Biologically-inspired Neural Networks for Shape and Color Representation

Paria Mehrani

A Dissertation Submitted to the Faculty of Graduate Studies
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy

Graduate Program in
Electrical Engineering and Computer Science
York University
Toronto, Ontario

December 2021

Abstract

The goal of human-level performance in artificial vision systems is yet to be achieved. With this goal, a reasonable choice is to simulate this biological system with computational models that mimic its visual processing. A complication with this approach is that the human brain, and similarly its visual system, are not fully understood. On the bright side, with remarkable findings in the field of visual neuroscience, many questions about visual processing in the primate brain have been answered in the past few decades. Nonetheless, a lag in incorporating these new discoveries into biologically-inspired systems is evident. The present work introduces novel biologically-inspired models that employ new findings of shape and color processing into analytically-defined neural networks. In contrast to most current methods that attempt to learn all aspects of behavior from data, here we propose to bootstrap such learning by building upon existing knowledge rather than learning from scratch. Put simply, the processing networks are defined analytically using current neural understanding and learned where such knowledge is not available. This is thus a hybrid strategy that hopefully combines the best of both worlds. Experiments on the artificial neurons in the proposed networks demonstrate that these neurons mimic the studied behavior of biological cells, suggesting a path forward for incorporating analytically-defined artificial neural networks into computer vision systems.

To mom and dad,

Rezvan and Hossein,

for your ever-present love and support.

And, to my little love, Parsa,

for being an anchor to my existence.

Acknowledgments

The list is endless and emotions are in abundance. But I will try to keep it short. I would like to extend my thanks and appreciations:

To my supervisor and mentor, John Tsotsos, for his kindness, guidance, patience, support and much more (I promised to keep it short!). John, your impact on my life goes beyond a PhD and what it will bring to my life, and for that, I will remain indebted to you forever.

To my examination committee members, Michael Brown, Marcus Brubaker, Jeffrey Schall, and Thomas Serre for your time and valuable feedback.

To Minas Spetsakis, for all the many exciting discussions throughout my years at York and great advice and support when I needed it most.

To a sister of my heart, Vida Movahedi, for being an amazing friend and confidant.

To all my great colleagues, past and present, in Tsotsos lab for being great friends, scientifically and personally.

And finally, to my family:

To my mom and my dad for nurturing curiosity, love of learning, discovering and solving puzzles in me and for raising me to be strong.

To my brother, Pouria, for being my big brother and not solving my math problems, and for encouraging me to choose computer science. I delight in that choice every single day. Even oceans apart, my love reaches you.

To my son, Parsa: my little love, your contributions to this work were more than just pressing the enter button to create some of the figures. Your innocent smile on a spring afternoon made my path forward clear.

To my husband, Mohammad, for respecting my decisions and standing by my side during the

sixteen years we've been together. And, for buying the bananas in the color chapter. They turned out good.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

CNN	Convolutional neural networks
AI	Artifical Intellingence
RBO	The recurrent border ownership network
LGN	The lateral geniculate nucleus
RF	Receptive field
LMS	Long, Medium and Short wavelengths
MB	MacLeod and Boynton color space
RGB	Red, Green and Blue color space
HSL	Hue, Saturation and Lightness color space
mLGN	Model LGN cell
mV1	Model V1 cell
mV2	Model V2 cell
mV4	Model V4 cell
PCA	Principal Component Analysis
SVM	Support Vector Machines
HSV	Hue, Saturation and Value color space
V4SV	V4, Saturation and Value color space
$H_{smooth}SV$	Smoothed Hue, Saturation and Value color space
IoU	Intersection over Union
BO	Border Ownership
CRF	Conditional Random Fields
BSD	Berkeley Segmentation Dataset

DOC	Deep OCclusion network
RBO	Recurrent Border Ownership
mvSimple	Model ventral Simple cell
mvComplex	Model ventral Complex cell
mdSimple	Model dorsal Simple cell
mMT	Model MT cell
mBO	Model Border Ownership cell
DoG	Difference of Gaussians
RL	Relaxation Labeling
ROPD	Randomly Overlapping Polygons Dataset
NDR	Normalized Difference of Responses
VMI	Vector Modulation Index
APC	Angular Position and Curvature space
HMAX	The Hierarchical MAX model
2DSIL	The 2D SILhouette model
ML	Machine Learning
mEsDeg	Model Endstopped curvature Degree cell
mEsDir	Model Endstopped curvature Direction cell
mEsSign	Model endstopped curvature Sign cell
mLocalCurv	Model Local Curvature cell
SparseShape	The sparse shape model

Chapter 1

Introduction

1.1 Motivation

Successful theories and models, regardless of how they are derived, are intended to explain observed data and to predict new and not yet discovered aspects of the observed systems. In her paper "To Explain or to Predict?", Shmueli [30] defines explanation as "causal explanation" and explanatory modeling as "the use of statistical models for testing causal explanations". She then proceeds and defines predictive modeling as any approach that produces a prediction for new or unobserved input values. Although both modeling approaches might employ similar tools, for example, fitting a regression model, the distinguishing factor is the goal in each case: whereas in explanatory models the focus is on recovering an underlying causal function, in predictive approaches the data and generating good predictions are of interest. The distinct goals in each approach determines every step of a study from model design to data collection, experimental setting and measuring of performance. With dissimilar outcomes from each modeling approach, determining the goal of a study is an essential first step. Yet, regardless of the approach, models are evaluated based on both their explanatory power and predictive power as a means to gauge a model against others. Models that are far from the predictive benchmark suggest that "there are substantial practical and theoretical gains to be had from further scientific development," leading to refinement of the theory or model.

In many visual tasks, which are the focus of the present work, the human visual system with its complicated cognitive capabilities has been the predictive benchmark; a goal is to achieve human-level performance. Although some models claimed to have reached that goal [31], more recent investigations pointed out how those models fail to perform like humans and how vulnerable as visual systems those models are in certain circumstances [32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47]. In other words, there is substantial distance to reach the predictive benchmark.

To achieve this goal, a natural course of action is to design models that are inspired by the human visual system. After all, the biological machine solves many visual problems, and a goal is to match its performance. Therefore, building a vision system which imitates visual processing in the human brain has the potential to reach the goal. A difficulty in proceeding in this path is that the human brain is not fully understood, which impedes any attempts in simulating the biological visual system. But on the bright side, in the past few decades, neuroscience has come a long way

(See [48] for a summary of major advances) with remarkable new discoveries that enhanced our comprehension of the brain. Despite these advancements, biologically-inspired models that are faithful to known visual processing in the brain have not progressed accordingly. In particular, the new neuroscience findings have not been actively incorporated into biologically-inspired models, and thus, this area of research is left under-explored.

In this work, biologically-inspired models resembling visual processing in the human brain entail those that are inspired by and analytically defined according to known neural mechanisms in the visual system. Such models mimic visual areas known to be involved in certain visual tasks and are analytically-constrained with revealed properties of those visual areas. This group of models excludes Convolutional Neural Networks (CNNs) for reasons that are explained in detail in Section 1.2. Biologically-inspired models as considered in this work and at the intersection of computer vision and computational neuroscience can serve dual purposes: first, they introduce analytically-defined neural networks as automatic vision systems, second, as Brown [49] argues, they help in "understand[ing] how the elements of the brain work together to form functional units and ultimately generate the complex cognitive behaviors we study."

Despite significant progress in modern neuroscience, there is yet a tremendous amount that is unknown about the visual system; a fact that cannot be overlooked. Subsequently, any biologically-inspired model as defined above will lack the still unknown processing aspects, which inevitably impedes its performance. Yet, discovering unknown processing aspects in the biological visual system might take decades and when the goal is having a readily available biologically-inspired vision system, a prominent question is how to account for those unknown mechanisms in the model? An easy answer is: analytically define the model with known processing aspects, learn the unknowns from the data. Interestingly, with experimental data, the gains are substantial: the learned parameters of the model suggest plausible mechanisms of the visual system, serving the second purpose of biologically-inspired models mentioned above.

In this dissertation, the primary objective is to incorporate recent advances in the field of visual neuroscience into biologically-inspired computer vision systems. In particular, the goal is to lay out how certain visual features can be represented in a computational vision model that implements known mechanisms and processing in the primate visual system. As a second goal, we seek to surmount the still-unknown obstacle of the biological visual system by combining biologically-

inspired models with learning algorithms. With this latter step, we demonstrate how existing biologically-inspired models can be extended to account for the unknowns and how new falsifiable theories can be unveiled.

1.2 CNNs as Computational Vision Models

CNNs, as the mainstream architectures in many current vision applications, are hierarchical models whose computations where inspired by those of simple and complex cells in the visual cortex [50]. Additionally, these networks were reported to have achieved human-level performance in certain visual tasks [31]. Putting these two facts together brings up the matter of parallelism between the artificial and biological neural networks. In other words, a prominent question is: "How are CNNs and the brain related?"

The pursuit to discover the relationship between CNNs and the brain spans a large literature from behavioral studies to more physiological ones. In many studies, the internal representations of CNNs were compared to those of the ventral stream [51, 52, 53, 54, 55, 56, 57, 58, 59] with the suggestion of similarities. Some even reported layer-by-layer mappings between CNN layers and brain areas along the ventral stream [54, 59] hinting on similar hierarchical processing in both networks. With the advancement of computer vision and introduction of myriad variations of the underlying CNN architecture, it is unclear if the analogy remains intact for the newly-introduced architectures. In a recent attempt, Schrimpf *et al.* [60] introduced a platform[1] that scores a total of 149 hierarchical architectures, including influential biologically-inspired models such as HMAX, state-of-the-art CNNs and other non-convolutional neural networks. These models were scored against a number of neural and behavioral benchmarks. In the specific case of neural benchmarks, Schrimpf *et al.* [60] compared the representations in all layers of these architectures with V1, V2, V4 and IT representations and reported the best matching score. They found that the internal representations of CNNs were positively correlated with responses of neurons in these areas. Interestingly, they also found that hierarchical models with Transformer blocks [61] such as [62, 63] ranked lower on the list (rank 98 onward). This finding alone suggests that models with features that were inspired by the brain are more relatable to biological neural responses.

[1] https://www.brain-score.org/

These findings give the perception that all we need to build a model of the brain is enough computational power and massive training data. Having a model of the brain is significant, especially in the field of neuroscience, since if researchers can establish that CNNs give a model of the brain, then, these architectures can be employed as the general framework to further our understanding of the brain. This is an active area of research and despite reported similarities between CNN and brain representations, many other studies suggest that there are fundamental differences between CNNs and the brain [37, 38, 39, 40, 41, 42, 43, 44]. For example, Baker *et al.* [64, 65] performed a series of experiments which showed that disrupting global shape, which tremendously affects human accuracy, does not affect CNN predictions. Similarly, Geirhos *et al.* [66] demonstrated another difference between humans and CNNs in image recognition tasks where texture bias in CNNs causes these networks to recognize texture rather than shapes. In a recent study, Geirhos *et al.* [67] suggested that while the performance gap between machines and humans in presence of certain distortions to input images (for example, replacing natural images with sketches) are closing, there are still large gaps in some other distortion cases. In another work, Kim *et al.* [43] showed that CNNs fail in simple same-different tasks that are trivial to humans and suggested that the lack of feedback mechanisms could be a contributing factor. Similarly, Kar *et al.* [68] found fast recurrence mechanisms are needed for robust object recognition.

Finding dissimilarities between CNNs and the human/primate brain are consequential as the gap between the best available model and the predictive benchmark, *i.e.*, human performance, suggests that "there are substantial practical and theoretical gains to be had from further scientific development" [30]. Even with the knowledge of fundamental differences between CNNs and the human brain, CNNs might still be good models of the brain in certain processing aspects. In other words, failing to replicate the brain in a number of studies does not contradict the claim that we have a computational model of the human brain. Instead, identifying those differences, as part of a scientific endeavor, helps researchers to further improve CNNs and make those architectures even more brain-like. Currently, however, there is no agreement among researchers if CNNs model the brain [69]. Therefore, we take a step back and look at this problem from the specific angle of interest in this work. In particular, we ask: "Do CNNs, in their current state, constitute a mechanistic model of the brain?"

To answer the question put forth, one needs to have a definition of a "mechanistic model."

