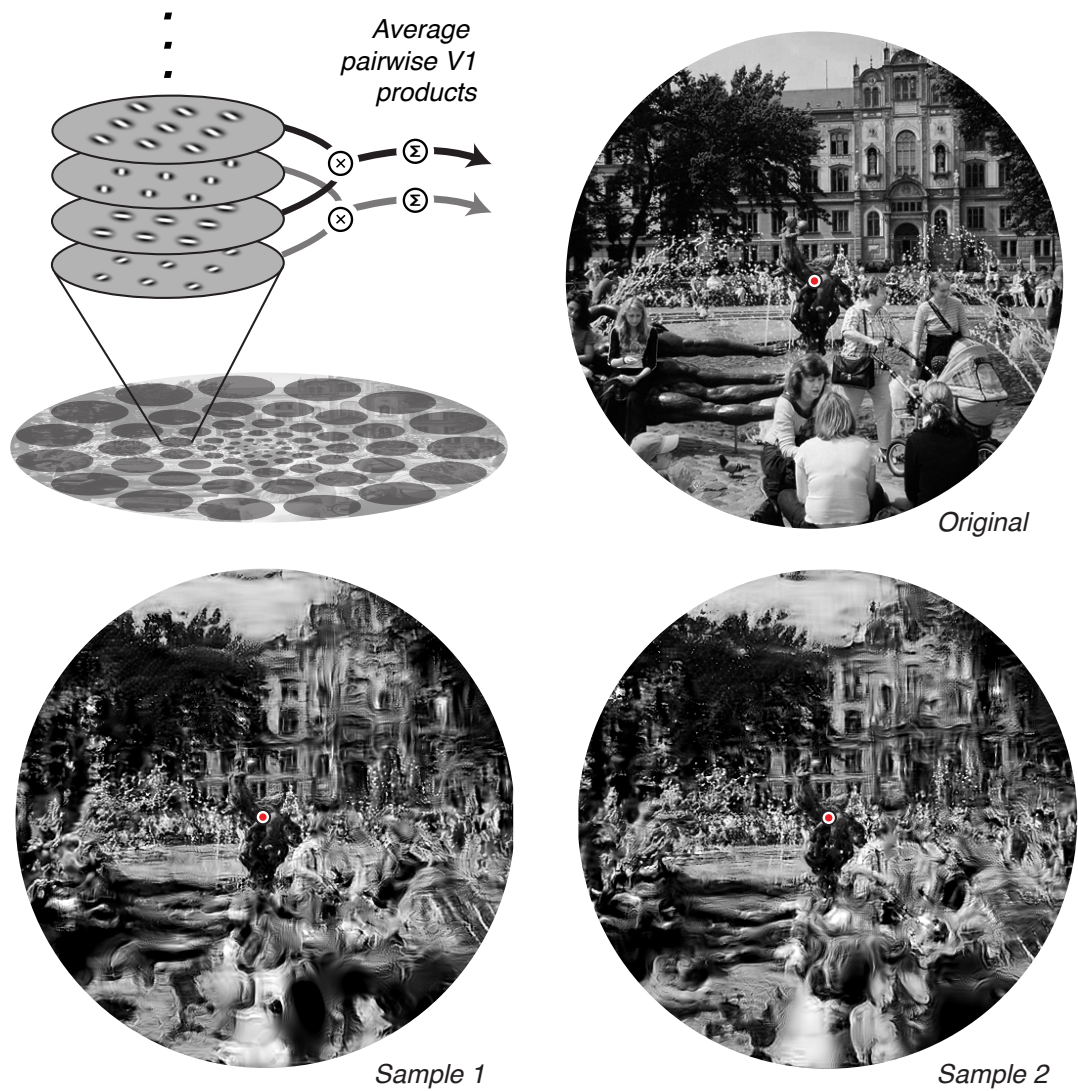


Figure 5.4: V2 model responses (upper left) were computed on an original image (upper right), and new samples (bottom) were generated by matching an image of Gaussian noise for the responses computed on the original.

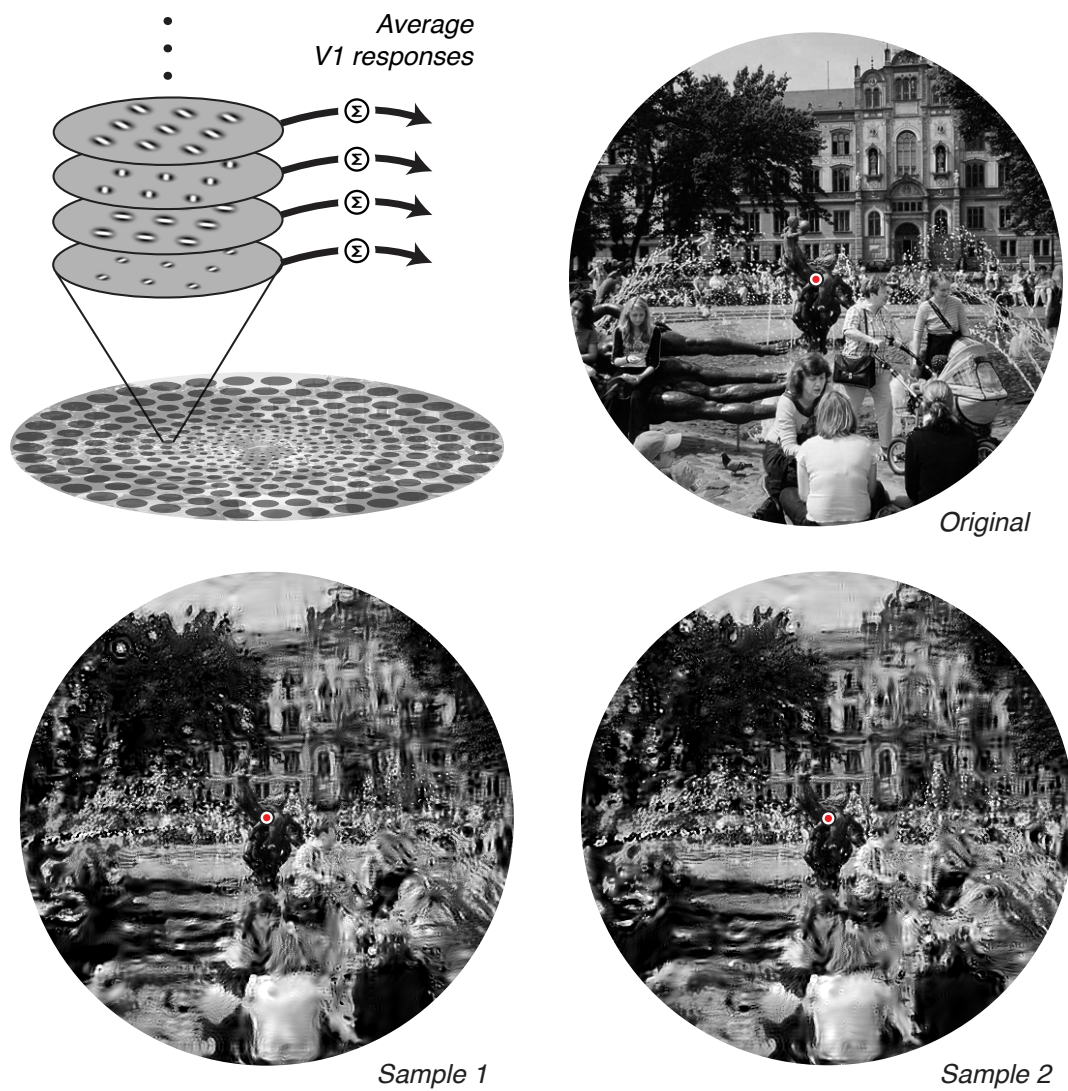


Figure 5.5: *Metamers for the “V1 model” were generated by only computing (and matching) spatial averages of the V1 responses. Images were again generated by matching a noise image to the model responses computed on an original image.*

responses [41] (i.e., local spectral energy). This model used the same components as the “V2 model”, but did not include the local correlations. Like the V2 model, the V1-only model collapsed the computation into a single stage of pooling, instead of building the V2 model on the responses of a pooled V1 stage (and previous stages, such as the retina and LGN). This kind of simplification is common in modeling sensory representations, and allowed us to develop a tractable synthesis procedure. The “V1-only model” experiment was identical, except that it included 9 scaling conditions, resulting in 384 trials per block. Observers performed three blocks (1152 trials). The critical scaling estimated for these stimuli should match the receptive field sizes of area V1. Since the V2 model includes a larger and more complex set of responses than the V1 model, we know a priori that the critical scaling for the V2 model will be as large or larger than for the V1 model, but we do not know by how much.

For each model, we measured the ability of human observers to distinguish synthetic images generated for a range of scaling constants. All four observers exhibited monotonically increasing performance as a function of scaling constant (Figure 5.6). Chance performance (50%) indicates that the stimuli are metameric, and roughly speaking, the critical scaling is the value at which each curve first rises above chance.

To obtain an objective estimate of the critical scaling values, we derived an observer model that used the same representation used to generate the matched images. Our model assumes that an observer’s performance in the ABX experiment is determined by a population of model neurons whose receptive fields grow with eccentricity according to scaling parameter s_0 , and their performance depends on the total squared difference of those responses computed on the two presented images generated with model critical scaling s . Here we derive a closed-form approximation

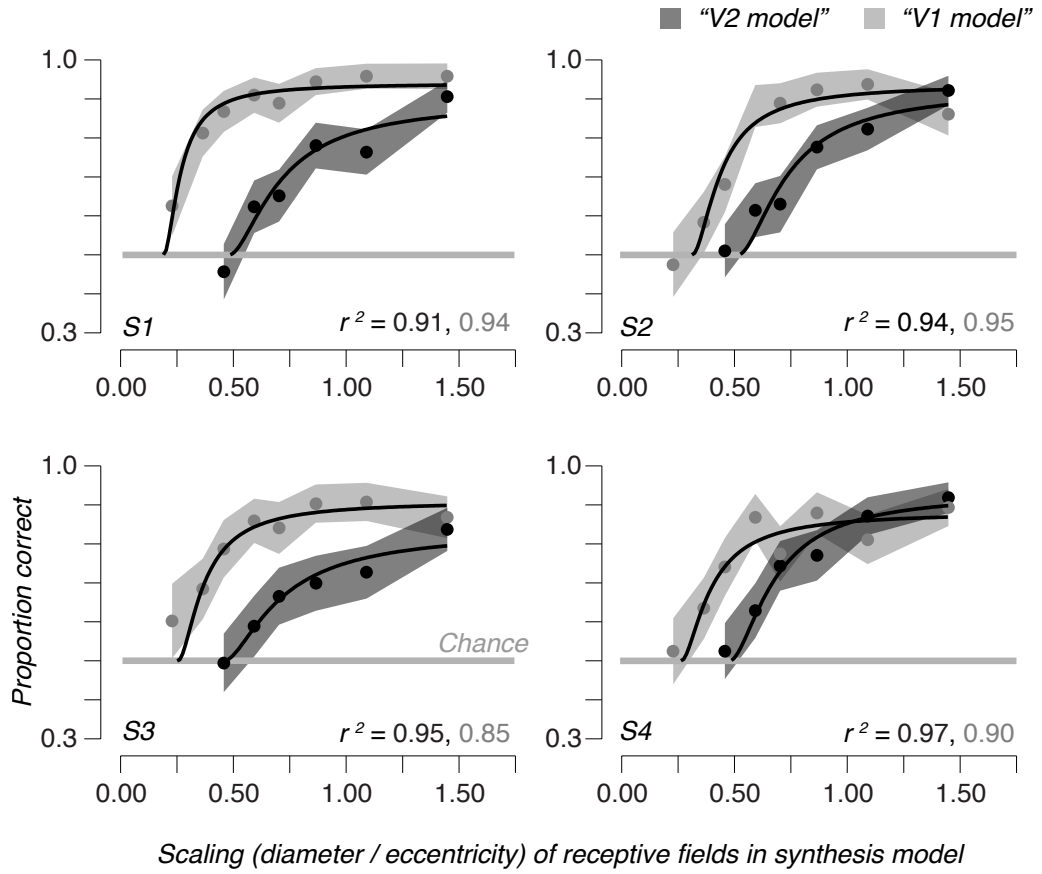


Figure 5.6: Psychometric functions for four observers showing performance as a function of the scaling of receptive fields in the model used to generate stimuli, for the V1 and V2 models. Shaded regions, 68% bootstrapped confidence interval. Black lines, performance of observer model with critical scaling and gain parameters chosen to maximize the likelihood of the data.

to that squared difference as a function of s_0 and s . Let \vec{x} be a vector of values from an original image to be locally averaged (e.g., a vector containing pairwise products of two orientation subbands). Let M be a matrix whose rows contain the weighting functions (with sizes scaling according to s), that are used to compute local averages. Assume that a second vector \vec{y} was initially set to a vector of white noise samples, \vec{n} , and then adjusted so that $M\vec{x} = M\vec{y}$, i.e., the two images match

with respect to the local averages computed by M . Define the projection matrix $P = M^T (MM^T)^{-1} M$, which projects vectors into the space spanned by M . We can rewrite \vec{y} as the sum of two components,

$$\vec{y} = (I - P)\vec{n} + P\vec{x} \quad (5.12)$$

where I is the identity matrix, the first term is the component of \vec{n} that lies in the null space of M , and the second is constrained by the fact that \vec{y} is matched to \vec{x} (i.e., $M\vec{x} = M\vec{y}$).