As mentioned above, Brown [49] defines a mechanistic model of the brain as one which helps in "understand[ing] how the elements of the brain work together to form functional units and ultimately generate the complex cognitive behaviors we study." Kay [69] also gives a similar definition as "a mechanistic model attempts to also use components that parallel the actual physical components of the system." In the particular case of the brain, Kay specifies a mechanistic model as one "that not only characterizes stimulus-response transformations, but also the way in which the brain accomplishes those transformations." The Oxford Dictionary of English [70] defines the word mechanistic as, "relating to theories which explain phenomena in purely physical or deterministic terms," hinting on the goal of explanation. Putting these together, a mechanistic model is beyond an input-output mapping and incorporates the various components that contribute to the formation of this transformation for the purpose of giving an *explanation* of the input-output relationship. These definitions suggest that mechanistic models can be put under the category of explanatory models.

With mechanistic models as those that explain the underlying observed system, we can rephrase the earlier question and ask: "Are CNNs explanatory models?" CNNs are impressive predictive models. The primary goal in the design of these architectures is to improve performance with little to no constraints on the model itself so far as the available computational power and the data allow. In other words, despite being inspired by the brain, CNNs do not incorporate our current knowledge of processing in the brain. Sejnowski [71] called this view as "brains are not relevant" and attributed it in part to the massive amount of unknowns about the brain in the early days of appearance of these models. This view is also manifested in the data-driven and end-to-end employment of CNNs. Nevertheless, these architectures might enjoy high explanatory power.

For CNNs to explain the brain, these architectures are required to reveal the various components that contribute to the stimulus-response transformation in the brain. However, a complete understanding of how these networks reach their reported performance eludes scientists. CNNs are employed in an end-to-end fashion and mainly as mysterious models that perform well. What features are learned inside these networks is still unknown and this lack of knowledge hinders our understanding of the various components contributing to the final output of these models. Moreover, why certain choices of architecture perform better than others is another open question. Therefore, it would not be surprising to consider CNNs as black-box models.

A common debate among researchers is that some argue that an understanding of CNNs and their internal representations is not necessary when it comes to having a mechanistic model of the brain so long as we have "a mathematical function" that gives the input-output mapping. In this case, one must remember that both explanatory and predictive models each yield "a function" that maps input to output. The difference is that in explanatory models the goal is for the recovered function to match the underlying phenomena, to reproduce what has already been observed. With this goal, the input-output mapping in explanatory models produces outputs from the same distribution as the training set. In contrast, predictive models are not limited to the same distribution as that of the training data and can yield input-output mappings for data outside the training distribution. Therefore, with both types of models having a function, one can not draw the mechanistic model conclusion.

It goes without saying that a predictive model might also be mechanistic. An example includes the selective tuning model of attention [72] predicting new observations that had not yet been tested and for which no data was available at the time its proposal, yet explaining the causal effects. It is in fact the ultimate goal to have a model that has the best of both worlds: to explain and to predict. The message, however, remains the same: availability of a function for a model does not imply explanatory or predictive modeling.

A mechanistic model of the brain, as defined above, reveals the various components that contribute to the stimulus-response transformation. CNNs, however, are still black-boxes that lack revealing such causal relationships. In recent years, a group of approaches have attempted to explain black-box CNNs by evaluating the network components with hope to better understand these models. These approaches will be referred to as CNN-explanatory models in what remains in this chapter. The common strategy to provide explanations for CNN predictions has been to solve an optimization problem and find regions in the input image that most influence the network output [73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 3]. Two issues are notable here:

1. When a prediction does not match the attributed region in the input, is it an error in the CNN or the CNN-explanatory method? In other words, in case of mismatch between the attribution and the output, it is difficult to distinguish the point of failure between the CNN and the CNN-explanatory model. In effect, it is unclear what the "correct" explanation is.

Figure 1.1: Examples of highlighted image regions considered as explanations for certain CNN predictions. Each region within the red curve shows a subset of image as the region influencing the CNN prediction. Each sub-region contains a wealth of information and it is not clear which piece is most affecting the prediction (adapted from [3]).

2. Often, CNN-explanatory approaches select a region in the input image as the attribution most affecting the CNN prediction. Consider, for example, the images with highlighted regions in Figure 1.1. In each row, parts of the input image are identified with red boundaries for a single class predictions. As an example, in the second row, the face of a cat is selected. Looking at the selected regions, the question is what feature in these regions influenced the prediction? Was it the existence of light-colored pixels or it was the configuration of the eyes, nose and the mouth? Or, perhaps it was the fur texture or presence of whiskers? Maybe it is the combination of all these features? These questions make it clear that simply selecting regions is not enough for understanding what CNNs see in an image and in short, selecting regions in the input image does not provide a clear explanation of which set of features in the input *caused* the observed output.

Perhaps the set of studies on understanding CNNs from Torralba's lab [88, 89, 90, 91] could be set apart from other approaches. In these studies, Torralba and colleagues take an approach similar to how neuroscientists study the brain by directly studying activations of individual units in the network to various concepts. They introduced four categories of concepts: objects, parts, materials and colors. Then, they segment the activation map of each unit given images in each concept

category and measure the intersection over union between the segment and the segmented region corresponding to the concept in the input image. This approach has the advantage of avoiding solving an optimization problem that could otherwise inject its error in the explanation. However, one cannot help noticing that first, the stimuli employed in these studies are natural images to which the same argument from above applies. Specifically, with natural images, it is not clear to which visual feature network units respond. Second, the set of concepts employed in these studies are quite high-level concepts, with the exception of color. Undoubtedly, it is reassuring to observe emergence of objects and parts in higher layers of CNNs combined with the observed edge and border selectivity in the first layer of CNNs. However, that is as far as our conclusion about CNNs can go, leaving us with close to no explanations for the intermediate layers and the components that contribute to the input-output transformation.

As a summary, CNNs were initially inspired by the brain, but with the goal of prediction. Despite reports of accurate predictions of neural data, CNNs were shown to have fundamental processing differences with the human brain. Whether these discrepancies between CNNs and the brain are due to the lack of certain mechanisms in CNNs or other factors, such as learning mechanisms and patterns that have no biological correlates, remains as a quest in both computer vision and neuroscience communities. Indeed the hierarchical architecture of these networks provides a mapping from the stimulus to neural responses. However, having a function in hand does not make a model a "mechanistic" one. A mechanistic model is transparent to reveal the causal relationship between the stimulus and the observed responses, even if the revealed function itself is not easy to understand. Given the current state of our understanding of CNNs, it is still not easy to infer a clear causal relationship between input and output in these networks and by the same token these hierarchical models do not provide a clear explanation of how the brain achieves the observed responses. Therefore, in the present work, we refrain from including these networks in the set of biologically-inspired models with the properties defined earlier.

1.3 Contributions and Significance

In this dissertation, analytical hierarchical neural networks that adhere to the discoveries of shape and color processing mechanisms in the primate visual systems are introduced. With the said

constraint, the proposed models enjoy the benefit of representations that are based on a working intelligent system geared toward visual tasks. Additionally, meeting the existing knowledge of the biological visual system, the choice of architecture becomes irrelevant; we simply follow the steps of nature. To address the second objective for this work, a simple learning algorithm is combined with an analytical neural network modeling V4. With experimental data from V4 available, the learned model reveals shape processing mechanisms in this visual area.

The present work makes the following contributions:

- **The Hue Network:** Introduces the first hierarchical hue processing model that explicitly models hue processing in LGN, V1, V2, and V4 and reveals the contributions of each processing layer in the output hue encoding.

- **The RBO Network:** Presents a novel biologically-plausible model for figure border ownership assignment. The RBO network is the first to combine global and local context for border ownership assignment and lends detailed insight to the role each of the global and local context play in a quantified way.

- **The SparseShape Network:** Introduces the first biologically-inspired model that explicitly represents an object-centered shape encoding similar to those reported in V4. SparseShape suggests an approach for combining an analytically-defined network for shape processing with simple learning methodologies to recover sophisticated and long-range interactions between shape parts within V4 receptive fields.

The proposed models benefit from the following properties:

- **Direct understanding of model features:** By design, all three models provide clear explanations for the final representations as well as the features in the intermediate layers of the network.

- **Effortless verification of contributing factors:** With analytical models, the factors contributing to a specific representation can be easily tested. For example, in the hue network, the contributions from single-opponent and multiplicative model V2 neurons to model V4 cells were evaluated with a simple approach. As another example, in the RBO network, eval-

uating the effects of global versus local context did not require any complicated attempts; simply the strength of the signal before and after each step was compared.

- **Explicit explanation of learned features:** When combined with learning approaches, due to the easy interpretation of features, the learned mechanisms are easy to understand and explain. For example, in SparseShape, the recovered receptive fields of model V4 cells show which convex/concave shape parts affect responses of these cells.

In contrast to black-box networks, the proposed models with these properties provide a transparent picture of the implemented system, its various components, interactions among its components and their role in the abstraction of representations. Moreover, these models suggest a mechanistic model of how the different brain elements work together to achieve the reported observations for the studied visual area.

Since the analytical neural network introduced in this dissertation are constrained to match the known properties of the human and primate visual systems, the next section is dedicated to a brief overview of the visual cortex architecture. Details of functional properties in each visual area employed in and relevant to the proposed models will be discussed in the individual chapters.

1.4 Architecture of the Visual Cortex

When light hits neurons in the retina, visual information in the form of neuron activation signals pass through the optic nerve to the lateral geniculate nucleus (LGN), and later to the visual cortex. Several areas involved in visual processing are identified in the visual cortex [92] with connections between these areas. Figure 1.2 depicts the visual hierarchy areas and their connections. In this figure, feedforward and feedback connections between the various areas are identified. Missing in this figure are connections in local neighborhoods inside each area; even within a single area and among neighboring neurons that are laterally connected, visual information can be exchanged.

The visual hierarchy in Figure 1.2 looks complicated with many intra-area connections. A simplified version of a subset of those areas involved in early vision are shown on the sketch of a brain in Figure 1.3. This figure illustrates two main streams of visual areas: the ventral and dorsal pathways shown in magenta and blue colors respectively. The ventral stream is believed to be involved

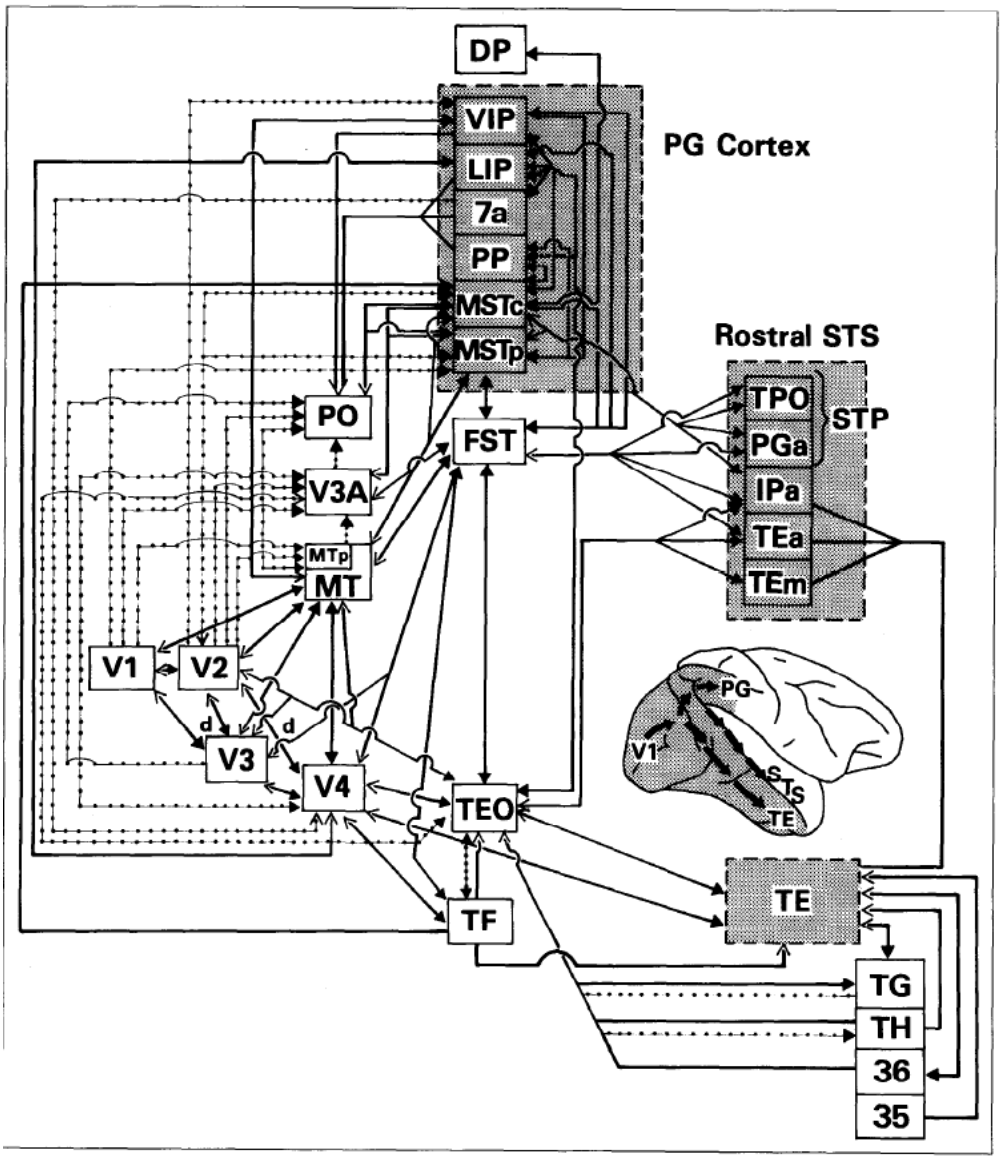


Figure 1.2: Visual hierarchy areas and their connections. This figure depicts several brain areas with feedforward (solid arrowheads) and feedback (open arrowheads) connections. Figure adapted from [4].

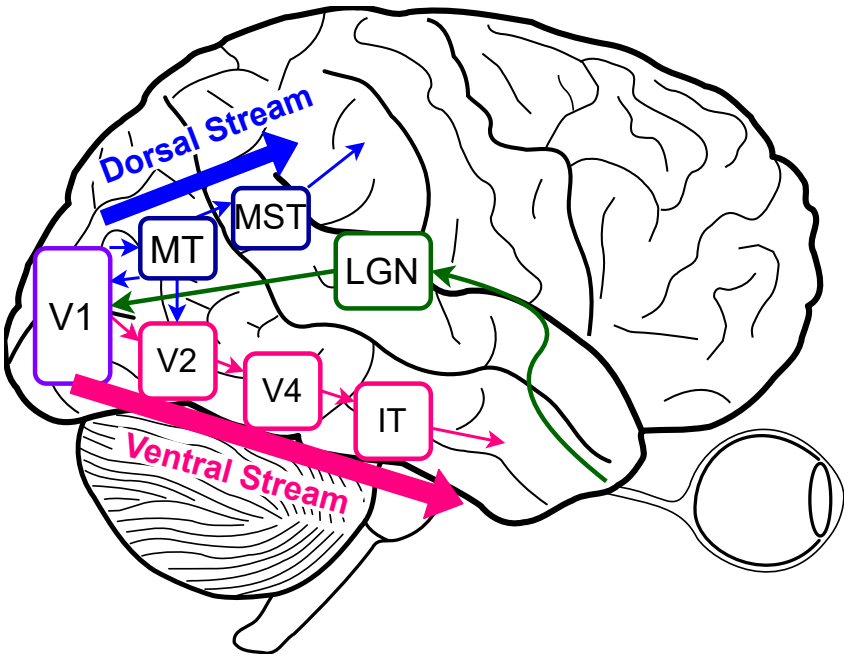


Figure 1.3: The visual processing network in humans and primates. Only the brain areas involved in the models studied in this dissertation are shown. The visual signal from the retina is relayed to the LGN and then the primary visual cortex. The brain areas shown in this figure have complicated connection and interactions with other areas shown here and those that are not. For simplicity, only the connections employed in the presented models are depicted. Figure modified from www.pngkey.com.

with recognition and discrimination of objects. Shape and color information processing, two visual representations studied in this dissertation, are performed in the ventral stream. The dorsal stream carries the signal related to spatial information and movement. Despite the dichotomy of the two pathways in terms of information processing, there is an increasing body of research in recent years providing evidence for shape selectivity to static stimuli in the dorsal stream [93, 94, 95, 17]. Such evidence causes researchers to question previous beliefs of segregated pathways and streams, each specialized for a specific processing type. Regardless of visual processing segregation into pathways, the visual hierarchy can be thought of a hierarchical network in which visual representations become more and more abstract as the signal moves higher in both streams. For example, in the ventral stream, V1 cells encode elongated bars and borders. After V1, in V2, angles and junctions within a contour are encoded by combining V1 inputs [96]. Higher in IT, neurons selective to objects and faces have been discovered [97, 98].

In addition to more abstract encodings, the receptive field (RF) sizes nearly double from one

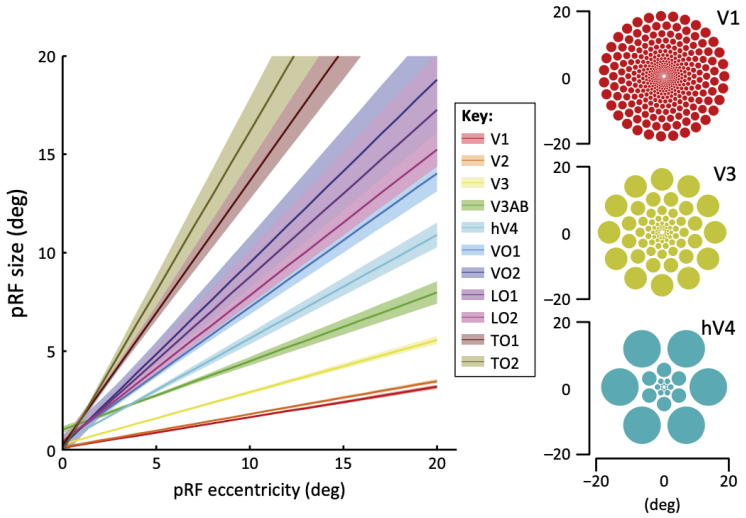


Figure 1.4: Human population receptive filed measured with fMRI. (left) Population RF plotted as a function of eccentricity demonstrates increase in receptive field sizes with eccentricity and in the hierarchy. (right) The spatial array of receptive fields in V1, V3, and hV4 shown as circles. Figure adapted from [5].

area to the next in the ventral stream [99]. Figure 1.4 includes diagrams of RF sizes in V1, V3, and V4 depicting the increase in RF sizes in the ventral stream. The center of each circular diagram in this figure represents the fovea. This figure also shows that the receptive field sizes within an area increase with eccentricity.

Neurons in the dorsal stream have larger RFs than those in similar hierarchical level in the ventral stream. For example, whereas MT receives its feedforward signal from V1, just like V2 in the ventral stream, the RF size of its neurons are comparable to those of V4. In addition to larger RFs, neurons in the dorsal stream are fast. It was in fact due to rapid responses that Bullier [15] called certain areas in the dorsal stream the "fast brain" and hypothesized that the asynchronous information transfer in the dorsal stream, compared to the ventral stream, is to "generate a first-pass analysis of the visual scene". Bullier suggested that the dorsal stream interjects rapid global information to V1 and V2 in the ventral stream "that act as an active blackboard". Later studies put forth evidence in favor of Bullier's fast brain and dorsal-to-ventral feedback hypothesis [100, 101, 102] and it is in this sense that MT is included in two of the models that will be discussed later in this dissertation.

Models in the present work are focused on early visual processing in V1 and V2 and mid-level representations of V4. In all the remaining chapters, part of the visual hierarchy with brain

areas modeled is shown as a network with each brain area represented as a layer of the network. Additionally, any brain area or cell type preceded with "m" refers to the corresponding model layer or neuron in the chapter. For example, mV1 cells refers to model V1 neurons. Biological cells are referred to without the prefix.

Each model introduced in this dissertation focuses on a certain processing aspect with neurons relevant to the studied representation in the visual cortex. Although each modeled visual area consists of layers of processing, I refrained from further extension of modeling to the granular level of layers within the included visual areas. Instead, I modeled the reported observations at a certain level of abstraction for each visual area to focus on the progression of the visual signal in a hierarchical network that can be tested against biological cells. In doing so, we can make progress and perhaps further predictions that can be tested to better enhance our understanding of processing in the brain. Previous computational models against which the proposed hierarchical networks are compared follow a similar level of abstraction and focus on modeling neurons in select visual areas rather than developing models for the formation of the signal in the various layers within a visual area.

1.5 Dissertation Organization

The remainder of this dissertation is organized as follows: the hue network is outlined in Chapter 2 followed by Chapter 3 providing an introduction to the Recurrent Border Ownership (RBO) network. In Chapter 4, the SparseShape hierarchy is explained in addition to how an analytical-defined model combined with a learning step can recover shape responses. Finally, Chapter 5 concludes the dissertation with some final remarks and future directions.

Chapter 2

Hue Processing: A Hierarchical Model

The work in this chapter has been published previously as the following:

Paria Mehrani, Andrei Mouraviev, and John K. Tsotsos, "Multiplicative modulations enhance diversity of hue-selective cells.", in *Scientific Reports*, 10, 8491 [2020].

and has been extended here for the purposes of this document. Andrei Mouraviev's contribution to this project was limited to developing the theoretical formalism and implementation of two types of hue-selective neurons in each of mLGN, mV1 and mV2 layers in addition to the six neuron types in mV4. Specifically, Andrei contributed to modeling and implementing single-opponent L-on and S-on neurons in each of mLGN, mV1 and mV2 layers in addition to the six neuron types in mV4. I was the sole contributor to adding single-opponent M-on, M-off, L-off and S-off cells in mLGN, mV1 and mV2 layers. Additionally, proposing and developing multiplicative neurons in the mV2 layer (Equation 2.5) were entirely carried out by me. Determining weights from mV2 layer to mV4 neurons (Equation 2.7) were also developed and implemented by me. Andrei contributed to producing part of the architecture and results depicted in Figure 2.3. I performed further theoretical development, implementation and experimental evaluation of the model to its final published state. Finally, I wrote the published manuscript.

Below is a summary of additions compared to the published version:

1. *The literature review is extended with additional details.*
2. *Two sets of experiments including Image Classification and Color Image Segmentation in Sections 2.4.9, 2.4.10 are added that were not originally part of the published manuscript.*

2.1 Motivation

Color is a contributing feature in many vision-related tasks. For example, color, and not shape, can help in discriminating between a raw banana and a ripe one, or choosing between white or brown eggs (See Figure 2.1 for examples). In this sense, there is a strong theoretical motivation for encoding color in computer vision systems. In this chapter, a novel hierarchical hue processing network is proposed that like the rest of the models in this document are informed by the human/primate visual system. More specifically, whereas color in the general sense, and depending on the space in which it is represented, can be comprised of different components, here, the focus is on perceptual hue encoding. As a result, in what follows, whenever the term "color" is used, it solely refers to the hue component. Moreover, it is imperative to note that the proposed hierarchy models hue-selective neurons with no orientation selectivity, resulting in no processing of color borders and form in this network.

2.2 Background

2.2.1 Hue Representation in the Ventral Stream

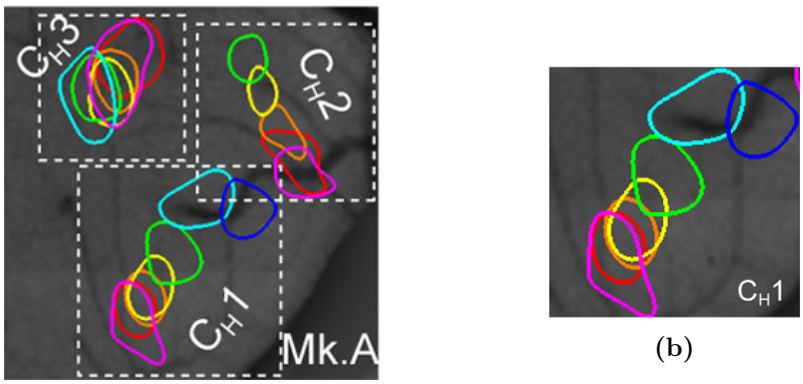
The color processing mechanisms in the primary visual cortex and later processing stages are a target of debate among color vision researchers. What is also unclear is which brain area represents unique hues, those pure colors unmixed with other colors. In spite of a lack of agreement on color mechanisms in higher visual areas, human visual system studies confirm that color encoding starts with three types of cones sensitive to long (L), medium (M) and short (S) wavelengths, forming the LMS color space. Cones send feedforward signals to LGN cells with single-opponent receptive fields [103, 104, 105]. Cone-opponent mechanisms such as those in LGN were the basis of the "Opponent Process Theory" of Hering [106], where he introduced unique hues. The four unique hues, red, green, yellow and blue were believed to be encoded by cone-opponent processes. Later studies [107, 108, 109], however, confirmed that the cone-opponent mechanisms of earlier processing stages do not correspond to Hering's red vs. green and yellow vs. blue opponent processes. In fact, they observed that the color coding in early stages is organized along the two dimensions of the MacLeod and Boynton (MB) [110] diagram, where abscissa and ordinate represent L vs. M and S

Figure 2.1: Color is a contributing feature in visual tasks. Where all the bananas are of the same shape, their color help in distinguish between raw and ripe bananas.

vs. LM activations respectively.