Now let R be the matrix that the observer uses to compute averages over regions scaling with s_0 . We assume the discriminability of the two stimuli depends on the sum of squared differences between these averages. We can express the expected value of this quantity, taken over instantiations of \vec{x} and \vec{y} that match the same model measurements, as:

$$\begin{aligned} d^2 &= \mathbf{E} [||R\vec{x} - R\vec{y}||^2] \\ &= \mathbf{E} [||R((I - P)\vec{x} + P\vec{x}) - ((I - P)\vec{n} + P\vec{x})||^2] \\ &= \mathbf{E} [||R(I - P)(\vec{x} - \vec{n})||^2] \end{aligned} \quad (5.13)$$

where we use the definition of \vec{y} from Eq 5.12 and rewrite \vec{x} in a similar form. Assuming that \vec{x} and \vec{n} are independent and have the same covariance matrix C , we obtain:

$$\begin{aligned} d^2 &= \text{Tr} \left(\mathbf{E} \left[R(I - P)(\vec{x} - \vec{n})(\vec{x} - \vec{n})^T (I - P^T)R^T \right] \right) \\ &= \text{Tr} \left(R(I - P)2C(I - P^T)R^T \right) \\ &= \text{Tr} \left((R - RM^T(MM^T)^{-1}M)2C(R^T - M^T(MM^T)^{-1}MR^T) \right) \end{aligned} \quad (5.14)$$

We can obtain a simple functional form for this expression by assuming that C is

a multiple of the identity matrix. In general, the components of \vec{x} (and \vec{n}) are not decorrelated, but the predicted discriminability is still valid within a scale factor, as can be verified through simulation. After some matrix algebra, we obtain

$$d^2 \propto \text{Tr}(RR^\top) - \text{Tr}\left(R^\top RM^\top(MM^\top)^{-1}M\right) \quad (5.15)$$

This provides a closed-form expression for the overall error as a function of the measurement matrices M and R . Finally, we wish to express this result in terms of the scaling parameters for the synthesis model and the observer. This is easily obtained from Eq 5.15 if we assume that (i) M and R compute local means within blocks of fixed sizes m and r , respectively, (ii) m is an integer multiple of r (iii) both m and r divide evenly into n , the length of \vec{x} . For matrices with this structure, we can express d^2 as a function of m :

$$d^2(m) \propto \begin{cases} \frac{n}{r^2} \left(1 - \frac{r}{m}\right) & m > r \\ 0 & m \leq r \end{cases} \quad (5.16)$$

This expression has a natural continuous generalization to handle smoothly overlapping averages and non-integer ratios. The radial extent of our model pooling regions is proportional to the scaling s , so the average region size will be proportional to s^2 , with a proportionality constant that depends on the shape of the region. Replacing m with s^2 , and r with s_0^2 , and absorbing the factor of n/r^2 into a single scale constant, gives the closed form approximation:

$$d^2(s) \approx \begin{cases} \alpha_0(1 - s_0^2/s^2) & s > s_0 \\ 0 & s \leq s_0 \end{cases} \quad (5.17)$$

We empirically verified that this approximation holds for the smooth weighting functions used in our model implementation. The proportionality factor, α_0 , is likely to differ for each measurement in the model. If we assume that the observer performs a weighted sum of the squared errors over the full set of measurements, then the overall error will be of the same form as that of Eq 5.17. Notice that α_0 scales the magnitude of the squared difference, without affecting the point at which the curve first exceeds 0 (i.e., $s = s_0$). Thus, when fitting the data, the gain parameter captures variability in overall performance across observers and presentation conditions. Finally, signal detection theory [139] describes the probability of a correct response $P_C(s)$ in the ABX task as a function of the underlying difference $d^2(s)$,

$$P_c(s) = \Phi\left(d^2(s)/\sqrt{2}\right) \Phi(d^2(s)/2) + \Phi\left(-d^2(s)/\sqrt{2}\right) \Phi(-d^2(s)/2) \quad (5.18)$$

where Φ is the CDF of the Normal distribution. We used the MATLAB `fminsearch` routine to find the values of the gain factor (α_0) and the critical scaling (s_0) that maximized the likelihood of the data (proportion correct responses for each scaling) under this model, for each subject and condition. We used bootstrapping to obtain 95% confidence intervals for the parameter estimates: we resampled the individual trials with replacement, and refit the resampled data to reestimate the parameters.

The observer model provided an excellent fit to individual observer data for both the V1 and V2 experiments (Figure 5.6). Critical scaling values (s_0) were highly consistent across observers, with most of the between-subject variability captured by differences in overall performance (α_0). As expected, the simpler V1-only model required a smaller scaling to generate metameric images. Specifically, critical scaling values for the V1 model were 0.26 ± 0.05 (mean \pm sd), whereas values for the mid-ventral model were roughly twice as large (0.48 ± 0.02).

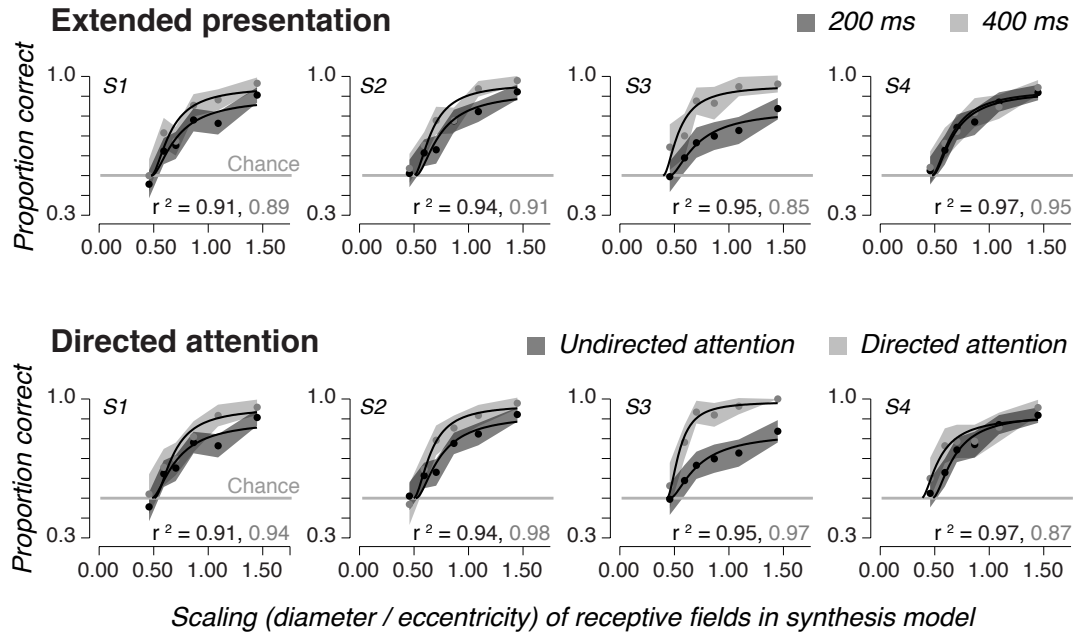


Figure 5.7: Critical scaling was measured under a bottom-up manipulation (extended presentation time) and a top-down manipulation (directed attention) of performance. Psychometric functions for the 200 ms, undirected attention condition are replotted from Figure 5.6, with the same conventions.

5.3.4 Robustness to manipulations of performance

If metamerism reflects a structural limitation of the visual system, governed by the eccentricity-dependent scaling of receptive field sizes, the effects should be robust to experimental manipulations that alter observer performance without changing the spatial properties of the stimuli. To test this, we performed two variants of the experiment, designed to alter performance through bottom-up and top-down manipulations of the experimental task.

First, we repeated the original experiment with doubled presentation times (400 ms instead of 200 ms). Each observer performed either two or three blocks (576 or 864 trials). Fitting the observer model to data from four observers (Figure 5.7),

we found that the gain parameter (α_0) was generally larger to account for increases in performance, but that the critical scaling (s_0) was statistically indistinguishable from that estimated in the original experiment ($P = 0.18$, paired t -test).

In a second control experiment, we manipulated endogenous attention. At the onset of each trial, a small arrow (1° long) was presented at fixation, pointing toward the region in which the two subsequently presented stimuli differed most. It was presented for 300 ms, with a 300 ms blank period before and after. On each trial, we computed the squared error (in the pixel domain) between the two to-be-presented images, and averaged the squared error within each of six radial sections. The line cue pointed to the section with largest squared error. Each observer performed two blocks (576 trials). The fitted gain parameter was again generally larger, accounting for improvements in performance, but the critical scaling was statistically indistinguishable from that estimated in the original experiment (Figure 5.7, $P = 0.30$; paired t -test). In both control experiments, the increase in gain varies across observers, and depends on their overall performance in the original experiment (some observers already have near-maximal performance).

5.4 Estimation of physiological locus

We compared the psychophysically estimated scaling parameters to physiological estimates of receptive field size scaling in different cortical areas. Functional magnetic resonance imaging has been used to measure “population receptive fields” in humans by estimating the spatial extent of a stimulus that contributes to the hemodynamic response across different regions of the visual field [62]. Although these sizes grow with eccentricity, and across successive visual areas, they include additional factors such as variability in receptive field position and non-neural hemodynamic effects,

which may depend on both eccentricity and visual area. We thus chose to compare our results to single-unit electrophysiological measurements in non-human primates.

5.4.1 Physiological measures of receptive field scaling

We performed a meta-analysis to estimate the relationship between physiologically measured receptive field size and eccentricity in non-human primates. Measurements of receptive field sizes are variable across different experiments because different labs use different stimuli and mapping procedures [199, 227, 40]. To compare our psychophysics to physiology, we considered a wide range of data sets: four in V2 [81, 82, 30, 5], five in V1 [81, 83, 40, 220, 6], and three in V4 [82, 140, 57]. Two of these data sets were from owl monkey [5, 6], one from capuchin [83], and the rest were from macaque.

For each visual area, we combined data across experiments and estimated variability by pooling the raw data (rather than the fits), matching sample sizes, and resampling multiple times to obtain a 95% confidence interval on the slopes (Figure 5.8). Specifically, we determined the minimum number of cells across the data sets, and on each iteration of a bootstrap, resampled that number with replacement from each data set, and reestimated the slope of size versus eccentricity from the pooled data. We fit the data with a two-parameter hinged line, with a constant minimum size over some small range of eccentricities, followed by a linear relationship with some slope (examples in Figure 5.1). For consistency, we used this “hinged line” model to estimate all slopes, but we obtained similar results when using a linear fit through 0. We also considered a straight line with variable intercept and slope [62], but the hinged line fit the data well (error was comparable for the two fits) and better matched the parameterization of our model. Variability across data sets

tended to be largest at far eccentricities, and given that our visual stimuli only extended to 12.25 deg, we restricted our analysis of the physiology data to this range. In some of the cited studies [81, 83, 82, 220, 30, 57], rectangular receptive field sizes were mapped using a minimum response field procedure. To convert these numbers to diameters of circular receptive fields, and partially compensate for the bias toward smaller values inherent in this mapping technique [227, 40], we took the average of the diameter associated with the corners and sides of the squares (i.e., we multiplied the reported diameters by $(1 + \sqrt{2})/2$). Small modifications to any of these aspects of the data analysis did not qualitatively change the comparison between our psychophysics and the physiology (Figure 5.8).

5.4.2 Comparison to psychophysics

The meta-analysis of physiological data yielded scaling values of 0.21 ± 0.07 for receptive fields in V1, 0.46 ± 0.05 for those of V2, and 0.84 ± 0.06 for those of V4 (mean with 95% confidence intervals). Moreover, for studies that used comparable methods to estimate receptive fields in both V2 and V1, the average receptive field sizes in V2 were approximately twice the size of those in V1, for both macaque and human [81, 62, 199].