Beyond LGN, studies on multiple regions in the ventral stream show an increase in nonlinearity in terms of cone responses from LGN to higher brain areas [111] and a shift of selectivity toward intermediate hues in the MB diagram with more diverse selectivity in later processing stages [112]. Specifically, some suggested that neurons in V1 encoding local hue have single-opponent receptive fields similar to LGN cells [113, 114]. Lennie *et al.* [115] observed similar chromatic selectivity as LGN neurons in V1 cells and suggested a rectified sum of the three cone types describing the responses of V1 neurons. Hanazawa *et al.* [111] found both linear and nonlinear V1 and V2 responses with respect to the three cone types, with gradual increase in percentage of nonlinearity from LGN to higher layers. A similar finding by Kuriki *et al.* [112] suggested that there are selectivities to both intermediate and cone-opponent hues in areas V1, V2, V3 and V4, with more diverse selectivity in V4 compared to the other areas. De Valois *et al.* [108] suggested perceptual hues can be encoded in V1, obtained by combining LGN responses in a nonlinear fashion. Wachtler *et al.* [116] found that the tunings of V1 neurons are different from LGN, and that the responses in V1 are affected by context as well as remote color patches in both suppressive and enhancement manners.

Although Namima *et al.* [117] found neurons in V4, AIT and PIT to be luminance-dependent, others reported that in the Macaque extrastriate cortex, millimeter-sized neuron modules called globs, have luminance-invariant color tunings [118]. Within globs in V4, clusters of hue-selective patches with sequential representation following the color order in the hue, saturation and lightness (HSL) space were identified, which were called "rainbows of patches" [6]. An example of an identified map of three clusters is shown in Figure 2.2a. In Figure 2.2b a larger view of one of the clusters is shown. Conway and Tsao [119] suggested that cells in a glob are clustered by color preference and form the hypothesized color columns of Barlow [120]. Comparable findings in V2 were reported



(a) (b)

Figure 2.2: Color map of V4 neurons in three clusters of patches (adapted from [6]). (a) Combined map of hue patches in three clusters, (b) Larger view of a cluster.

in [121]. Patches in each cluster have the same visual field location and overlap greatly in their visual field with their neighboring patches [6, 118]. Moreover, Li *et al.* [6] observed that each color activates 1-4 overlapping patches and neighboring patches are activated for similar hues. Consequently, Li *et al.* [6] suggested that different multi-patch patterns represent different hues, and such a combinatorial color representation could encode the large space of physical colors, given the limited number of neurons in each cortical color map. Other studies also suggested that glob populations uniformly represent the color space [122] with narrow tunings for glob cells [122, 123, 124]. Similar findings for IT cells were reported by Zaidi *et al.* [125].

Not only is there disagreement about the color processing mechanisms in the visual cortex, but also which region in the brain represents unique hues. Furthermore, transformation mechanisms from cone-opponent responses to unique hues are unclear. While unique red is found to be close to the +L axis in the MB diagram, unique green, yellow and blue hues cluster around intermediate directions [126], not along cone-opponent axes. Perhaps clustering of the majority of unique hues along intermediate directions could describe the suggestion by Wuerger *et al.* [127] who proposed that the encoding of unique hues, unlike the tuning of LGN neurons, needs higher order mechanisms such as a piecewise linear model in terms of cone inputs. The possibility of unique hue representations in V1 and V2 was rejected in [128], but like others [124, 129], they observed neurons in PIT show selectivities to all hue angles on the color wheel and that there are more neurons selective to those close to unique hues. These findings were challenged in [130]. Similarly, Zaidi *et al.* [131] observed no significance of unique hues in human subjects and responses of IT neurons, which was

confirmed in more recent studies [132, 133].

2.2.2 Computational Models for Hue Encoding

Color is a key contributing representation for visual tasks; hence, the many various color spaces that are introduced to represent color. Some attempts in quantifying color are specific to human color vision. In these approaches the goal is to model perceptual hues. However, a recent study [132] warns against assuming the geometry of certain color spaces as accurate estimates of the perceptual color space. In that case, one wonders if the imperfect estimation could explain the conclusion of the study by Rasouli *et al.* [134] who found no optimal color space that would be suitable for detection and discrimination of all color group.

In CNNs, Bau *et al.* [91] found neurons responding to color concepts. In fact, color was the only concept that persistently appeared in all the convolutional layers. Similar observations were reported in other work [135, 136]. Rafegas *et al.* [136] mentioned similarities between color processing in CNNs and the primate brain. Here, we refrain from repeating the issues with such suggestions in favor of all that was mentioned in detail in Chapter 1.

Among all the biologically-inspired approaches that attempt to understand the neural processes for transformation from cone-opponent to perceptual colors, a number of computational models suggested mechanisms for this problem and other aspects of color representation in higher areas [137, 138, 139, 140, 141]. For example, the role of context in color constancy and color induction in V4 was studied by Dufort and Lumsden [137]. Their model combines L+M- and S+(L+M)- LGN neurons to achieve color constancy. Later, Courtney *et al.* [138] achieved this goal in V4 by means of a "silent" surround [123] for providing context, while [139] introduced a model of feedforward, feedback and lateral connections for this purpose.

The most relevant computational model to ours with focus on hue encoding would be the work of De Valois *et al.* [140], in which they introduced a network consisting of three stages: cone stage, cone-opponent stage, and perceptual color representation stage. Introduction of the third stage was inspired by the noted discrepancy between the axes of cone-opponency and perceptual hues. They proposed that S-modulations in the third stage rotate the cone-opponent axes in a way that results in perceptual hue representation. The third stage responses are obtained by a linear combination of cone-opponent signals, followed by a half-wave rectification. De Valois *et al.*

did not explicitly mention to which of the visual processing layers in the brain this third stage corresponds. Here, we examine two possibilities. For the first possibility, suppose the perceptual hue stage corresponds to V1, which is downstream of LGN. This is not an invalid assumption as this stage combines the linearly modeled cone-opponent responses. In this case, perceptual hues are encoded in V1, as found in a later study by the authors [108], in contrast with other findings for this visual area [142, 143, 7, 115]. A similar argument applies to assuming correspondence with V2 [127]. For the second possibility, suppose their third stage is modeling the neurons in V4. In that case, they provide no explicit model of V1 or V2 processes, even though these levels are believed to play the role of gradually increasing the nonlinearity in the processing of cone inputs [111]. Omitting the nonlinear layers of V1 and V2 results in wide tunings, an outcome which is not supported by previous and recent discoveries [122, 124].

Overall, all the mentioned computational models are one-layer formulations of perceptual hue encoding, or simply put, the totality of processing in these models is compressed into a single layer process. The end result may indeed provide a suitable model in the sense of its input-output characterization. However, it does not make an explicit statement about what each of the processing areas are contributing to the overall result and while their goal was to reveal biological color processing, they do not shed light upon the mystery of color representation mechanisms in the brain.

2.3 The Hue Network Model

In this chapter, we introduce a computational color processing model inspired by neural mechanisms in the visual system, that explicitly models neurons in each of LGN, V1, V2, and V4 areas and reveals how each visual cortical area participates in the process. In this model, nonlinearity is gradually increased in the hierarchy as observed by [111]. Specifically, while a half-wave rectifier unit keeps the V1 tunings similar to those of LGN [115], it makes V1 neurons nonlinear in terms of cone inputs. In V2, besides single-opponent cells, we propose employing neurons with multiplicative modulations [144, 145, 146], which not only increase nonlinearity but also allow neuronal interactions by mixing the color channels and narrows the tunings. De Valois *et al.* [140] suggested that additive or subtractive modulation of cone-opponent cells with S-opponent cell responses ro-

tates the cone-opponent axes to red-green and blue-yellow directions. We achieved this rotation with multiplicative modulations of L- and M-opponent cell activations with S-opponent neuron responses. We call these cells "multiplicative V2" neurons. Finally, V4 responses are computed by linearly combining V2 activations with weights determined according to tuning peak distances of V2 cells to the desired V4 neuron tuning peak.

Figure 2.3a depicts our proposed model and Figure 2.3b demonstrates our network in action. Each layer of this model implements neurons in a single brain area. Each map within a layer consists of neurons of a single type with receptive fields spanning the visual field of the model; for example, a map of neurons selective to red hue in model layer V4. The leftmost layer in this figure shows the input to the network with the LMS cone activations. In the event that the presented stimulus was available in other color spaces, we converted those to standard RGB followed by a conversion into LMS channels using the transformation algorithm proposed by [147] (we used the C code provided by the authors). This algorithm converts RGB values using the phosphor emission function of a CRT monitor and then multiplies those values with the sensitivity functions of L, M and S cones. In this sense, we are assuming that the proposed hue network, similar to the experimental settings with monkeys [6], sees the stimuli displayed on an imaginary CRT monitor with emission functions specified in [147] and their accompanying code. As a result, one can think of the presented stimulus to the network as the activations of three cone types. These cone activations are fed to single-opponent mLGN cells, which in turn feed single-opponent mV1 cells with nonlinear rectification. In the mV2 layer, single-opponent neurons replicate the activations of those of mV1, but with larger receptive fields. Later, we refer to these single-opponent cells as "additive mV2 neurons". "Multiplicative mV2" neurons form when single-opponent mV1 cells with L and M cone inputs are multiplicatively modulated by mV1 neurons with S-cone input. This approach is in a sense similar to S-modulations proposed by De Valois *et al.* [140], but in a multiplicative manner and not additive or subtractive. Finally, the hue-sensitive neurons in mV4 receive feedforward signal from additive and multiplicative mV2 cells. Figure 2.4 depicts the computational steps of our model in each layer.

Our model was implemented in TarzaNN [148]. The neurons in all layers are linearly rectified.

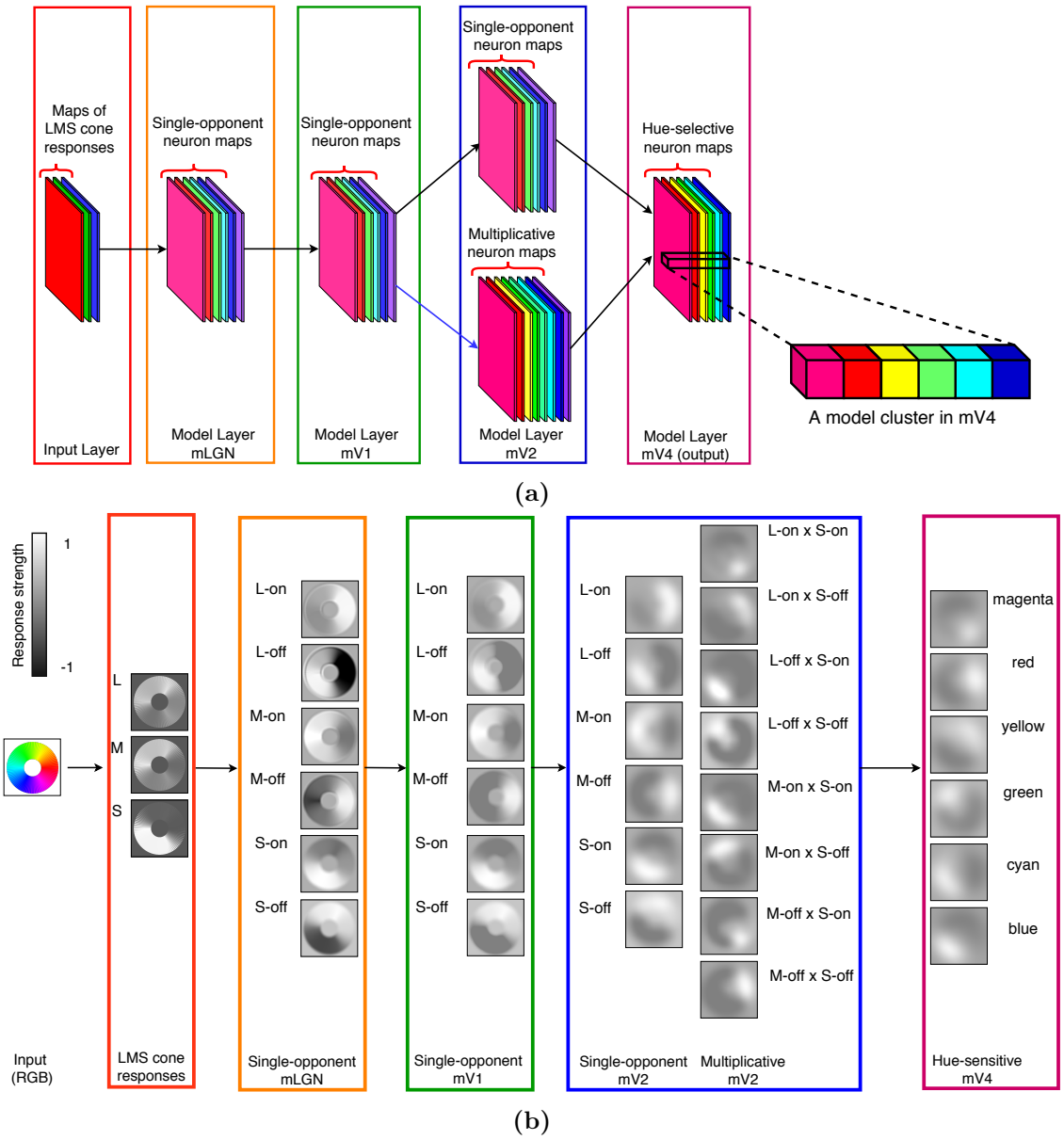


Figure 2.3: The proposed hierarchical model. (a) Our model architecture for local hue representation (best seen in color). Each layer of this model implements neurons in a single brain area. The prefix "m" in each layer name represents model layers. Each map within a layer consists of neurons of a single type with receptive fields spanning the visual field of the model. The leftmost layer shows the input to the network with the LMS cone activations. The color employed for each feature map in the model is figurative and not a true representation of the hue-selectivity of its comprising neurons. In layer mV4, a larger view of an example model cluster is shown, similar to the clusters found in monkey V4 [6]. Each model cluster corresponds to a column of the three-dimensional matrix obtained by stacking mV4 maps. (b) An example of our model responses. We show each layer of the network within a rectangle, with a number of feature maps inside the rectangle. The selectivity of neurons is written next to each map. The receptive field of each neuron in these maps is centered at the corresponding pixel location. The neuron responses are shown in greyscale, with a minimum response as black, and maximum activation as white. The range of responses shown on the upper left part of (b) represents the minimum and maximum firing rate of model neurons. The dark border around each feature map is shown only for the purpose of this figure and is not part of the activity map.

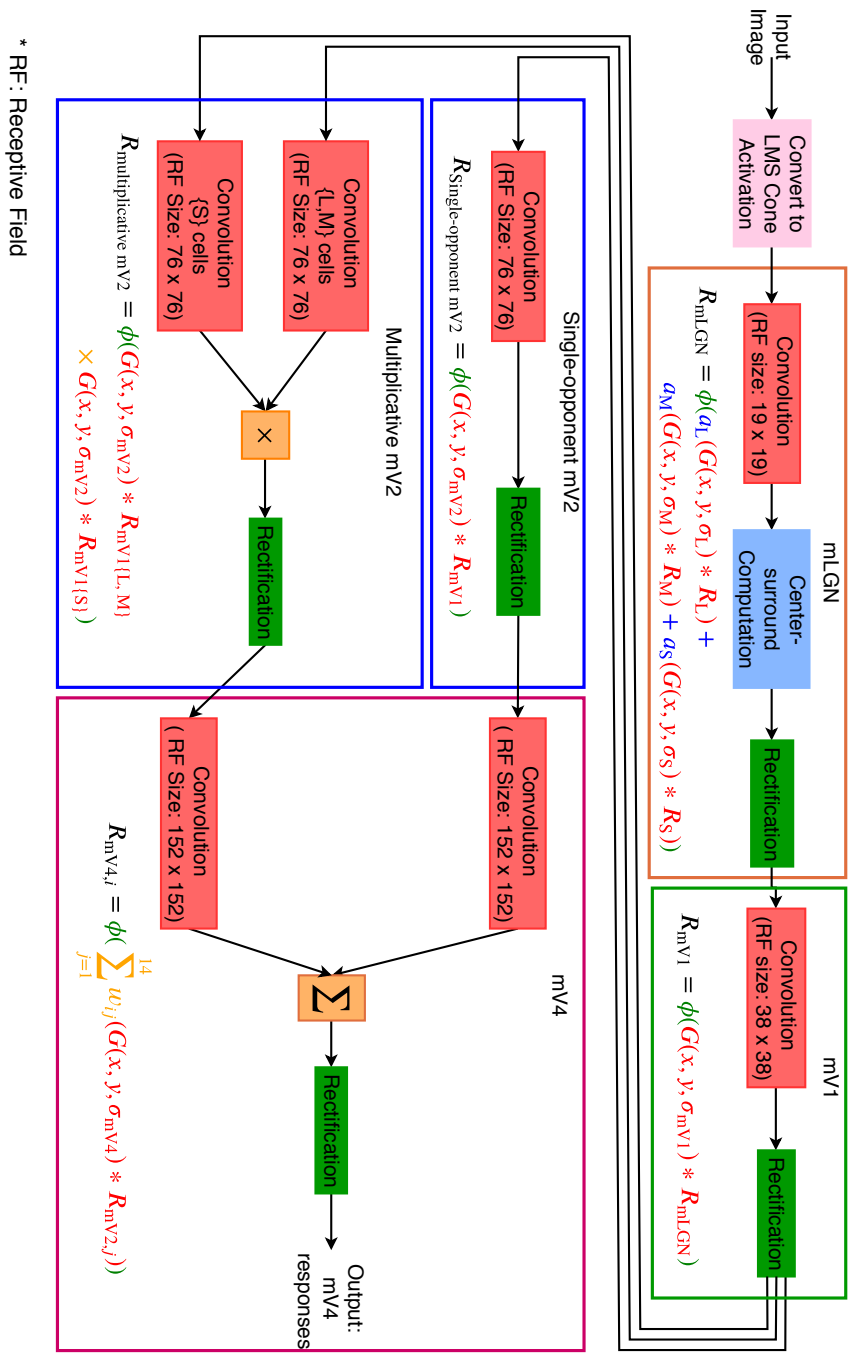


Figure 2.4: Hue processing network with details of computational blocks. Each layer is enclosed in a colored rectangle similar to those shown in Figure 2.3. Each little box in this figure represents one computational step in the model. The mathematical equation for each layer is included within the layer's rectangle. The font color of each component in the equation matches the color of its corresponding computational box in the layer. For example, the font color of the rectification function ϕ matches the green color of the rectification box in mLGN, or the yellow color for $\times$ in the multiplicative mV2 equation matches the yellow color of the multiplication box in this layer. Note the arguments are delimited by the same-color brackets. In convolution boxes, RF stands for receptive field.

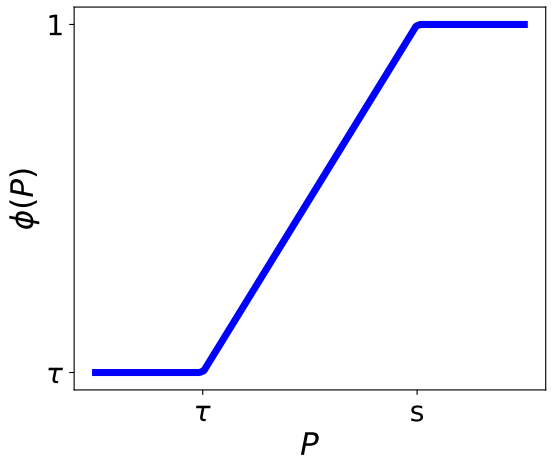


Figure 2.5: The rectification function, ϕ, applied to neurons in the proposed hue network (Equation 2.1). This rectifier maps neuron activations, P, to $[\tau, 1]$.

The rectification was performed using:

$$\phi(P) = \begin{cases} \tau, & \text{if } mP + b < \tau \\ mP + b, & \text{if } \tau \leq mP + b \leq s \\ 1, & \text{otherwise,} \end{cases} \tag{2.1}$$

where P is neuron activity, and m and b are the slope and base spike rate respectively, τ is a lower threshold of activities and s represents the saturation threshold. As depicted in Figure 2.5, this rectifier maps responses to $[\tau, 1]$. Depending on the settings of parameters τ and s, and the range of activations for the model neurons, the rectifier might vary from being linear to nonlinear. Wherever this rectifier is employed in the rest of this chapter, we mention the settings of the parameters, and whether parameter settings resulted in neuron activations to become linear or nonlinear in terms of their input.

2.3.1 Model LGN cells

The first layer of the hierarchy models single-opponent LGN cells. The mLGN cells are characterized by their opponent inputs from cones. For example, mLGN cells receiving excitatory input from L cones to their centers and inhibitory signals from L and M cones from the surround are known as

L-on cells. Model mLGN cell responses were computed by [149]:

$$\begin{aligned} R_{\mathrm{mLGN}} \quad = \phi(\quad & a_{\mathrm{L}}(G(x,y,\sigma_{\mathrm{L}}) * R_{\mathrm{L}}) + \\ & a_{\mathrm{M}}(G(x,y,\sigma_{\mathrm{M}}) * R_{\mathrm{M}}) + \\ & a_{\mathrm{S}}(G(x,y,\sigma_{\mathrm{S}}) * R_{\mathrm{S}}) - \\ & a'_{\mathrm{L}}(G(x,y,\sigma'_{\mathrm{L}}) * R_{\mathrm{L}}) - \\ & a'_{\mathrm{M}}(G(x,y,\sigma'_{\mathrm{M}}) * R_{\mathrm{M}}) - \\ & a'_{\mathrm{S}}(G(x,y,\sigma'_{\mathrm{S}}) * R_{\mathrm{S}})), \end{aligned} \tag{2.2}$$

where $*$ represents convolution. In this equation, mLGN response, R_{mLGN}, is computed by first linearly combining cone activities R_{L}, R_{M}, and R_{S}, convolved with normalized Gaussian kernels, G, of different standard deviations, σ, followed by a linear rectification, ϕ. For mLGN neurons, we set $\tau = -1$ and $s = 1$ to ensure the responses of these neurons are linear combinations of the cone responses [108, 115]. The differences in standard deviations of the Gaussian kernels ensure different spatial extents for each cone as described in [103]. Each weight in Equation (2.2), determines presence/absence of the corresponding cone. The weights used for mLGN cells were set following the findings of [104] and [105]. For example, for an L-on neuron, L-cones have excitatory contributions to the center and inhibitory surround contributions are from both L- and M- cones. In this case, for the center, a_L is positive with a_M and a_S set to zero, while in the surround only a'_S is zero. In total, we modeled six different mLGN neuron types, L-on, L-off, M-on, M-off, S-on, and S-off. As an example, consider M-on cells. These neurons receive opposite contributions from L and M cones, while S cones with weight 0 exhibit no contribution. That is, M and L cones have excitatory and inhibitory effects respectively, while S cones are absent. This type of neuron is known to best respond to cyan-like hues [142]. A relatively similar hue selectivity is observed in L-off cells, which receive excitatory and inhibitory contributions from M and L cones respectively. Receiving such a pattern of input signal results in selectivities to cyan-like hues for both cell types. Despite this similarity, however, their responses indicate different messages. Specifically, strong responses of an M-on cell conveys high confidence in the existence of an M cone signal with less impact from L cone responses within its receptive field. In contrast, strong activations of an L-

off cell indicates the existence of M cone signal with a confident message that almost no L cone activities exist within its receptive field. In other words, even a small amount of L cone responses within the receptive field of an L-off cell strongly suppresses the activation of this neuron. Looking at the feature maps for these two cell types in Figure 2.3b reveals these slight differences in their selectivities. Note that while both neurons have relatively strong positive responses to the green and blue regions in the input, activations of the L-off cells, unlike the responses of M-on neurons, in the yellow region are strongly suppressed.