The full set of psychophysically-estimated critical scalings, across all of our observers and experiments, are summarized in Figure 5.8, along with these physiological estimates of receptive field scaling. As expected, the critical scaling value estimated from the “V1-model” experiment were well matched to the physiological estimates of receptive field scaling for V1 neurons. For the “V2-model” experiment, the critical scaling was roughly twice that of the V1 model, was well matched to receptive field sizes of V2 neurons, and was substantially smaller than those

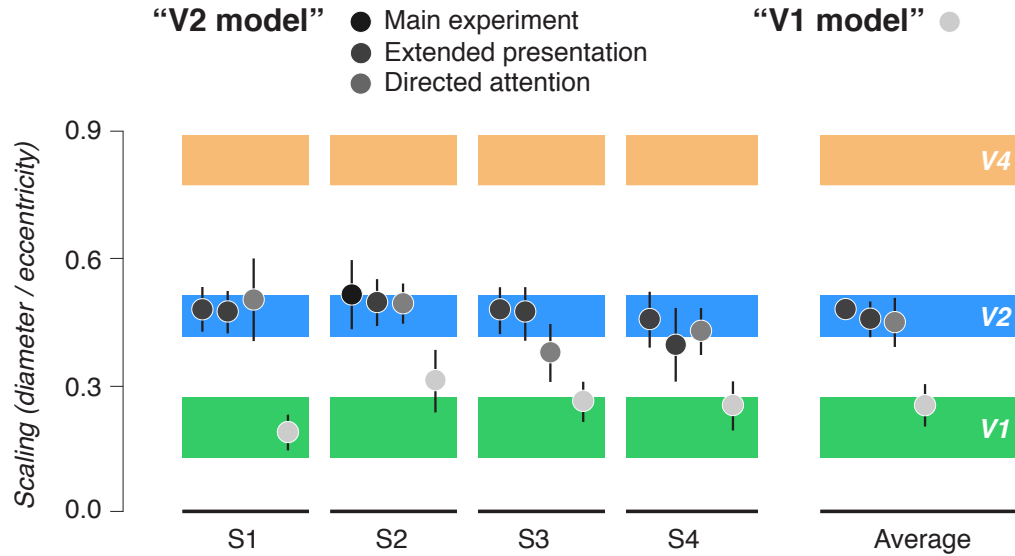


Figure 5.8: Summary of fitted critical scaling parameters for all experiments. Error bars, 95% confidence intervals on parameter estimates obtained through bootstrapping. Colored horizontal bars, physiologically measured receptive field scaling based on a meta-analysis of ten datasets. Bar thickness, 95% bootstrapped confidence interval.

of V4. The scaling for the two control experiments were similar to those of the original experiment, were closely matched to the scaling of receptive fields found in area V2, and were much greater than the scaling found in the V1 metamer experiment ($P = 0.0064$, extended presentation task, $P = 0.0183$, attention task; paired t -test). We take this as compelling evidence that the metamerism of images synthesized using our model arises in area V2.

5.5 Relationship to visual crowding

Our model implies severe perceptual deficits in peripheral vision, some of which are revealed in the well-studied phenomenon known as visual crowding [169, 132]. Crowding has been hypothesized to arise from pooling or statistical combination in

the periphery [131, 164, 168, 91, 11], and thus emerges naturally from our model. Crowding is typically characterized by asking observers to recognize a peripheral target object flanked by two distractors at varying target-to-flanker spacings. To qualitatively link our model to visual crowding, we generated metameric synthetic images based on stimuli commonly used in crowding experiments. Figure 5.9 shows many such demonstrations (adapted from [169]). The column on the left depicts a series of crowding demonstrations. In the middle column, we synthesized metamers using a fixation location close to the objects. In the right most column, we synthesized metamers using a fixation far from the objects. Only for the far fixation is there apparent jumbling or scrambling of the objects, suggesting the appearance of crowded objects. This link is only qualitative, but suggests that the information lost by the model, and the distortions that result, are consistent with the phenomenological experience of crowding [12].

To establish a quantitative link, we exploited the well-characterized dependence of crowding of spacing and eccentricity. Our model parameterizes the scale of eccentricity-dependent pooling. Crowding experiments have similarly estimated the scale of crowding effects, by measuring the “critical spacing” between target and flankers at which performance reaches threshold. That critical spacing is largely independent of stimulus size and increases proportional to eccentricity [132, 169], with reported rates ranging from 0.3 to 0.6. Our estimates of critical scaling for the V2 model metamers lie within this range, but the substantial variability (which arises from different choices of stimuli, task, number of targets and flankers, and threshold) renders this comparison equivocal. Moreover, a direct comparison may not even be warranted, because it implicitly relies on an unknown relationship between the pooling of model responses and the degradation of recognition performance.

So, we performed an additional experiment to determine directly whether our V2

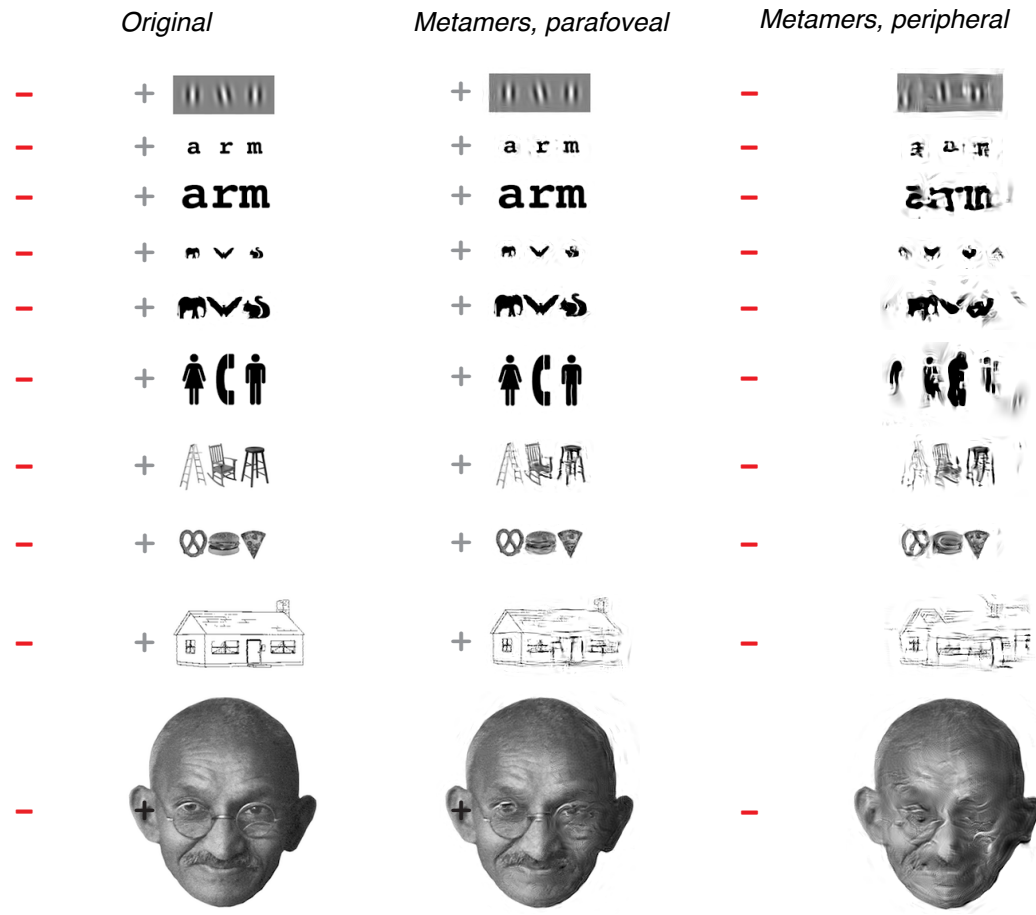


Figure 5.9: Crowding was simulated by applying the model to a set of crowding demos (adapted from [169]). In the column on the left, for each triplet of objects (each row), it should be easy to recognize the middle object when fixating near it (the plus) but not when fixating far away (the minus). Metamers were generated using the same two fixation locations (middle and right columns); only for the far fixation are the objects scrambled and jumbled.

model could predict recognition performance in a crowding task. The experimental design was inspired by a previous study linking statistical pooling in the periphery to crowding [11]. Briefly, the link relies on comparing recognition in two experiments: one that measures performance identifying a crowded, peripheral target object among distractors, and another that measures performance identifying the same object in a metamer generated from the peripheral display.

5.5.1 Crowding methods

Five observers participated in the crowding experiments (one of whom also participated in the experiments described above) (ages 24-32, three male, two female). Stimuli were presented on a 13" flat screen LCD monitor at a distance of 38 cm. Each observer performed two tasks: a peripheral recognition task on triplets of letters, and a foveal recognition task on synthesized stimuli, similar to the experiments in described in [11]. In the first task, each trial began with a 200 ms presentation of three letters in the periphery, arranged along the horizontal meridian. Letters were uppercase, in the Courier font, and 1° in height. The target letter was centered at 6° eccentricity, and the two flanker letters were presented left and right of the target. All three letters were drawn randomly from the alphabet without replacement. We varied the center-to-center spacing between the letters, from 1.1° to 2.8° (all large enough to avoid letter overlap). Observers had 2 s to identify the target letter with a key press (1 out of 26 possibilities, chance = 4%). Observers performed 48 trials for each spacing. For each observer, performance was fit with a cumulative Weibull function by maximizing likelihood. Spacings of 1.1, 1.5, and 2° corresponded to approximately 50%, 65%, and 80% performance, respectively; these spacings were used to generate synthetic stimuli for the foveal task (see below). To extend our

range of performance, two observers were run in an additional condition (8° eccentricity, 0.8° , 1° spacing) yielding approximately 20% performance. For these observers, the same condition was included in the foveal task.

We used our V2 model to synthesize stimuli from letter triplets (comparable to the letter rows from Figure 5.9). To reduce the number of syntheses, we synthesized stimuli containing triplets along eight radial arms, but eccentricity, letter size, font, and letter-to-letter spacing were otherwise identical. For each image of triplets we generated nine different synthetic stimuli: three different spacings (1.1 , 1.5 , 2°) for each of three different model scalings (0.4 , 0.5 , 0.6) centered roughly around the average critical scaling estimated in our metamer experiment. We synthesized stimuli for 56 unique letter triplets; letter identity was balanced across experimental manipulations. On each trial of the foveal recognition task, one of the triplets from the synthesized stimuli was presented for 200 ms, and the observer had 2 s to identify the middle letter. The observer saw each unique combination of triplet identity, spacing, and scaling only once. Trials with different spacings were interleaved, but the three different model scalings were performed in separate blocks (with random order).

5.5.2 Crowding results

First, we measured observers' ability to recognize target letters presented peripherally (6°) between two flanking letters, varying the target-to-flanker spacing to obtain a psychometric function (Figure 5.10, left). We then generated metamers and measured the ability of observers to recognize the letters in the metamers under foveal viewing. Recognition failure (or success) for a single metamer cannot alone

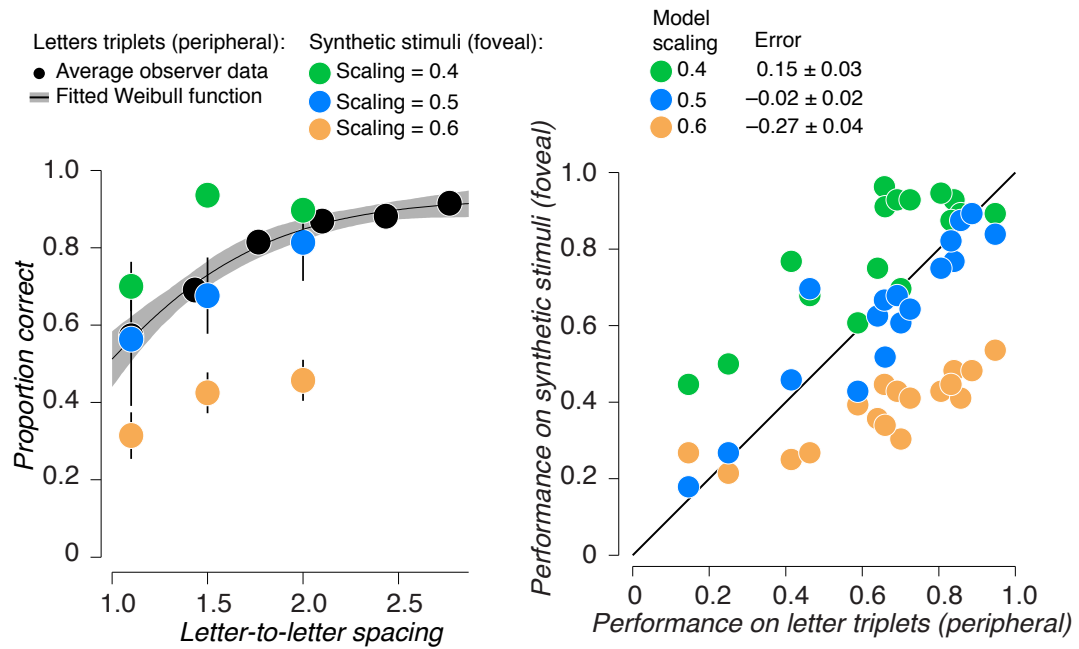


Figure 5.10: (Left) Performance recognizing the central letter in a triplet varied as a function of letter-to-letter spacing (black dots). Black line, best fitting Weibull function; gray shaded region, 95% confidence interval obtained through bootstrapping. Metamers were synthesized for each stimulus, with three different scalings, and observers recognized the letter in the synthetic image (colored dots). (Right) Performance on the two tasks was only similar for a scaling of 0.5, closely matched to V2.

indicate crowding (or lack thereof), but average performance across an ensemble of metamers quantifies the limitations on recognizability imposed by the model.