In what follows, whenever we refer to a cell as L, M, or S in layers mLGN and higher, we will be referring to the pair of on and off neurons in that layer. For instance, S-on and S-off neurons in mLGN will be called S neurons in this layer, for brevity.

2.3.2 Model V1 cells

Local hue in V1, as suggested in [114] and [113], can be encoded by single-opponent cells. To obtain such a representation in the mV1 layer, the responses are determined by convolving input signals with a Gaussian kernel. Note that since single-opponency is implemented in the mLGN layer, by simply convolving mLGN signals with a Gaussian kernel, we will also have single-opponency in mV1. The local hue responses of mV1 were obtained by:

$$R_{\text{mV1}} = \phi(G(x, y, \sigma_{\text{mV1}}) * R_{\text{mLGN}}), \tag{2.3}$$

where ϕ is the rectifier in Equation (2.1). With $\tau = 0$ and $s = 1$ for the rectifier, our mV1 neurons will be nonlinear functions of cone activations. In Equation (2.3), substituting R_{mLGN} with any of the six mLGN neuron type responses will result in a corresponding mV1 neuron type. Therefore, there are six neuron types in layer mV1 corresponding to L-on, L-off, M-on, M-off, S-on, and S-off.

2.3.3 Model V2 cells

In our network, the mV2 layer consists of two types of hue selective cells: single-opponent and multiplicative. The single-opponent neurons are obtained by:

$$R_{\text{additive mV2}} = \phi(G(x, y, \sigma_{\text{mV2}}) * R_{\text{mV1}}), \tag{2.4}$$

where ϕ is the rectifier in Equation (2.1). With $\tau = 0$ and $s = 1$ for the rectifier, the single-opponent mV2 cells are nonlinear functions of cone activations. In Equation (2.4), substituting R_{mV1} with each of the six mV1 neuron type responses will yield a mV2 neuron type with similar selectivities, but with a larger receptive field. To be more specific, the responses of single-opponent mV2 neurons can be considered as a linear combination of mV1 activations.

To increase the nonlinearity as a function of cone activations in mV2, as observed by Hanazawa *et al*. [111], and also to nudge the selectivities further toward intermediate hues, as found by Kuriki *et al*. [112], we introduce multiplicative mV2 neurons. These cells not only add another form of nonlinearity to the model, other than that obtained by the rectifier in mV1, but also mix the different color channels from mV1 and exhibit a decrease in their tuning bandwidths. In their model, De Valois *et al*. [140] suggested that S-modulated neurons rotate the cone-opponent axes to perceptual-opponent directions. Their modulations with S activations were in the form of additions and subtractions, which does not add to the nonlinearity of neuron responses. We leverage their observation, but in the form of multiplicative modulations for additional nonlinearity. That is, each mV2 multiplicative cell response is the result of multiplying L or M neurons from mV1 with a mV1 S cell activations. For example, in Figure 2.3b, "L-off × S-off" is for a cell obtained by modulating a mV1 L-off cell response by a mV1 S-off neuron activation. In our model, the multiplicative mV2 neurons are computed as:

$$\begin{aligned} R_{\text{multiplicative mV2}} &= \phi(G(x, y, \sigma_{\text{mV2}}) * R_{\text{mV1\{L, M\}}} \times \\ &\quad G(x, y, \sigma_{\text{mV2}}) * R_{\text{mV1\{S\}}}), \end{aligned} \tag{2.5}$$

where $\times$ represent multiplicative modulation, and $R_{\text{mV1\{L, M\}}}$ and $R_{\text{mV1\{S\}}}$ are for responses of an L or M cell and S neuron from mV1 respectively. As before, ϕ is the rectifier from Equation (2.1) with the same parameters as those of the additive mV2 cells. Multiplicative mV2 cells are nonlinear with respect to cone inputs and bilinear with regards to mV1 activations.

Multiplicative mV2 neurons have narrower bandwidths than those of additive mV2 cells, which will be shown quantitatively in Section 2.4. However, consider the multiplicative mV2 maps in Figure 2.3b for a brief qualitative explanation. For the hue wheel as input, relatively high responses of the single-opponent mV2 cells span a larger region of their map compared to multiplicative mV2

cells. This is an indication that multiplicative mV2 cells are selective to a narrower range of hue angles. As an example, both L-off and S-off mV1 cells have high activations for relatively large regions of the input respectively. However, when multiplied, the resulting neuron, *i.e.* the "L-off $\times$ S-off" cell has strong responses for regions with greenish color, and the activation of the L-off mV1 cell to bluish regions is suppressed to the extent that the L-off $\times$ S-off cell shows close to no responses.

As a summary, in layer mV2, a total of 14 neuron types are implemented: 6 additive and 8 multiplicative cell types.

2.3.4 Model V4 cells

We modelled mV4 neurons representing local hue using a weighted sum of convolutions over mV2 neuron outputs. More specifically, responses of the i-th mV4 neuron, $R_{\mathrm{mV4},i}$, are computed as:

$$R_{\mathrm{mV4},i} = \phi(\sum_{j=1}^{14} w_{ij}(G(x, y, \sigma_{\mathrm{mV4}}) * R_{\mathrm{mV2},j})), \tag{2.6}$$

where $R_{\mathrm{mV2},j}$ represents the responses of the j-th mV2 neuron, and ϕ is the rectifier introduced in Equation (2.1), with $\tau = 0$ and $s = 1$. As a result of this parameter setting for the rectifier, each mV4 cell is a linear combination of mV2 cell responses and hence, nonlinear in terms of cone inputs. The set of weights $\{w_{ij}\}_{j=1,\ldots,14}$ determine the hue to which the i-th model mV4 neuron shows selectivity.

In layer mV4, we implemented six different neuron types according to distinct hues: red, yellow, green, cyan, blue, and magenta. The chosen hues are 60 deg apart on the hue circle of HSL, with red at 0 deg. These hues were also employed in the V4 color map study [6] and for comparison purposes, we utilize these hues. When the six mV4 colors are mapped to the MB space, the hue angles are shifted with respect to those of HSL, with red, yellow and cyan hues close to cone-opponent axes in the MB space, and green, blue and magenta along intermediate directions. Figure 2.6 illustrates 60 uniformly sampled hues from the HSL space mapped to the MB space with a clear nonuniform distribution and a shift, which are caused by the mapping and not the sampling.

Although here we limit the number of modeled neuron types in this layer to six, we would like to emphasize that changes in combination weights will lead to neurons with various hue selectivities in

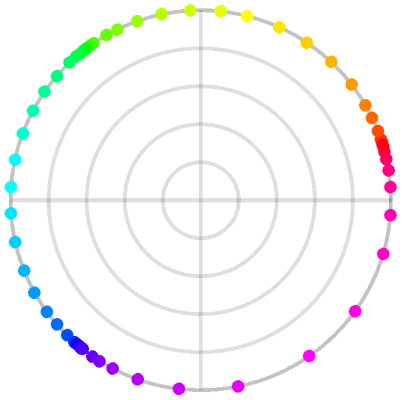


Figure 2.6: The 60 uniformly sampled hues from the HSL space mapped onto a unit circle in the MacLeod and Boynton space. As shown, the hues are not uniformly spaced on the unit circle. The hues are rotated on the unit circle, with different hue angles, compared to their corresponding angles on the hue wheel as a result of a non-uniform mapping. The four cardinal axes in this space correspond to reddish, lime, cyan-like and lavender hues.

this layer. Modeling neurons with selectivities to a wide variety of hues with yet narrower tunings could be accomplished in higher layers, such as IT, by combining hue-selective model neurons in mV4.

In order to determine the weights from mV2 to mV4 neurons, the w_{ij}'s in Equation (2.6), we considered the distance between peak activations of mV2 neurons to the desired hue in a mV4 cell. The hue angle between these two hues on the hue circle is represented by d_{ij}. Then, the weight w_{ij} from mV2 neuron j to mV4 neuron i is determined by:

$$w_{ij} = \frac{\mathcal{N}(d_{ij}; 0, \sigma)}{Z_i}, \tag{2.7}$$

where $\mathcal{N}(.; 0, \sigma)$ represents a normal distribution with 0 mean and σ standard deviation, and Z_i is a normalizing constant obtained by

$$Z_i = \sum_{j=1}^{14} \mathcal{N}(d_{ij}; 0, \sigma). \tag{2.8}$$

The weights used for each of mV4 neuron types are summarized in Figure 2.7. In this figure, each row represents the weights for a single mV4 cell, and the columns are for mV2 cells. Note that all mV2 to mV4 weights are normalized to sum to 1.0. That is, the sum of weights in each row is 1. In

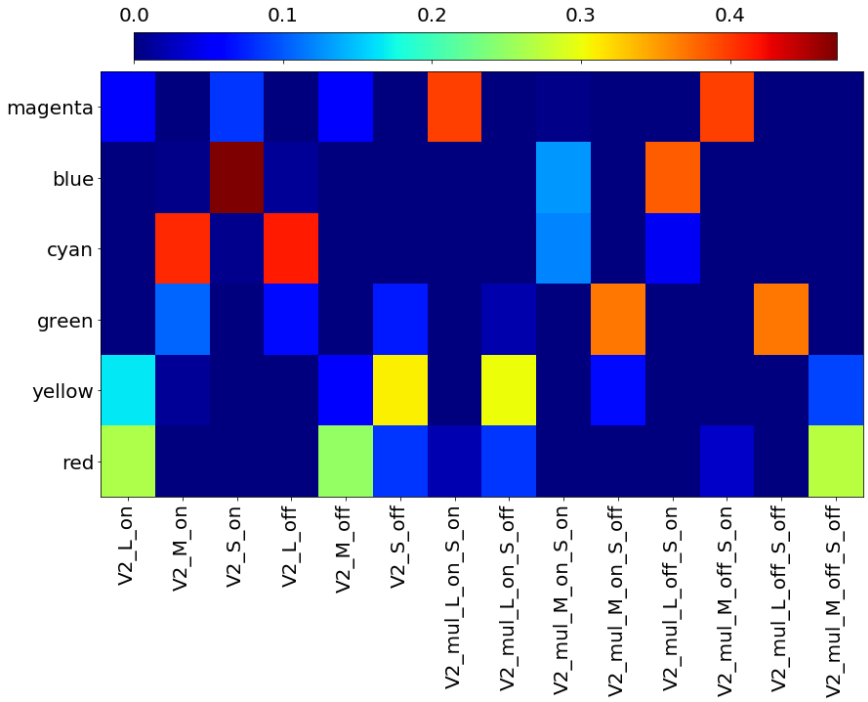


Figure 2.7: The relative contributions of individual mV2 cells in the encoding of six distinct hues in our mV4 layer (best seen in color). Each row represents a hue-sensitive mV4 cell, and each column shows a mV2 cell. Multiplicative mV2 cells are indicated as V2_mul_x_y, where x and y signify the type of mV1 neurons sending feedforward signal to the multiplicative mV2 cell. The weights are normalized to sum to 1.0. Determining weights from mV2 to mV4 layers in our network is described in detail in the Methods section.

this figure, dark red shows a large contribution, while dark blue represents close to no input from the relevant mV2 neuron. Consider, for example, the weights for the red mV4 cells. This neuron has relatively large weights from mV2 L-on, M-off, and M-off $\times$ S-off cells. In other words, cells with large contributions from L cones. This observation is not surprising as previous research by Webster *et al*. [126] found that unique red in human subjects has the largest contributions from L cones.

In Figure 2.3b, at the mV4 layer, from top to bottom, the neurons selective to magenta, red, yellow, green, cyan, and blue are displayed. As expected, mV4 yellow neurons, for instance, show activations across red, yellow, and green regions of the stimulus, with stronger activations in the yellow segment.

2.4 Evaluation

We designed simulation experiments to make two important aspects of our model clear:

1. at the single-cell level, our model neurons perform similarly to their biological counterparts.

2. at the system level, our hierarchical model, due to a gradual increase in nonlinearity, makes it possible to model neurons with narrow bandwidths, represent hues in intermediate directions, and represent unique hues in our layer mV2.

As a result, our experiments bridge a mixture of single-cell examinations to evaluations of the hierarchical model as a whole.

2.4.1 Stimuli

To test the effectiveness of our approach in hue encoding, we examined the tuning of each hue-selective neuron in all layers of our network. We sampled the hue dimension of the HSL space, with saturation and lightness values constant and set to 1 and 0.5, respectively, following [6]. We sampled 60 different hues in the range of $[0, 360)$ degrees, separated by 6 degrees. When these HSL hue angles are mapped to a unit circle in the MB space, they are not uniformly spaced on the circle and are rotated (depicted in Figure 2.6). For example, the red hue in the HSL space at 0 deg corresponds to the hue at about 18 deg in the MB space.

The input to the hierarchical network was always resized to 256×256 pixels. The receptive field sizes, following [99], double from one layer to the one above. Specifically, the receptive field sizes we employed were 19×19, 38×38, 76×76, and 152×152 pixels for mLGN, mV1, mV2, and mV4 layers respectively.

2.4.2 Model neuron tunings

We presented all sampled hues to the model and recorded the activities of model neurons with responses shown in Figures 2.8 and 2.9. Our single-opponent cells implement two types of center-surround receptive fields: X-on and X-off cells, where X represents one of the three cone types which dominates the input to the receptive field center and on/off signifies the excitatory/inhibitory contribution of that cone. The receptive field profiles of these neurons are borrowed from the studies by Wool *et al.* [104, 105].

In each plot, the circular dimension represents the hue angle in the MB diagram, and the radial dimension represents the neuron's response level. We found mLGN and mV1 tunings look relatively similar with differences due to the nonlinear rectifier imposed on mV1 responses. We plotted the responses of mV1 neurons in both negative and positive ranges for comparison purposes with those of mLGN. For mV2 and mV4, the responses are shown in the positive range. Although it might not be evident from tunings of mV1 cells and single-opponent mV2 neurons in Figure 2.8, due to the plotted range of responses in these figures, we emphasize that the tunings of these cells look identical when plotted in the same range, as the two examples in the bottom panel of Figure 2.8 show. The average difference of responses between pairs of mV1 and their corresponding single-opponent mV2 cells is on the order of 10^{-6}.

Comparing single-opponent and multiplicative mV2 tunings gives a clear image of narrower tunings in the S-modulated mV2 neurons. Not only do these cells have narrower tunings, but they generally peak close to intermediate directions. For instance, the mV2 L-off $\times$ S-off cell has a narrow tuning that peaks close to the unique green hue angle.

In mV4, we implemented six different neuron types according to distinct red, yellow, green, cyan, blue and magenta, which are 60 deg apart on the HSL hue circle. The weights from mV2 cells to mV4 neurons were determined according to the distance between peak activations of mV2 neurons to the desired hue in a mV4 cell. Tunings of mV4 neurons, depicted in Figure 2.9, show

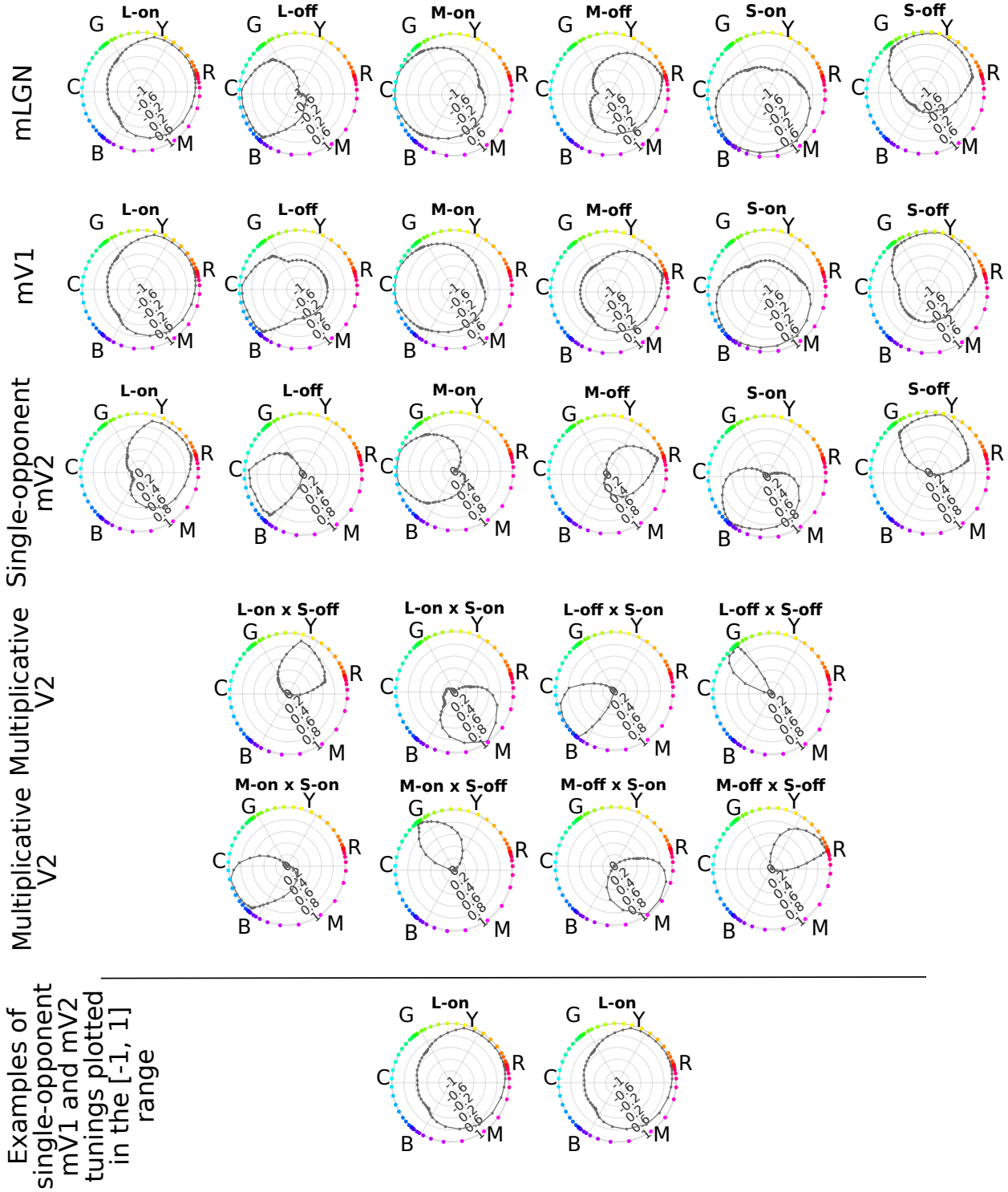


Figure 2.8: Caption on the next page.

Figure 2.8: (Previous Page). Model neuron tunings. Responses of neurons in layers mLGN, mV1, and mV2 to 60 hues sampled from the hue dimension in the HSL space. Each sampled hue is mapped to its corresponding hue angle in the MB space and is shown by a colored dot corresponding to the sampled hue on the circumference of a unit circle in the MB space. In each plot, the circular dimension represents the hue angle in the MB space. The level of responses is shown in the radial dimension in these plots with high activations away from the plot center. In each row, the model layer to which the neurons belong is specified on the left edge of the row. The neuron type is mentioned below each plot. In the bottom panel, we compare tunings of single-opponent mV1 and mV2 L-on neurons and show almost identical curves for single-opponent mV1 and mV2 cells when plotted in the same range of $[-1, 1]$.

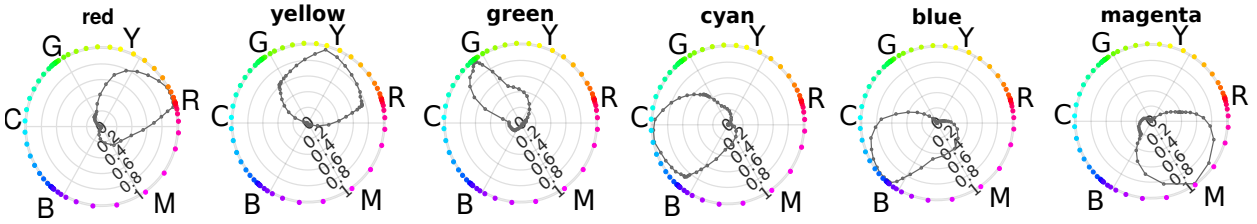


Figure 2.9: Model mV4 neuron responses to hues sampled from the hue dimension in the HSL space. The sampled hues are 6 degrees apart. In each plot, the angular dimension shows the hue angles in the MacLeod and Boyton diagram, and the radial dimension represents the response level of the neuron where larger responses are further away from the center. The dots on the circumference of each circle are representative of hues at sampled angles. The selectivity of each neuron is specified above its plot.

a distinct peak for each cell close to its desired selectivity, with narrower tunings compared to single-opponent mV2 cells.

2.4.3 Tuning bandwidths

Kiper *et al.* [7] defined tuning bandwidth as "the angular difference between the color vector giving the peak response, and that vector where the response had fallen to 50% of the difference between peak vector and 90 deg away". They reported tuning bandwidth as a metric for tuning narrowness and observed separate bandwidth distributions of linear and nonlinear V2 cells. To obtain a quantitative evaluation of our observation with regards to narrower tunings due to multiplicative modulations, we computed the tuning bandwidth of mV2 neurons, following Kiper *et al.* [7]. For neurons exhibiting maximum activation at a range of neighboring hues (see mV2 M-on tuning in Figure 2.8 as an example), we take the mean hue as the representative hue with peak response for computation of bandwidth and later for tuning peak. While our goal is to verify whether multiplicative modulations narrow the tunings, we did not hypothesize a model as Kiper *et al.* [7] who

computed the bandwidth threshold analytically for linear and nonlinear tunings, nor did we have a population of neurons to report the percentage of linear/nonlinear cells. Instead, we simply computed the bandwidth of responses for each cell type in the mV2 layer and plotted the distribution of tuning bandwidths. Specifically, we computed the tuning bandwidth of 6 single-opponent and 8 multiplicatively modulated mV2 neurons in our model. The bandwidth distributions are depicted in Figure 2.10a, where we also plotted the distributions of linear/nonlinear biological V2 cell population extracted from Kiper *et al*. [7], their Fig. 3. Interestingly, bandwidth distributions of mV2 neurons are separate and overlap with that of biological linear and nonlinear V2 cells reported by Kiper *et al*. [7]. Although the distribution of single-opponent mV2 neurons is slightly shifted toward smaller bandwidths compared to that of biological linear V2 cells, both distributions are spread over a wide range of bandwidths. While the biological nonlinear V2 distribution is squeezed compared to that of multiplicative mV2 neurons, both distributions peak at similar bandwidths. Note that the area under all curves are normalized, hence, the higher peak in the nonlinear distribution. Single-opponent mV2 neurons are mainly clustered around large bandwidths, between 36 and 66 deg (mean = 52.98 deg). Tuning bandwidths of multiplicative mV2 cells vary between 9 and 60 deg with an apparent density toward bandwidths smaller than 40 deg (mean = 30.17 deg). These results confirm our previous observation that multiplicative modulations lead to narrower tunings.