Average recognition performance for the metamers was well matched to that of their corresponding letter stimuli (Figure 5.10, left), for metamers synthesized with scaling parameter $s = 0.5$ (the average critical scaling estimated for our human observers). For metamers synthesized with scaling parameters of $s = 0.4$ or $s = 0.6$, performance was significantly higher or lower, respectively ($P < 0.0001$; paired t -test). These results were consistent across all observers, at all spacings, and for two different eccentricities, as summarized in Figure 5.10.

5.5.3 Consequences for other visual tasks

The fact that the model operates on arbitrary photographic images allows generalization of the laboratory phenomenon of crowding to complex scenes and everyday visual tasks. For example, crowding places limits on reading speed, because only a small number of letters around each fixation point are recognizable [167]. Model-synthesized metamers can be used to examine this “uncrowded” window (Figure 5.11). We envision that the model could be used to optimize fonts, letter spacings, or line spacings for robustness to crowding effects, potentially improving reading performance. There is also some controversial evidence linking dyslexia to crowding with larger-than-normal critical spacing [84, 169, 144, 242], and our model and paradigm might serve as a useful tool for investigating this hypothesis; Figure 5.11, for example, shows a metamer of text generated with a larger-than-normal critical scaling, which reveals a much smaller window around fixation within which letters are recognizable. Our experimental paradigm could also be used to measure and compare critical scaling in normal and dyslexic subjects, using natural photographic stimuli rather than letters, which could help isolate differences in basic perceptual processing. Corresponding fMRI experiments could measure receptive field scaling in early visual areas [62], and possibly identify physiological correlates of variability in critical scaling across subjects.

Figure 5.12 provides additional metamers, depicting how camouflaged objects, which are already difficult to recognize foveally, blend into the background when viewed peripherally. These images suggest how peripheral representations may impede performance when searching for an object amongst a background of distractors. Along similar lines, Rosenholtz et al. (2012) showed that texture statistics can capture a variety of well-known effects and asymmetries involving visual search.

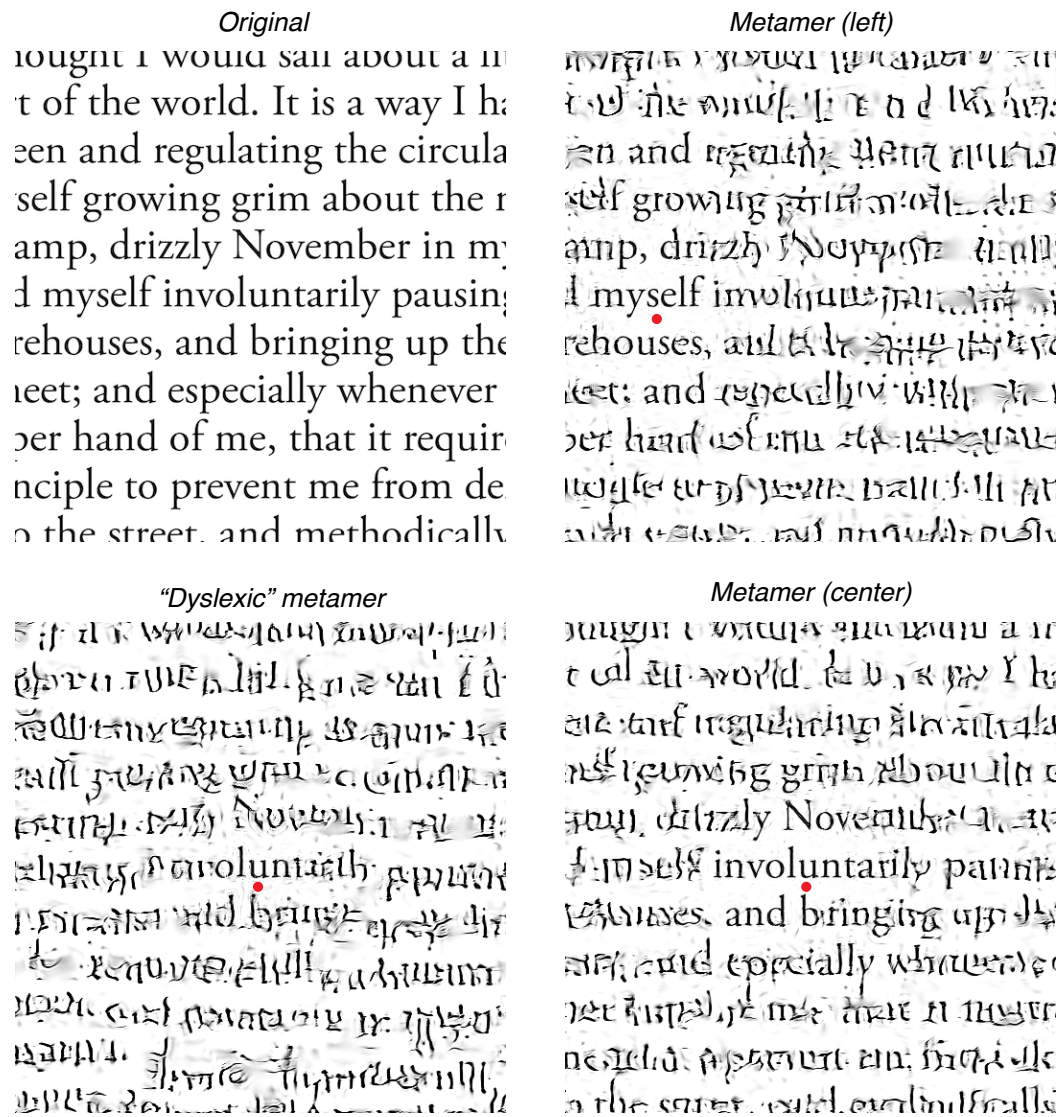


Figure 5.11: Dyslexia may in part reflect excessive pooling in the periphery. Metamers were synthesized from a block of text (from Herman Melville's *Moby Dick*), at two locations (red dots) reflecting the typical distances observers traverse while reading. Two metamers were generated using the V2 scaling (right), and one (lower left) was generated using a higher scaling (0.75). Larger scaling (and thus more crowding) may be a factor in dyslexia.

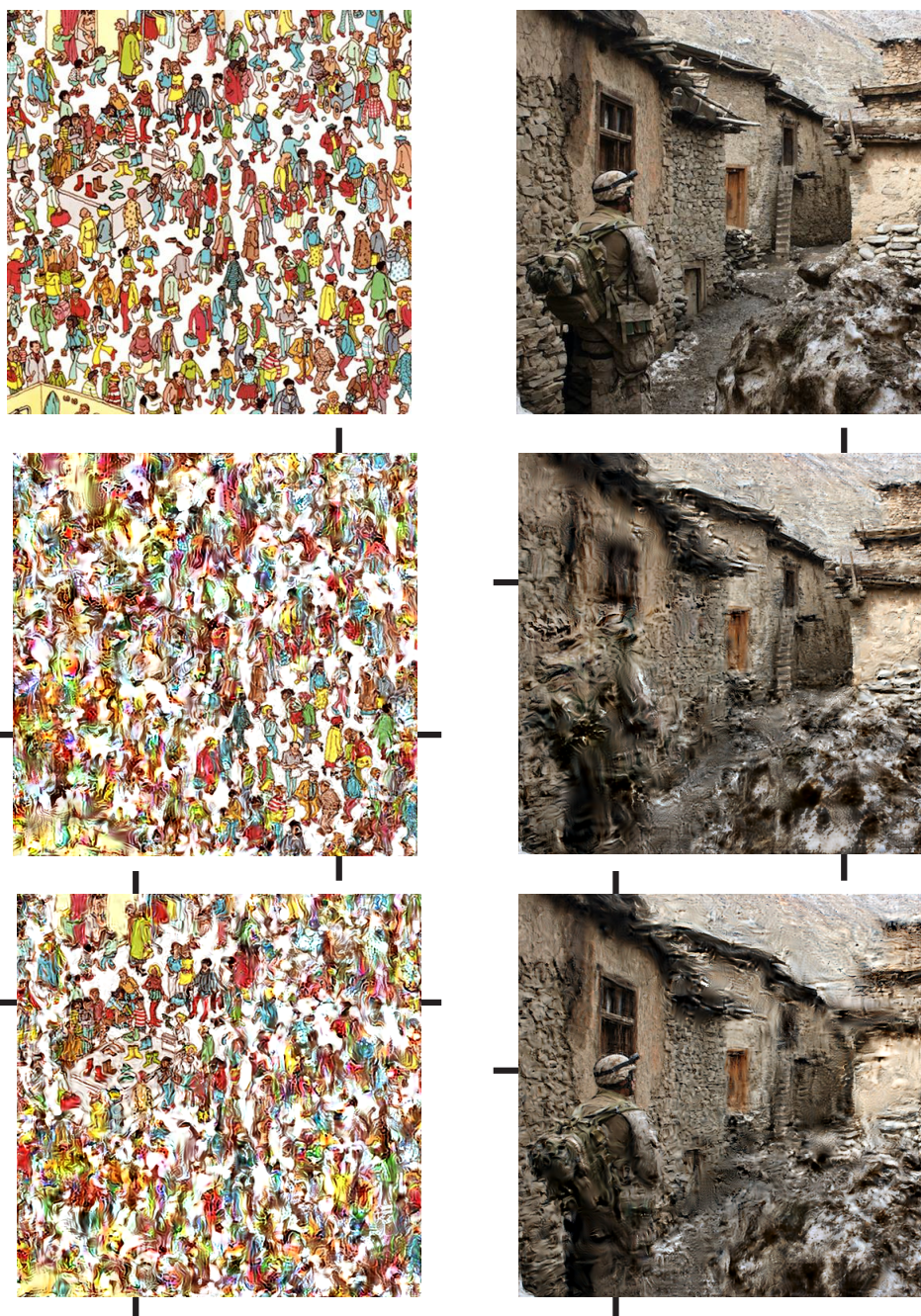


Figure 5.12: (Left) The notoriously hard to find Waldo blends into the distracting background and is only recognizable when fixated. Cross-hairs indicate the location of the fovea used by the model during synthesis. (Right) A soldier in Afghanistan wears patterned clothing to blend into the stony texture of the environment.

5.6 Discussion

We have constructed a model for visual scene representation in V2, based on local correlations among V1 responses within eccentricity-dependent pooling regions. We have developed a method for generating complex heterogeneous images with identical model responses, and used these synthetic images to show that: (1) when the pooling region sizes of the model were set correctly, images with identical model responses were indistinguishable (metameric) to human observers, despite substantial variation among features in the periphery; (2) the critical pooling size required to produce metamerism was robust to bottom-up and top-down manipulations of discrimination performance; (3) critical pooling sizes were consistent with the eccentricity dependence of receptive field sizes of neurons in ventral visual area V2; and (4) the model predicted degradations of peripheral recognition due to visual crowding, as a function of both spacing and eccentricity.

Perceptual deficits in peripheral vision have been recognized for centuries. Most early literature focuses on the loss of acuity that results from eccentricity-dependent sampling and blurring in the earliest visual stages. Crowding is a more complex deficit [21]. In a prescient article in 1976, Jerome Lettvin gave a subjective account of this phenomenon, describing letters embedded in text as having “lost form without losing crispness,” and concluding that “the embedded [letter] only seems to have a statistical existence” [131]. Lettvin’s article seems to have drifted into obscurity, but these ideas have been formalized in recent literature that explains crowding in terms of excessive averaging or pooling of features [164, 168, 91, 11]. Balas et. al. (2009), in particular, hypothesized that crowding is a manifestation of the representation of peripheral visual content with local summary statistics. They showed that human recognition performance for crowded letters was matched to that of foveally viewed

images synthesized to match the statistics of the original stimulus (computed over a localized region containing both the letter and flankers).

Our model provides an instantiation of these pooling hypotheses that operates over the entire visual field, which, in conjunction with our synthesis approach, enabled several scientific advances. First, we validated the model with a metamer discrimination paradigm, which provides a more direct test than comparisons to recognition performance in a crowding experiment. Second, the parameterization of eccentricity dependence allowed us to estimate the size of pooling regions, and thus to associate the model with a distinct stage of ventral stream processing. Third, our implementation allowed us to examine crowding in stimuli extending beyond a single pooling region, and thus to account for the dependence of recognition on both eccentricity and spacing – the defining properties of crowding [169].

The interpretation of our experimental results relies on assumptions about the representation of, and access to, information in the brain. This is perhaps best understood by analogy to trichromacy [229]. Color metamers occur because information is lost by the cones and cannot be recovered in subsequent stages. But color appearance judgements clearly do not imply direct, conscious, access to the responses of those cones. Analogously, our experiments imply that the information loss ascribed to areas V1 and V2 cannot be recovered or accessed by subsequent stages of processing (two stimuli that are V1 metamers, for example, should also be V2 metamers). But this does not imply that observers directly access the information represented in V1 or V2. Indeed, if observers could access V1 responses, then any additional information loss incurred when those responses are combined and pooled in V2 would have no perceptual consequence, and the stimuli generated by the V2 model would not appear metameric!

The loss of information in our model arises directly from its architecture – the

set of statistics, and the pooling regions over which they are computed – and this determines the set of metameric stimuli. Discriminability of non-metameric stimuli depends on the strength of the information preserved by the model, relative to noise. As seen in the presentation time and attention control experiments, manipulations of signal strength did not alter the metamerity of stimuli, and thus did not affect estimates of critical scaling. These results are also consistent with the crowding literature. Crowding effects are robust to presentation time [214], and attention can increase performance in crowding tasks while yielding small or no changes in critical spacing [132, 196]. Certain kinds of exogenous cues, however, may reduce critical spacing [238], and perceptual learning has been shown to reduce critical spacing through several days of intensive training [44]. If either manipulation were found to reduce critical scaling (as estimated from a metamer discrimination experiment), we would interpret this as arising from a reduction in receptive field sizes, which could be verified through electrophysiological measurements.

From a physiological perspective, our model is deliberately simplistic: We expect that incorporating more realistic response properties (e.g., spike generation, feedback circuitry) would not significantly alter the information represented in model populations, but would render the synthesis of stimuli computationally intractable. Despite the simplicity of the model, the metamer experiments do not uniquely constrain the response properties of individual model neurons. This may again be understood by analogy with the case of trichromacy: color matching experiments constrain the linear subspace spanned by the three cone absorption spectra, but do not uniquely constrain the spectra of the individual cones [229]. Thus, identification of V2 as the area in which the model resides does not imply that responses of individual V2 neurons encode local correlations.

The experiments described in Chapters 2 and 3 provided more direct evidence

linking this model to V2, and to specific perceptual capabilities that may rely on the responses of individual V2 neurons to homogenous texture patterns. In particular, as described in Chapter 3, statistically-matched homogenous textures yielded similar neuronal population responses in V2. That result suggests a specific neurophysiological basis for the metamerism described here, but a more direct test would be to measure the responses of entire neuronal populations in V2 while presenting full-field metamer stimuli. This could be accomplished through array recordings or imaging in non-human primates, or through fMRI adaptation experiments in humans. At the same time, even the single-unit physiological experiments described in Chapters 3 and 4 did not constrain precisely the computations performed by individual neurons in V2. Those experiments established an overall differential response to stimuli with the naturalistic features captured by the model, which suggests, but does not demonstrate, that individual V2 neurons are selective to particular model components.

Finally, one might ask why the ventral stream discards such a significant amount of information. Theories of object recognition posit that the growth of receptive field sizes in consecutive areas, as well as with eccentricity, confers invariance to geometric transformations, and cascaded models based on filtering, simple nonlinearities, and successively broader spatial pooling have been used to explain such invariances measured in area IT [179, 241, 197, 183]. Our model closely resembles the early stages of these models, but our inclusion of eccentricity-dependent pooling, and the invariance to feature scrambling revealed by the metamerism of our synthetic stimuli, seems to be at odds with the goal of object recognition [189, 60, 61]. One potential resolution of this conundrum is that the two forms of invariance arise in distinct parallel pathways. An alternative possibility is that a texture-like representation in the early ventral stream provides a substrate for object representations in

later stages. Such a notion was suggested by Lettvin, who hypothesized that “texture, somewhat redefined, is the primitive stuff out of which form is constructed” [131]. If so, the metamer paradigm introduced here may provide a powerful tool for exploring the nature of invariances arising in subsequent stages of the ventral stream.

Notes

¹⁵There are six ways to order two of the three images, which determines the first two images in the ABX sequence. In each case, there are two choices for the probe image. So there are 12 unique sequences total.

Chapter 6

Conclusion

In this thesis we have described a set of novel, behaviorally-relevant response properties that distinguish neurons in the second visual area of both macaques and humans. We accomplished this by using a model of natural image statistics to generate experimental stimuli, and used those stimuli in physiological and perceptual experiments in humans and macaques. We were able to interpret the unique physiological responses in V2 in terms of a family of hierarchical models. We also linked the physiological responses to perceptual capabilities, thereby establishing a functional role for neuronal responses in V2. In both perceptual and physiological experiments, we described a form of perceptual invariance related to the processing of these stimuli that is notable both for its similarities to, and differences from, conceptions of invariance employed in the study of shape and object representation. In the context of a population model for V2, we showed that these invariances have profound consequences for the capabilities and limits of everyday vision.

Below, we summarize the four most important conceptual and methodological advances of the work, and then note its limitations and discuss four avenues for future exploration.

6.1 Summary of contributions

6.1.1 Artificial versus natural stimuli

A widespread debate in sensory neuroscience is whether to use natural or artificial stimuli. Advocates of natural stimuli recognize that they become unwieldy when fitting functional models to describe the responses of cortical neurons, but claim that only they contain the features that will drive neurons with complex selectivity, like those in V2 or higher areas, to respond. Advocates of artificial stimuli have recognized that they may not contain all of those features, but prefer to retain the control that artificial stimuli afford. Our success in V2 using synthetic “naturalistic” stimuli may point to a new middle ground in this debate that should prove useful not only in vision, but also in other sensory domains [145]

6.1.2 Framework for hierarchical modeling

The experimental approach we used is unusual in sensory neuroscience, in that we used a model to generate targeted experimental stimuli for testing hypotheses about neurons, rather than to describe their responses to arbitrary inputs. But we exploited our understanding of the stimulus generation – the statistical properties imposed in synthetic images – to propose a family of computations capable of explaining some of the response properties we observed. The proposed computations fit broadly within the tradition of feedforward models of cortical processing, but also emphasize new components that have been ignored in previous efforts to describe V2, although discussed widely in the context of image modeling and computer graphics. Contextualizing models of neuronal function among studies of image statistics [195]

is a powerful approach that will prove important as we further develop the model proposed here.

6.1.3 Using perception to guide physiology

This thesis began with physiology, followed by perceptual correlates and consequences. But we performed the research in the opposite order. We began with the perceptual demonstration of information loss described in Chapter 5, and were inspired by those results to use related stimuli in physiological experiments. That behavioral demonstrations of information loss can place powerful constraints on physiological representation is well known to students of trichromacy, but has been used rarely in the study of pattern vision. Furthermore, many of the perceptual experiments described in Chapter 4 were performed alongside the initial physiology experiments, and helped guide their design. This work thus stands as a testament to the usefulness of perception in guiding investigations of intermediate representations deep within an information processing system.

6.1.4 Emphasizing stuff over things

Our success in V2 stemmed partly from critiquing the intuition that V2 neurons must encode the the feature combinations that are useful for segmenting scenes and individuating shapes and objects. The reality of vision is that the world largely consists of messy, complicated stuff [3, 1]. Our experience of apprehending objects leads us to believe that we delineate the mess into precise contours and boundaries, but that does not imply that the visual system does it that way, especially not in its intermediate stages. We were able to differentiate V2 neurons from V1 using stimuli based more on stuff than on things. Although we do not yet understand how

sensitivity to the statistical properties of these stimuli relates to object recognition, if it does at all, we suspect that it reflects a key stage of intermediate representation in the primate visual system.

6.2 Future work

Despite the advances presented in this thesis, we have only opened a crack in a very large door. Compared to our understanding of earlier areas, we are far from understanding neurons in V2. We do not yet know what mechanisms give rise to the distinctive responses we measured in V2, and we do not know along what dimensions individual neurons in V2 are selective or invariant. Below are four directions for future work, inspired by the approaches taken thus far.

6.2.1 Fitting nonlinear hierarchical models

We proposed a hierarchical “subunit” model capable of explaining the differential response to naturalistic stimuli found in V2. In its most general form, the model can exhibit both selectivity and invariance to complex magnitude relationships among orientations, spatial frequencies, and positions. The model reflects existing efforts in statistical modeling of images [195, 123], computer graphics [175], and canonical cascade models of cortical visual processing [36]. In future work, we should further explore the relationship between these efforts, and assess the consequences of incorporating additional components, like normalization [97], to our proposed model. Of particular theoretical interest is how implicit representations of magnitude dependencies, e.g. through normalization in V1, relate to explicit representations that may appear in downstream areas. In the physiology, we can explore these and other

ideas by directly fitting the model to data, and by relating it to the circuitry between V1 and V2. Similar questions may also be fruitfully explored in other systems where the underlying circuitry is more tractable [162, 160, 161, 71].

6.2.2 Generating new synthetic stimuli

We generated synthetic naturalistic stimuli using an existing model [175]. Having established response properties in V2 related to these stimuli, it would be useful to develop a new model for image synthesis, alongside, and tightly yoked, to more direct physiological characterization. Integrating model and stimuli has many advantages: it makes it more tractable to fit a model to neural responses, it clarifies how particular aspects of the stimuli might bias the fits, it allows expansions of the model to occur alongside concomitant changes to the stimuli, and it helps ensure that the stimuli effectively sample the space described by the model. The latter benefit could be enhanced through online synthesis during experiments. Adaptive stimuli have long been used in psychophysics [233], but only recently in physiology [134, 38, 108]. Our paradigm, emphasizing stimulus synthesis in complement to model fitting, is well-posed to incorporate online stimulus design.

6.2.3 Reconceptualizing invariance

The statistical invariance we identified with our textures seems complementary to the invariance to changes in position and size and other physical properties described in studies of object representation. Invariances to physical changes can be readily expressed in terms of the transformations, but it is non-trivial to specify a computation that would yield a constant and selective response in spite of the transformation; indeed, this is the “heart” of the object recognition problem [60, 61]. In

contrast, for an invariant representation of statistical properties, we can specify a computation that yields a constant response to a self-similar family of stimuli – the texture model employed in Chapters 2 and 3 provides such an example – but it is difficult, if not meaningless, to express the corresponding physical transformations; try describing how one patch of grass turns into another. Are these two forms of invariance encoded separately in the visual system, or are they intermingled? If they are learned through experience, is the form of learning similar? Studying the transformation from V2 to V4 could be particularly informative, as V4 neurons encode contours and exhibit some of the position and size insensitivity found downstream, but presumably receive inputs that signal the kinds of statistical selectivities and invariances identified here.

6.2.4 From textures to objects?

We have focused almost entirely on texture in V2, but the brain must eventually encode information about the shapes and identities of objects. It remains unclear how texture contributes to that goal. One possibility is that there begins in V2 a fundamental separation between two modes of processing. There is a long history in visual neuroscience in drawing a set of related distinctions: things/stuff, objects/texture, attentive/preattentive, conscious/unconscious, recognition/navigation, foveal/peripheral. It seems plausible that the representations we have identified are specific to the parafovea, and used primarily for downstream signals that emphasize navigation and coarse scene geometry, whereas processing in the fovea is something else entirely. But another possibility is that, at least at the level of V2, there is no hard line between the specific feature combinations that comprise objects, and the statistics of local features that characterize textures. They

are intermingled in V2, and both subserve downstream processing. Characterizing the heterogeneity of selectivity across large populations of V2 neurons, including those closer to the fovea, is an important goal for future work.

As we and others make progress ascending the hierarchy from V1 to V2, and as other groups continue to explore representations further downstream, the goal of achieving a complete understanding of computational transformations along the primate visual pathway begins to feel within reach.

Bibliography

- [1] E. H. Adelson. On seeing stuff: the perception of materials by humans and machines. *Proceedings of the SPIE*, 2001.
- [2] E. H. Adelson and J. R. Bergen. Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America*, 2(2):284–99, 1985.
- [3] E. H. Adelson and J. R. Bergen. *Computational Models of Visual Processing*, chapter The plenoptic function and the elements of early vision, pages 3–20. MIT Press, Cambridge, MA, 1991.
- [4] J. M. Ales, F. Farzin, B. Rossion, and A. M. Norcia. An objective method for measuring face detection thresholds using the sweep steady-state visual evoked response. *Journal of Vision*, 12(10), 2012.
- [5] J. M. Allman and J. H. Kaas. Representation of the visual field in striate and adjoining cortex of the owl monkey (*aotus trivirgatus*). *Brain Research*, 35(1):89–106, 1971.
- [6] J. M. Allman and J. H. Kaas. The organization of the second visual area (v ii) in the owl monkey: a second order transformation of the visual hemifield. *Brain Research*, 76(2):247–65, 1974.

- [7] A. Angelucci. Contribution of feedforward, lateral and feedback connections to the classical receptive field center and extra-classical receptive field surround of primate v1 neurons. *Progress in Brain Research*, 2006.
- [8] A. Anzai, X. Peng, and D. C. Van Essen. Neurons in monkey visual area v2 encode combinations of orientations. *Nature Neuroscience*, 10(10):1313–21, 2007.
- [9] C. I. Baker, J. Liu, L. L. Wald, K. K. Kwong, T. Benner, and N. Kanwisher. Visual word processing and experiential origins of functional selectivity in human extrastriate cortex. *Proceedings of the National Academy of Sciences*, 104(21):9087–92, 2007.
- [10] B. Balas. Attentive texture similarity as a categorization task: comparing texture synthesis models. *Pattern Recognition*, 2008.
- [11] B. Balas, L. Nakano, and R. Rosenholtz. A summary-statistic representation in peripheral vision explains visual crowding. *Journal of Vision*, 2009.
- [12] B. J. Balas. Texture synthesis and perception: using computational models to study texture representations in the human visual system. *Vision Research*, 46(3):299–309, 2006.
- [13] H. B. Barlow. Possible principles underlying the transformations of sensory messages. *Sensory Communication*, pages 1–10, 1961.
- [14] H. O. D. Beeck and R. Vogels. Spatial sensitivity of macaque inferior temporal neurons. *Journal of Comparative Neurology*, 426(4):505–18, 2000.

- [15] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–59, 1995.
- [16] A. J. Bell and T. J. Sejnowski. The "independent components" of natural scenes are edge filters. *Vision Research*, 37(23):3327–38, 1997.
- [17] J. R. Bergen. Texture perception: filters, non-linearities and statistics. In *Proceedings Investigative Ophthalmology and Visual Science (ARVO)*, volume 35, page 1477, 1994.
- [18] J. R. Bergen and E. H. Adelson. Early vision and texture perception. *Nature*, 333(6171):363–4, 1988.
- [19] J. R. Bergen and B. Julesz. Parallel versus serial processing in rapid pattern discrimination. *Nature*, 303(5919):696–8, 1983.
- [20] A. B. Bonds. Temporal dynamics of contrast gain in single cells of the cat striate cortex. *Visual Neuroscience*, 6(3):239–55, 1991.
- [21] H. Bouma. Interaction effects in parafoveal letter recognition. *Nature*, 226(5241):177–8, 1970.
- [22] G. M. Boynton, J. B. Demb, G. H. Glover, and D. J. Heeger. Neuronal basis of contrast discrimination. *Vision Research*, 39(2):257–69, 1999.
- [23] G. M. Boynton, S. A. Engel, G. H. Glover, and D. J. Heeger. Linear systems analysis of functional magnetic resonance imaging in human v1. *Journal of Neuroscience*, 16(13):4207–21, 1996.

- [24] C. E. Bredfeldt, J. C. A. Read, and B. G. Cumming. A quantitative explanation of responses to disparity-defined edges in macaque v2. *Journal of Neurophysiology*, 101(2):701–13, 2009.
- [25] S. L. Brincat and C. E. Connor. Underlying principles of visual shape selectivity in posterior inferotemporal cortex. *Nature Neuroscience*, 7(8):880–6, 2004.
- [26] 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.
- [27] 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.
- [28] G. J. Brouwer and D. J. Heeger. Decoding and reconstructing color from responses in human visual cortex. *Journal of Neuroscience*, 29(44):13992–4003, 2009.
- [29] R. W. Buccigrossi and E. P. Simoncelli. Image compression via joint statistical characterization in the wavelet domain. *IEEE Transactions on Image Processing*, 8(12):1688–701, 1999.
- [30] A. Burkhalter and D. C. V. Essen. Processing of color, form and disparity information in visual areas vp and v2 of ventral extrastriate cortex in the macaque monkey. *Journal of Neuroscience*, 6(8):2327–51, 1986.
- [31] L. Busse, A. R. Wade, and M. Carandini. Representation of concurrent stimuli by population activity in visual cortex. *Neuron*, 64(6):931–42, 2009.

- [32] 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.
- [33] T. Caelli and B. Julesz. On perceptual analyzers underlying visual texture discrimination: part i. *Biological Cybernetics*, 28(3):167–75, 1978.
- [34] T. Caelli, B. Julesz, and E. Gilbert. On perceptual analyzers underlying visual texture discrimination: Part ii. *Biological Cybernetics*, 29(4):201–14, 1978.
- [35] 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, 2005.
- [36] M. Carandini and D. J. Heeger. Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1):51–62, 2012.
- [37] M. Carandini, D. J. Heeger, and J. A. Movshon. Linearity and normalization in simple cells of the macaque primary visual cortex. *Journal of Neuroscience*, 17(21):8621–44, 1997.
- [38] E. T. Carlson, R. J. Rasquinha, K. Zhang, and C. E. Connor. A sparse object coding scheme in area v4. *Current Biology*, 21(4):288–93, 2011.
- [39] J. R. Cavanaugh, W. Bair, and J. A. Movshon. Nature and interaction of signals from the receptive field center and surround in macaque v1 neurons. *Journal of Neurophysiology*, 88(5):2530–46, 2002.
- [40] J. R. Cavanaugh, W. Bair, and J. A. Movshon. Selectivity and spatial distribution of signals from the receptive field surround in macaque v1 neurons. *Journal of Neurophysiology*, 88(5):2547–56, 2002.

- [41] X. Chen, F. Han, M.-M. M. Poo, and Y. Dan. Excitatory and suppressive receptive field subunits in awake monkey primary visual cortex (v1). *Proceedings of the National Academy of Sciences*, 104(48):19120–5, 2007.
- [42] E. J. Chichilnisky. A simple white noise analysis of neuronal light responses. *Network: Computation in Neural Systems*, 12(2):199–213, 2001.
- [43] C. Chubb and M. Landy. *Computational Models of Visual Processing*, chapter Orthogonal distribution analysis: A new approach to the study of texture perception, pages 291–301. MIT Press, Cambridge, MA, 1991.
- [44] S. T. L. Chung. Learning to identify crowded letters: does it improve reading speed? *Vision Research*, 47(25):3150–9, 2007.
- [45] M. R. Cohen and J. H. R. Maunsell. A neuronal population measure of attention predicts behavioral performance on individual trials. *Journal of Neuroscience*, 30(45):15241–53, 2010.
- [46] 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.
- [47] D. D. Cox, P. Meier, N. Oertelt, and J. J. DiCarlo. 'breaking' position-invariant object recognition. *Nature Neuroscience*, 8(9):1145–7, 2005.
- [48] E. Craft, H. Schütze, E. Niebur, and R. von der Heydt. A neural model of figure-ground organization. *Journal of Neurophysiology*, 97(6):4310–26, 2007.

- [49] Y. L. Cun, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, 1989.
- [50] A. M. Dale. Optimal experimental design for event-related fmri. *Human Brain Mapping*, 8(2-3):109–14, 1999.
- [51] S. V. David and J. L. Gallant. Predicting neuronal responses during natural vision. *Network: Computation in Neural Systems*, 16(2-3):239–260, 2005.
- [52] 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.
- [53] A. Dawid and A. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28, 1979.
- [54] G. C. DeAngelis, I. Ohzawa, and R. D. Freeman. Spatiotemporal organization of simple-cell receptive fields in the cat’s striate cortex. ii. linearity of temporal and spatial summation. *Journal of Neurophysiology*, 69(4):1118–35, 1993.
- [55] G. C. DeAngelis, J. G. Robson, I. Ohzawa, and R. D. Freeman. Organization of suppression in receptive fields of neurons in cat visual cortex. *Journal of Neurophysiology*, 68(1):144–63, 1992.
- [56] L. V. der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.

- [57] R. Desimone and S. J. Schein. Visual properties of neurons in area v4 of the macaque: sensitivity to stimulus form. *Journal of Neurophysiology*, 57(3):835–68, 1987.
- [58] R. DeValois and K. DeValois. *Spatial Vision*. Oxford University Press, New York, NY, 1988.
- [59] E. DeYoe, D. J. Felleman, D. C. Van Essen, and E. McClendon. Multiple processing streams in occipitotemporal visual cortex. *Nature*, 371:151–154, 1994.
- [60] J. J. DiCarlo and D. D. Cox. Untangling invariant object recognition. *Trends in Cognitive Sciences*, 11(8):333–41, 2007.
- [61] J. J. DiCarlo, D. Zoccolan, and N. C. Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–34, 2012.
- [62] S. O. Dumoulin and B. A. Wandell. Population receptive field estimates in human visual cortex. *Neuroimage*, 39(2):647–60, 2008.
- [63] Y. El-Shamayleh, R. Kumbhani, N. Dhruv, and J. Movshon. Visual response properties of v1 neurons projecting to v2 in macaque. *Society for Neuroscience Abstracts*, 404.3, 2009.
- [64] Y. El-Shamayleh and J. A. Movshon. Neuronal responses to texture-defined form in macaque visual area v2. *Journal of Neuroscience*, 31(23):8543–55, 2011.
- [65] Y. El-Shamayleh, J. A. Movshon, and L. Kiorpes. Development of sensitivity to visual texture modulation in macaque monkeys. *Journal of Vision*, 10(11):1–12, 2010.

- [66] C. Enroth-Cugell and J. G. Robson. The contrast sensitivity of retinal ganglion cells of the cat. *Journal of Physiology*, 187(3):517–52, 1966.
- [67] F. Fang, H. Boyaci, and D. Kersten. Border ownership selectivity in human early visual cortex and its modulation by attention. *Journal of Neuroscience*, 29(2):460–5, 2009.
- [68] D. J. Felleman and D. C. Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1(1):1–47, 1991.
- [69] G. Felsen and Y. Dan. A natural approach to studying vision. *Nature Neuroscience*, 8(12):1643–6, 2005.
- [70] G. Felsen, J. Touryan, F. Han, and Y. Dan. Cortical sensitivity to visual features in natural scenes. *PLoS Biology*, 3(10):1819–1828, 2005.
- [71] G. D. Field, J. L. Gauthier, A. Sher, M. Greschner, T. A. Machado, L. H. Jepson, J. Shlens, D. E. Gunning, K. Mathieson, W. Dabrowski, L. Paninski, A. M. Litke, and E. J. Chichilnisky. Functional connectivity in the retina at the resolution of photoreceptors. *Nature*, 467(7316):673–7, 2010.
- [72] J. Freeman, G. J. Brouwer, D. J. Heeger, and E. P. Merriam. Orientation decoding depends on maps, not columns. *Journal of Neuroscience*, 31(13):4792–4804, 2011.
- [73] J. Freeman, T. H. Donner, and D. J. Heeger. Inter-area correlations in the ventral visual pathway reflect feature integration. *Journal of Vision*, 11(4), 2011.
- [74] J. Freeman and D. G. Pelli. An escape from crowding. *Journal of Vision*, 7(2):22–22, 2007.

- [75] J. Freeman and E. P. Simoncelli. Metamers of the ventral stream. *Nature Neuroscience*, 14(9):1195–201, 2011.
- [76] J. Freeman and C. M. Ziemba. Unwrapping the ventral stream. *Journal of Neuroscience*, 31(7):2349–2351, 2011.
- [77] R. D. Freeman, I. Ohzawa, and G. Walker. Beyond the classical receptive field in the visual cortex. *Progress in Brain Research*, 134:157–70, 2001.
- [78] K. Fukushima. Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, 1980.
- [79] J. L. Gallant, J. Braun, and D. C. Van Essen. Selectivity for polar, hyperbolic, and cartesian gratings in macaque visual cortex. *Science*, 259(5091):100–3, 1993.
- [80] J. L. Gardner, E. P. Merriam, J. A. Movshon, and D. J. Heeger. Maps of visual space in human occipital cortex are retinotopic, not spatiotopic. *Journal of Neuroscience*, 28(15):3988–99, 2008.
- [81] R. Gattass, C. G. Gross, and J. H. Sandell. Visual topography of v2 in the macaque. *Journal of Comparative Neurology*, 201(4):519–39, 1981.
- [82] R. Gattass, A. P. Sousa, and C. G. Gross. Visuotopic organization and extent of v3 and v4 of the macaque. *Journal of Neuroscience*, 8(6):1831–45, 1988.
- [83] R. Gattass, A. P. Sousa, and M. G. Rosa. Visual topography of v1 in the cebus monkey. *Journal of Comparative Neurology*, 259(4):529–48, 1987.

- [84] G. Geiger, J. Y. Lettvin, and O. Zegarra-Moran. Task-determined strategies of visual process. *Brain Research*, 1(1):39–52, 1992.
- [85] W. S. Geisler and D. G. Albrecht. Cortical neurons: isolation of contrast gain control. *Vision Research*, 32(8):1409–10, 1992.
- [86] W. S. Geisler, J. S. Perry, B. J. Super, and D. P. Gallogly. Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research*, 41(6):711–24, 2001.
- [87] N. Graham. *Visual pattern analyzers*. Oxford University Press, New York, NY, 1989.
- [88] N. Graham. *Computational models of visual processing*, chapter Complex channels, early local nonlinearities, and normalization in texture segregation, pages 273–290. MIT Press, Cambridge, MA, 1991.
- [89] N. Graham. Breaking the visual stimulus into parts. *Current Directions in Psychological Science*, 1992.
- [90] G. Granlund. In search of a general picture processing operator. *Computer Graphics and Image Processing*, 1978.
- [91] J. A. Greenwood, P. J. Bex, and S. C. Dakin. Positional averaging explains crowding with letter-like stimuli. *Proceedings of the National Academy of Sciences*, 106(31):13130–5, 2009.
- [92] K. Grill-Spector and R. Malach. The human visual cortex. *Annu Rev Neurosci*, 27:649–77, 2004.

- [93] Y. Gu, D. E. Angelaki, and G. C. Deangelis. Neural correlates of multisensory cue integration in macaque mstd. *Nature Neuroscience*, 11(10):1201–10, 2008.
- [94] R. Haefner, S. Gerwinn, J. Macke, and M. Bethge. Inferring decoding strategy from choice probabilities in the presence of noise correlations. *Nature Precedings*, 2012.
- [95] L. E. Hallum, M. S. Landy, and D. J. Heeger. Human primary visual cortex (v1) is selective for second-order spatial frequency. *Journal of Neurophysiology*, 105(5):2121–31, 2011.
- [96] J. H. Hedges, A. A. Stocker, and E. P. Simoncelli. Optimal inference explains the perceptual coherence of visual motion stimuli. *Journal of Vision*, 11(6), 2011.
- [97] D. J. Heeger. Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, 9:181–197, 1992.
- [98] D. J. Heeger and J. R. Bergen. Pyramid-based texture analysis/synthesis. *Proceedings of SIGGRAPH*, pages 229–238, 1995.
- [99] D. J. Heeger, G. M. Boynton, J. B. Demb, E. Seidemann, and W. T. Newsome. Motion opponency in visual cortex. *Journal of Neuroscience*, 19(16):7162–74, 1999.
- [100] D. J. Heeger, A. C. Huk, W. S. Geisler, and D. G. Albrecht. Spikes versus bold: what does neuroimaging tell us about neuronal activity? *Nature Neuroscience*, 3(7):631–3, 2000.

- [101] D. J. Heeger, E. P. Simoncelli, and J. A. Movshon. Computational models of cortical visual processing. *Proceedings of the National Academy of Sciences*, 93(2):623–7, 1996.
- [102] J. Hegdé and D. C. V. Essen. Selectivity for complex shapes in primate visual area v2. *Journal of Neuroscience*, 20(61):1–6, 2000.
- [103] J. Hegdé and D. C. V. Essen. A comparative study of shape representation in macaque visual areas v2 and v4. *Cerebral Cortex*, 17(5):1100–16, 2007.
- [104] J. M. Hillis, M. O. Ernst, M. S. Banks, and M. S. Landy. Combining sensory information: mandatory fusion within, but not between, senses. *Science*, 298(5598):1627–30, 2002.
- [105] G. Hinton and S. Roweis. Stochastic neighbor embedding. *Advances in neural information processing systems*, 15:833–840, 2002.
- [106] S. Hochstein and R. M. Shapley. Linear and nonlinear spatial subunits in y cat retinal ganglion cells. *Journal of Physiology*, 262(2):265–84, 1976.
- [107] G. D. Horwitz, E. J. Chichilnisky, and T. D. Albright. Cone inputs to simple and complex cells in v1 of awake macaque. *Journal of Neurophysiology*, 97(4):3070–3081, 2007.
- [108] G. D. Horwitz and C. A. Hass. Nonlinear analysis of macaque v1 color tuning reveals cardinal directions for cortical color processing. *Nature Neuroscience*, 15(6):913–9, 2012.
- [109] D. Hubel and T. Wiesel. Q&a: David hubel and torsten wiesel. *Neuron*, 75(2):182–4, 2012.

- [110] D. H. Hubel. Exploration of the primary visual cortex, 1955-78. *Nature*, 299(5883):515–24, 1982.
- [111] D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160:106–54, 1962.
- [112] D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *Journal of Neurophysiology*, 28:229–89, 1965.
- [113] D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology*, 195(1):215–43, 1968.
- [114] A. C. Huk and D. J. Heeger. Pattern-motion responses in human visual cortex. *Nature Neuroscience*, 5(1):72–5, 2002.
- [115] 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, 2005.
- [116] M. Ito and H. Komatsu. Representation of angles embedded within contour stimuli in area v2 of macaque monkeys. *Journal of Neuroscience*, 24(13):3313–24, 2004.
- [117] M. Ito, H. Tamura, I. Fujita, and K. Tanaka. Size and position invariance of neuronal responses in monkey inferotemporal cortex. *Journal of Neurophysiology*, 73(1):218–26, 1995.
- [118] Julesz. Visual pattern discrimination. *Information Theory, IRE Transactions on*, 8(2):84 – 92, 1962.

- [119] B. Julesz. Visual pattern discrimination. *Information Theory*, 1962.
- [120] B. Julesz. Textons, the elements of texture perception, and their interactions. *Nature*, 290(5802):91–7, 1981.
- [121] B. Julesz, E. N. Gilbert, and J. D. Victor. Visual discrimination of textures with identical third-order statistics. *Biological Cybernetics*, 31(3):137–40, 1978.
- [122] 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.
- [123] Y. Karklin and M. S. Lewicki. Emergence of complex cell properties by learning to generalize in natural scenes. *Nature*, 457(7225):83–6, 2009.
- [124] H. Knutsson and G. Granlund. Texture analysis using two-dimensional quadrature filters. *Architecture for Pattern Analysis*, 1983.
- [125] J. Koenderink and A. van Doom. Local image operators and iconic structure. *Algebraic Frames for the Perception-Action Cycle*, 1315:66–93, 1997.
- [126] M. S. Landy and J. R. Bergen. Texture segregation and orientation gradient. *Vision Research*, 31(4):679–91, 1991.
- [127] M. S. Landy and N. Graham. *The Visual Neurosciences*, chapter Visual perception of texture, pages 1106–1118. MIT Press, Cambridge, MA, 2004.
- [128] J. Larsson and D. J. Heeger. Two retinotopic visual areas in human lateral occipital cortex. *Journal of Neuroscience*, 26(51):13128–42, 2006.

- [129] T. S. Lee and M. Nguyen. Dynamics of subjective contour formation in the early visual cortex. *Proceedings of the National Academy of Sciences*, 98(4):1907–11, 2001.
- [130] P. Lennie and J. A. Movshon. Coding of color and form in the geniculostriate visual pathway. *Journal of the Optical Society of America*, 22(10):2013–33, 2005.
- [131] J. Lettvin. On seeing sidelong. *The Sciences*, 16(4):10–20, 1976.
- [132] D. M. Levi. Crowding—an essential bottleneck for object recognition: a mini-review. *Vision Research*, 48(5):635–54, 2008.
- [133] J. B. Levitt, D. C. Kiper, and J. A. Movshon. Receptive fields and functional architecture of macaque v2. *Journal of Neurophysiology*, 71(6):2517–42, 1994.
- [134] J. Lewi, R. Butera, and L. Paninski. Sequential optimal design of neurophysiology experiments. *Neural Computation*, 21(3):619–87, 2009.
- [135] N. Li and J. J. DiCarlo. Unsupervised natural experience rapidly alters invariant object representation in visual cortex. *Science*, 321(5895):1502–7, 2008.
- [136] N. Li and J. J. DiCarlo. Unsupervised natural visual experience rapidly reshapes size-invariant object representation in inferior temporal cortex. *Neuron*, 67(6):1062–75, 2010.
- [137] N. Li and J. J. DiCarlo. Neuronal learning of invariant object representation in the ventral visual stream is not dependent on reward. *Journal of Neuroscience*, 32(19):6611–20, 2012.

- [138] N. K. Logothetis and D. L. Sheinberg. Visual object recognition. *Annual Review of Neuroscience*, 19:577–621, 1996.
- [139] N. A. Macmillan, H. L. Kaplan, and C. D. Creelman. The psychophysics of categorical perception. *Psychological Review*, 84(5):452–71, 1977.
- [140] W. M. Maguire and J. S. Baizer. Visuotopic organization of the prelunate gyrus in rhesus monkey. *Journal of Neuroscience*, 4(7):1690–704, 1984.
- [141] L. E. Mahon and R. L. D. Valois. Cartesian and non-cartesian responses in lgn, v1, and v2 cells. *Visual Neuroscience*, 18(6):973–81, 2001.
- [142] R. Malach, J. B. Reppas, R. R. Benson, K. K. Kwong, H. Jiang, W. A. Kennedy, P. J. Ledden, T. J. Brady, B. R. Rosen, and R. B. Tootell. Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proceedings of the National Academy of Sciences*, 92(18):8135–9, 1995.
- [143] L. T. Maloney and J. N. Yang. Maximum likelihood difference scaling. *Journal of Vision*, 3(8):573–85, 2003.
- [144] M. Martelli, G. D. Filippo, D. Spinelli, and P. Zoccolotti. Crowding, reading, and developmental dyslexia. *Journal of Vision*, 9(4):14.1–18, 2009.
- [145] J. H. McDermott and E. P. Simoncelli. Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis. *Neuron*, 71(5):926–40, 2011.
- [146] W. H. Merigan, L. M. Katz, and J. H. R. Maunsell. The effects of parvocellular lateral geniculate lesions on the acuity and contrast sensitivity of macaque monkeys. *Journal of Neuroscience*, 11(4):994–1001, 1991.

- [147] L. Montaser-Kouhsari, M. S. Landy, D. J. Heeger, and J. Larsson. Orientation-selective adaptation to illusory contours in human visual cortex. *Journal of Neuroscience*, 27(9):2186–95, 2007.
- [148] I. Motoyoshi, S. Nishida, L. Sharan, and E. H. Adelson. Image statistics and the perception of surface qualities. *Nature*, 447(7141):206–9, 2007.
- [149] J. A. Movshon, E. H. Adelson, M. Gizzi, and W. T. Newsome. *The analysis of moving visual patterns*, volume 54, pages 117–151. Vatican Press, Rome, 1985.
- [150] J. A. Movshon, I. D. Thompson, and D. J. Tolhurst. Receptive field organization of complex cells in the cat’s striate cortex. *Journal of Physiology*, 283:79–99, 1978.
- [151] J. A. Movshon, I. D. Thompson, and D. J. Tolhurst. Spatial summation in the receptive fields of simple cells in the cat’s striate cortex. *Journal of Physiology*, 283:53–77, 1978.
- [152] O. Nestares and D. J. Heeger. Robust multiresolution alignment of mri brain volumes. *Magnetic Resonance in Medicine*, 43(5):705–15, 2000.
- [153] W. T. Newsome, K. H. Britten, and J. A. Movshon. Neuronal correlates of a perceptual decision. *Nature*, 341(6237):52–4, 1989.
- [154] H. Nienborg, M. R. Cohen, and B. G. Cumming. Decision-related activity in sensory neurons: correlations among neurons and with behavior. *Annual Review of Neuroscience*, 35:463–83, 2012.
- [155] H. Nienborg and B. G. Cumming. Macaque v2 neurons, but not v1 neurons, show choice-related activity. *Journal of Neuroscience*, 26(37):9567–78, 2006.

- [156] H. Nienborg and B. G. Cumming. Psychophysically measured task strategy for disparity discrimination is reflected in v2 neurons. *Nature Neuroscience*, 10(12):1608–14, 2007.
- [157] H. Nienborg and B. G. Cumming. Decision-related activity in sensory neurons reflects more than a neuron’s causal effect. *Nature*, 459(7243):89–92, 2009.
- [158] S. Nishimoto, T. Ishida, and I. Ohzawa. Receptive field properties of neurons in the early visual cortex revealed by local spectral reverse correlation. *Journal of Neuroscience*, 26(12):3269–80, 2006.
- [159] S. Ogawa, T. M. Lee, A. R. Kay, and D. W. Tank. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences*, 87(24):9868–72, 1990.
- [160] S. R. Olsen, V. Bhandawat, and R. I. Wilson. Divisive normalization in olfactory population codes. *Neuron*, 66(2):287–99, 2010.
- [161] S. R. Olsen, D. S. Bortone, H. Adesnik, and M. Scanziani. Gain control by layer six in cortical circuits of vision. *Nature*, 483(7387):47–52, 2012.
- [162] S. R. Olsen and R. I. Wilson. Cracking neural circuits in a tiny brain: new approaches for understanding the neural circuitry of drosophila. *Trends in Neurosciences*, 31(10):512–20, 2008.
- [163] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–9, 1996.

- [164] L. Parkes, J. Lund, A. Angelucci, J. A. Solomon, and M. Morgan. Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience*, 4(7):739–44, 2001.
- [165] A. Pasupathy and C. E. Connor. Shape representation in area v4: position-specific tuning for boundary conformation. *Journal of Neurophysiology*, 86(5):2505–19, 2001.
- [166] A. Pasupathy and C. E. Connor. Population coding of shape in area v4. *Nature Neuroscience*, 5(12):1332–8, 2002.
- [167] D. Pelli, K. Tillman, J. Freeman, M. Su, T. Berger, and N. Majaj. Crowding and eccentricity determine reading rate. *Journal of Vision*, 7(2), 2007.
- [168] D. G. Pelli, M. Palomares, and N. J. Majaj. Crowding is unlike ordinary masking: Distinguishing feature integration from detection. *Journal of Vision*, 4(12):12–12, 2004.
- [169] D. G. Pelli and K. A. Tillman. The uncrowded window of object recognition. *Nature Neuroscience*, 11(10):1129–35, 2008.
- [170] F. Pestilli, M. Carrasco, D. J. Heeger, and J. L. Gardner. Attentional enhancement via selection and pooling of early sensory responses in human visual cortex. *Neuron*, 72(5):832–46, 2011.
- [171] E. Peterhans and R. v. d. Heydt. Mechanisms of contour perception in monkey visual cortex. ii. contours bridging gaps. *Journal of Neuroscience*, 9(5):1749–63, 1989.

- [172] J. W. Pillow and E. P. Simoncelli. Dimensionality reduction in neural models: an information-theoretic generalization of spike-triggered average and covariance analysis. *Journal of Vision*, 6(4):414–28, 2006.
- [173] N. Pinto, D. D. Cox, and J. J. DiCarlo. Why is real-world visual object recognition hard? *PLoS Computational Biology*, 4(1):e27, 2008.
- [174] C. R. Ponce, S. G. Lomber, and R. T. Born. Integrating motion and depth via parallel pathways. *Nature Neuroscience*, 11(2):216–23, 2008.
- [175] J. Portilla and E. P. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, 40(1):49–70, 2000.
- [176] F. T. Qiu, T. Sugihara, and R. v. d. Heydt. Figure-ground mechanisms provide structure for selective attention. *Nature Neuroscience*, 10(11):1492–9, 2007.
- [177] D. Ress, B. T. Backus, and D. J. Heeger. Activity in primary visual cortex predicts performance in a visual detection task. *Nature Neuroscience*, 3(9):940–5, 2000.
- [178] D. Ress and D. J. Heeger. Neuronal correlates of perception in early visual cortex. *Nature Neuroscience*, 6(4):414–20, 2003.
- [179] M. Riesenhuber and T. Poggio. Are cortical models really bound by the "binding problem"? *Neuron*, 24(1):87–93, 111–25, 1999.
- [180] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–25, 1999.

- [181] M. Riesenhuber and T. Poggio. Neural mechanisms of object recognition. *Current Opinion in Neurobiology*, 12(2):162–8, 2002.
- [182] E. I. Rodionova, A. V. Revishchin, and I. N. Pigarev. Distant cortical locations of the upper and lower quadrants of the visual field represented by neurons with elongated and radially oriented receptive fields. *Exp Brain Res*, 158(3):373–7, 2004.
- [183] E. Rolls. The neurophysiology and computational mechanisms of object representation. *Object categorization: computer and human vision*, 2009.
- [184] R. Rosenholtz, J. Huang, A. Raj, B. J. Balas, and L. Ilie. A summary statistic representation in peripheral vision explains visual search. *Journal of Vision*, 12(4), 2012.
- [185] D. L. Ruderman and W. Bialek. Statistics of natural images: Scaling in the woods. *Physical Review Letters*, 73(6):814–817, 1994.
- [186] 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.
- [187] N. C. Rust and J. A. Movshon. In praise of artifice. *Nature Neuroscience*, 8(12):1647–50, 2005.
- [188] N. C. Rust, O. Schwartz, J. A. Movshon, and E. P. Simoncelli. Spatiotemporal elements of macaque v1 receptive fields. *Neuron*, 46(6):945–56, 2005.
- [189] N. C. Rust and A. A. Stocker. Ambiguity and invariance: two fundamental challenges for visual processing. *Current Opinion in Neurobiology*, pages 1–7, 2010.

- [190] K. S. Sasaki and I. Ohzawa. Internal spatial organization of receptive fields of complex cells in the early visual cortex. *Journal of Neurophysiology*, 98(3):1194–212, 2007.
- [191] K. S. Sasaki, Y. Tabuchi, and I. Ohzawa. Complex cells in the cat striate cortex have multiple disparity detectors in the three-dimensional binocular receptive fields. *Journal of Neuroscience*, 30(41):13826–37, 2010.
- [192] J. D. Schall, V. H. Perry, and A. G. Leventhal. Retinal ganglion cell dendritic fields in old-world monkeys are oriented radially. *Brain Research*, 368(1):18–23, 1986.
- [193] P. H. Schiller and J. G. Malpeli. The effect of striate cortex cooling on area 18 cells in the monkey. *Brain Research*, 126(2):366–9, 1977.
- [194] O. Schwartz, J. W. Pillow, N. C. Rust, and E. P. Simoncelli. Spike-triggered neural characterization. *Journal of Vision*, 6(4):484–507, 2006.
- [195] O. Schwartz and E. P. Simoncelli. Natural signal statistics and sensory gain control. *Nature Neuroscience*, 4(8):819–25, 2001.
- [196] M. Scolari, A. Kohnen, B. Barton, and E. Awh. Spatial attention, preview, and popout: which factors influence critical spacing in crowded displays? *Journal of Vision*, 7(2):7.1–23, 2007.
- [197] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *IEEE Trans Pattern Anal Mach Intell*, 29(3):411–26, 2007.

- [198] 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.
- [199] S. Shushruth, J. M. Ichida, J. B. Levitt, and A. Angelucci. Comparison of spatial summation properties of neurons in macaque v1 and v2. *Journal of Neurophysiology*, 102(4):2069–83, 2009.
- [200] E. P. Simoncelli. Statistical models for images: compression, restoration and synthesis. In *Asilomar Conference on Signals, Systems, and Computers*, pages 673–678, 1997.
- [201] E. P. Simoncelli and D. J. Heeger. A model of neuronal responses in visual area mt. *Vision Research*, 38(5):743–61, 1998.
- [202] E. P. Simoncelli and B. A. Olshausen. Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24:1193–216, 2001.
- [203] E. P. Simoncelli, L. Paninski, J. W. Pillow, and O. Schwartz. *The Cognitive Neurosciences*, chapter Characterization of neural responses with stochastic stimuli, pages 327–338. MIT Press, Cambridge, MA, 2004.
- [204] L. C. Sincich and J. C. Horton. The circuitry of v1 and v2: integration of color, form, and motion. *Annual Review of Neuroscience*, 28:303–26, 2005.
- [205] A. M. Smith, B. K. Lewis, U. E. Ruttimann, F. Q. Ye, T. M. Sinnwell, Y. Yang, J. H. Duyn, and J. A. Frank. Investigation of low frequency drift in fmri signal. *Neuroimage*, 9(5):526–33, 1999.
- [206] R. Sokal and F. Rohlf. *Biometry, the principles and practices of statistics in biological research*. W. H. Freeman, New York, NY, 1969.

- [207] M. W. Spratling. A single functional model accounts for the distinct properties of suppression in cortical area v1. *Vision Research*, 51(6):563–76, 2011.
- [208] K. Tanaka. Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, 19:109–39, 1996.
- [209] X. Tao, B. Zhang, E. L. Smith, S. Nishimoto, I. Ohzawa, and Y. M. Chino. Local sensitivity to stimulus orientation and spatial frequency within the receptive fields of neurons in visual area 2 of macaque monkeys. *Journal of Neurophysiology*, 107(4):1094–110, 2012.
- [210] M. G. Thomson, D. H. Foster, and R. J. Summers. Human sensitivity to phase perturbations in natural images: a statistical framework. *Perception*, 29(9):1057–69, 2000.
- [211] L. L. Thurstone. Psychophysical analysis. by I. I. thurstone, 1927. *American Journal of Psychology*, 100(3-4):587–609, 1987.
- [212] D. J. Tolhurst, Y. Tadmor, and T. Chao. Amplitude spectra of natural images. *Ophthalmic and Physiological Optics*, 12(2):229–32, 1992.
- [213] J. Touryan, B. Lau, and Y. Dan. Isolation of relevant visual features from random stimuli for cortical complex cells. *Journal of Neuroscience*, 22(24):10811–8, 2002.
- [214] J. Townsend, S. Taylor, and D. Brown. Lateral masking for letters with unlimited viewing time. *Perception & Psychophysics*, 10(5):375–378, 1971.
- [215] A. Treisman and H. Schmidt. Illusory conjunctions in the perception of objects. *Cognitive Psychology*, 14(1):107–41, 1982.

- [216] A. M. Treisman and N. G. Kanwisher. Perceiving visually presented objects: recognition, awareness, and modularity. *Current Opinion in Neurobiology*, 8(2):218–26, 1998.
- [217] D. Y. Tsao, W. A. Freiwald, R. B. H. Tootell, and M. S. Livingstone. A cortical region consisting entirely of face-selective cells. *Science*, 311(5761):670–4, 2006.
- [218] L. G. Ungerleider and J. Haxby. 'what' and 'where' in the human brain. *Current Opinion in Neurobiology*, 4:157–165, 1994.
- [219] D. C. Van Essen, C. H. Anderson, and D. J. Felleman. Information processing in the primate visual system: an integrated systems perspective. *Science*, 255(5043):419–23, 1992.
- [220] D. C. Van-Essen, W. T. Newsome, and J. H. Maunsell. The visual field representation in striate cortex of the macaque monkey: asymmetries, anisotropies, and individual variability. *Vision Research*, 24(5):429–48, 1984.
- [221] J. H. van Hateren and A. van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings Biological Sciences*, 265(1394):359–66, 1998.
- [222] J. D. Victor. Models for preattentive texture discrimination: Fourier analysis and local feature processing in a unified framework. *Spatial Vision*, 3(4):263–80, 1988.
- [223] J. D. Victor and R. M. Shapley. The nonlinear pathway of y ganglion cells in the cat retina. *Journal of General Physiology*, 74(6):671–89, 1979.

- [224] J. D. Victor and R. M. Shapley. Receptive field mechanisms of cat x and y retinal ganglion cells. *Journal of General Physiology*, 74(2):275–98, 1979.
- [225] B. Vintch, J. Movshon, and E. Simoncelli. Characterizing receptive field structure of macaque v2 neurons in terms of their v1 afferents. *Society for Neuroscience Abstracts*, 2010.
- [226] B. Vintch, A. Zaharia, J. Movshon, and E. Simoncelli. Efficient and direct estimation of a neural subunit model for sensory coding. *Accepted for presentation in Advances in Information Processing Systems*, 2012.
- [227] G. A. Walker, I. Ohzawa, and R. D. Freeman. Suppression outside the classical cortical receptive field. *Visual Neuroscience*, 17(3):369–79, 2000.
- [228] P. Wallisch and J. A. Movshon. Structure and function come unglued in the visual cortex. *Neuron*, 60(2):195–7, 2008.
- [229] B. Wandell. *Foundations of Vision*. Sinauer Associates, Sunderland, MA, 1995.
- [230] B. A. Wandell, S. O. Dumoulin, and A. A. Brewer. Visual field maps in human cortex. *Neuron*, 56(2):366–83, 2007.
- [231] H. X. Wang, D. J. Heeger, and M. S. Landy. Responses to second-order texture modulations undergo surround suppression. *Vision Research*, 62:192–200, 2012.
- [232] Z. Wang and E. P. Simoncelli. Maximum differentiation (mad) competition: a methodology for comparing computational models of perceptual quantities. *Journal of Vision*, 8(12):8.1–13, 2008.

- [233] A. B. Watson and D. G. Pelli. Quest: a bayesian adaptive psychometric method. *Perception and Psychophysics*, 33(2):113–20, 1983.
- [234] P. D. Weerd, R. Desimone, and L. G. Ungerleider. Cue-dependent deficits in grating orientation discrimination after v4 lesions in macaques. *Visual Neuroscience*, 13(3):529–38, 1996.
- [235] F. A. Wichmann and N. J. Hill. The psychometric function: I. fitting, sampling, and goodness of fit. *Perception and Psychophysics*, 63(8):1293–313, 2001.
- [236] B. D. B. Willmore, J. A. Mazer, and J. L. Gallant. Sparse coding in striate and extrastriate visual cortex. *Journal of Neurophysiology*, 2011.
- [237] B. D. B. Willmore, R. J. Prenger, and J. L. Gallant. Neural representation of natural images in visual area v2. *Journal of Neuroscience*, 30(6):2102–14, 2010.
- [238] Y. Yeshurun and E. Rashal. Precueing attention to the target location diminishes crowding and reduces the critical distance. *Journal of Vision*, 10(10):16, 2010.
- [239] B. Zenger-Landolt and D. J. Heeger. Response suppression in v1 agrees with psychophysics of surround masking. *Journal of Neuroscience*, 23(17):6884–93, 2003.
- [240] H. Zhou, H. S. Friedman, and R. v. d. Heydt. Coding of border ownership in monkey visual cortex. *Journal of Neuroscience*, 20(17):6594–611, 2000.

- [241] D. Zoccolan, M. Kouh, T. Poggio, and J. J. DiCarlo. Trade-off between object selectivity and tolerance in monkey inferotemporal cortex. *Journal of Neuroscience*, 27(45):12292–307, 2007.
- [242] M. Zorzi, C. Barbiero, A. Facoetti, I. Lonciari, M. Carrozzi, M. Montico, L. Bravar, F. George, C. Pech-Georgel, and J. C. Ziegler. Extra-large letter spacing improves reading in dyslexia. *Proceedings of the National Academy of Sciences*, 109(28):11455–9, 2012.