Another approach to construct neurons with narrow tunings is to pool the responses of cells with similar selectivities in local neighborhoods [150, 151]. Koch *et al*. [151] have shown this approach to narrow the tunings in orientation-selective cells. While multiplications cause a drop of firing rates and require an amplification of responses, the pooling approach does not reduce the firing rates. In order to compare the two methods and evaluate the effectiveness of each on narrowing the tunings, we implemented pooling in local regions of our mV2 layer, arguably the globs in V2. Our implementation follows the formulation of Koch *et al*. [151], but for color neurons. As the bandwidth distributions of mV2 neurons in Figure 2.10b shows, we found that even though pooling narrows the tunings, multiplication has a more significant effect in decreasing the bandwidths. In other words, the shift toward smaller bandwidths with respect to the bandwidth distribution of single-opponent mV2 cells is larger for the bandwidth distribution of multiplicative cells compared to that of pooling mV2 cells, as shown by the two arrows in Figure 2.10b. We stress here that

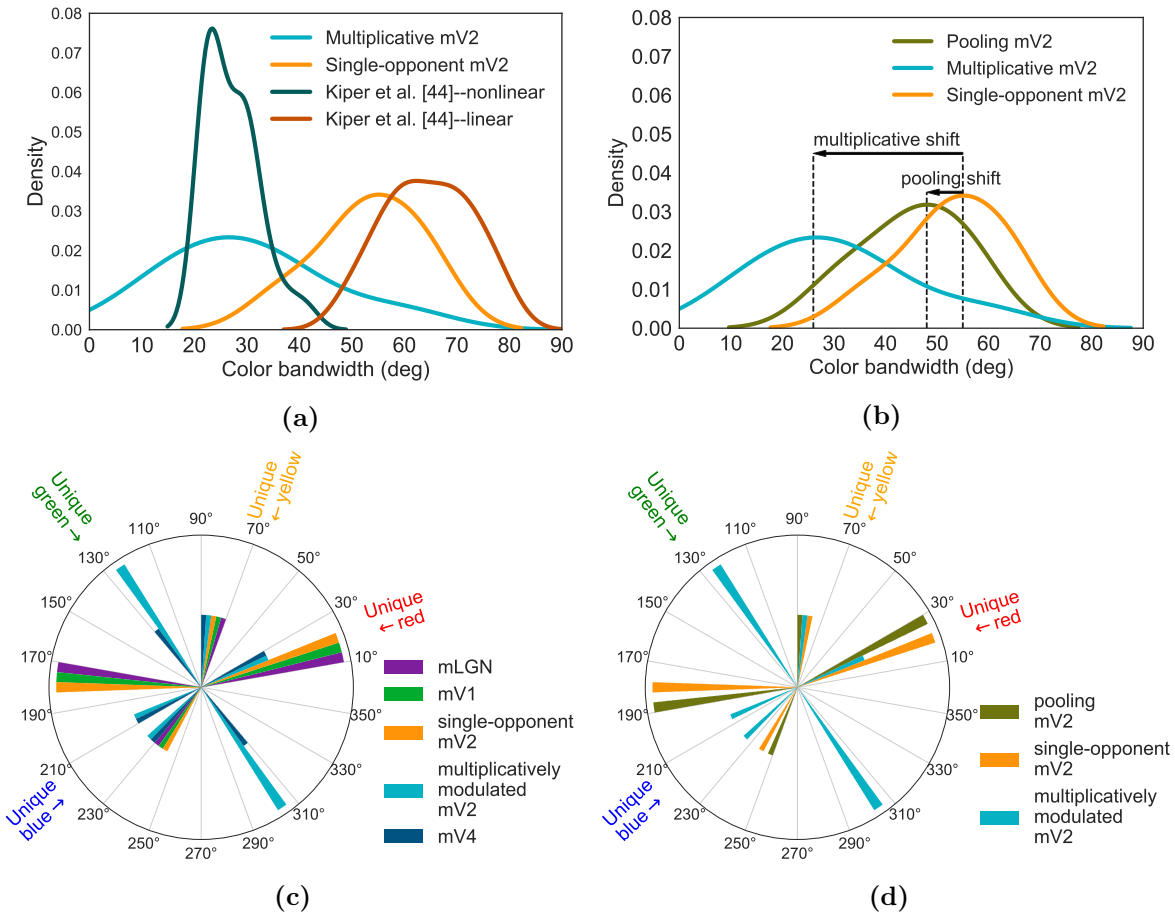


Figure 2.10: Tuning bandwidth and tuning peak analysis in the hierarchy. (a) Distributions of tuning bandwidths for mV2 neurons. Also, the bandwidth distributions of biological V2 cells reported by Kiper *et al.* [7] are plotted for comparison. The distribution of multiplicative mV2 neurons is clearly separate and shifted toward smaller bandwidths compared to single-opponent mV2 cells, similar to the separate distribution of the V2 population reported by Kiper *et al.* [7]. This confirms our observation that multiplicative modulations result in narrower tunings. Dashed vertical lines indicate the bandwidth at which each distribution peaks. While multiplicative mV2 and biological nonlinear V2 distributions peak at around similar bandwidths, our single-opponent mV2 distribution is slightly shifted compared to that of biological linear V2 neurons. (b) Distributions of tuning bandwidths of single-opponent, multiplicative and pooling mV2 neurons. Although pooling decreases the tuning bandwidths, the bandwidth distributions of these neurons has a large overlap with that of single-opponent mV2 cells. Multiplicative mV2 cells, in contrast, shift the distribution to smaller bandwidths, separating it from that of single-opponent mV2 cells. The shifts in distributions of multiplicative and pooling mV2 cells with respect to single-opponent mV2 neurons are shown by annotated black arrows. (c) A polar histogram of selectivity peaks for our model neurons. Our mLGN, mV1 and single-opponent mV2 neurons cluster around cone-opponent axis directions while our model multiplicative mV2 and mV4 cells have peaks both close to cone-opponent and off cone-opponent hue directions. Note specifically that all model layers have neurons in polar bins containing unique red and unique yellow hue angles. Bins including unique green and unique blue angles are limited to multiplicative mV2 and mV4 types. (d) Peak selectivities of single-opponent, multiplicative, and pooling mV2 cells shown in a polar histogram of selectivities. Despite decreasing tuning bandwidths, pooling mechanisms do not shift the selectivities to intermediate directions, whereas multiplicative modulations both decrease the tuning bandwidths and shift selectivities to intermediate hue directions.

our results of narrower tunings of multiplicative mechanisms versus pooling do not indicate that the hue selective neurons do not employ pooling mechanisms. However, our model confirms that a simple feedforward S-modulated multiplication has more success in decreasing the bandwidth compared to pooling in globs.

2.4.4 Tuning peaks

Watcher *et al*. [116] observed that most neurons in V1 peak around non-opponent directions, while Kiper *et al*. [7] reported that cells in V2 exhibited no preference to any particular color direction and no obvious bias to unique hues. We tested the tuning peaks of model neurons to examine for cone-opponent vs. intermediate selectivities. Figure 2.10c shows a polar histogram of tuning peaks of all types of neurons in all layers of our model, where each sector of the circle represents a bin of the polar histogram. This figure clearly demonstrates that the majority of mLGN, mV1, and single-opponent mV2 cells (5 out of 6 neuron types) peak close to cone-opponent axes. In contrast, the family of multiplicative mV2 cells and hue-sensitive mV4 neurons peak at both cone-opponent and intermediate hues, as reported in [7] and [122]. Simply put, with the multiplication mechanism, the representation of hues along intermediate directions starts to develop. This is in fact another advantage of multiplication over pooling of responses; multiplying the responses of cells with dissimilar selectivities rotates the peak responses to intermediate directions. In pooling, however, no shifting in the peak selectivities is achieved, as depicted in Figure 2.10d.

2.4.5 Unique hue representation

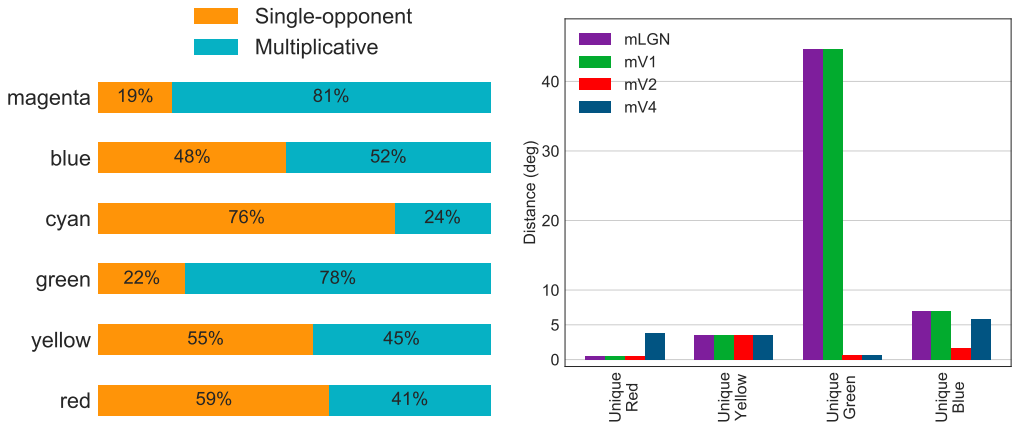
In Figure 2.10c, each mV4 bar in the polar histogram is paired with a multiplicative mV2 bar. Wondering about the contribution of multiplicative mV2 cells to the responses of mV4 neurons, we compared the sum of single-opponent mV2 cell weights against that of multiplicative mV2 neurons, depicted in Figure 2.11a. Interestingly, mV4 cells selective to magenta, blue, and green hues, which are off the cone-opponent directions, have significant contributions from multiplicative mV2 cells. In green and magenta, in particular, multiplicative cells make up more than 78% of mV2-mV4 weights. The rest of hues, which are close to the cone-opponent directions in the MB space, receive a relatively large feedforward input from the single-opponent mV2 cells. In short, multiplicative cells play a significant role in the representation of hues in intermediate directions,

while single-opponent cells have more substantial contributions to hues along cone-opponent axes.

Next, we asked the question of how well neurons in each of our model layers represent unique hues? To answer this, we computed the distance of peak angles for our model neurons to the unique hue angles reported by Miyahara [8]. For each unique hue, we report the minimum distance of peak angle among all neuron types in each layer to the given unique hue angle. As Figure 2.11b shows, the distances to unique red and yellow are relatively small in all the four model layers, at less than 5 deg. This could be due to the fact that unique red and yellow reported in [8] are close to cone-opponent axes in the MB space, in agreement with findings of [126] for unique red. For unique green, there is a 44 deg drop in mV2 and mV4 distances compared to that of mV1. Also, the distance to unique blue further decreases in mV2, although with an increase in mV4. The increase in mV4 to unique blue and unique red, however, does not indicate that mV4 neurons cannot represent these unique hues as well as mV2 cells. Additional mV4 neurons with significant inputs from mV2 cells representing these unique hues can have a distance similar to mV2 cells. To summarize, we observed a gradual development in the exhibition of selectivity to unique green and blue hues, while selectivity to unique red and yellow was observed in early as well as higher layers. Moreover, these results suggest that mV2 cells with peak selectivities at less than 5 deg distance to unique hue angles, and consequently, neurons in higher layers, have the capacity to encode unique hues. Our results do not suggest any significance of unique hues, as was also reported in previous research [131, 133, 132]. We emphasize that our model neurons were not tuned to achieve unique hue representations. Instead, the results of this analysis, put together with that of Figure 2.10c, indicate that by introducing multiplicative nonlinearities, neurons with selectivities spanning the color space start to emerge, some of which could potentially represent hues deemed as unique hues.

2.4.6 Hue distance correlation

Li *et al*. [6] found a correlation between pairs of stimulus hue distances and the cortical distances of maximally activated patches in each cluster. For this analysis, Li *et al*. [6] employed an ordered representation for hues and defined the hue distances accordingly. To test for a similar relationship between hue distances and the pattern of activities of mV4 neurons, we stacked our mV4 maps, in the order shown in Figure 2.3a, beginning with magenta, red, and so on. Stacking these maps results in a three-dimensional array, each column of which can be interpreted as a cluster of hue-



(a) (b)

Figure 2.11: Multiplicative neurons contribute greatly in intermediate and unique hue representations. (a) Relative contributions of single-opponent and multiplicative mV2 cells to each hue-sensitive mV4 neuron. The orange and blue bars show the percentage of contribution from single-opponent and multiplicative mV2 cells to each specific mV4 cell, respectively. Multiplicative mV2 neurons have more substantial input weight to mV4 cells with selectivity in intermediate hue directions. Single-opponent mV2 cells make up most of the input weight to mV4 cells with peaks close to cone-opponent axes. (b) Distances of our model neurons, layer-by-layer, to unique hue angles reported by Miyahara [8]. Unique hue representation develops in the hierarchy. Note the significant drop in the peak distance for mV2 and mV4 neurons to unique green, achieved by increasing the nonlinearity in those layers.

Figure 2.12: Hue distance correlation analysis between the hue distances and model patch distances in each model cluster. Our hierarchical model demonstrates a clear correlation between hue distances and the pattern of activations in mV4 neurons.

selective patches, with neighboring patches sensitive to related hues, similar to those observed in V4 [6]. We call each column of our stacked maps a model cluster and each element of the column a model patch. An example of a mV4 cluster in a larger view is shown in Figure 2.3a. For a given model cluster and a pair of stimulus hues, we compute the distance of the two maximally activated model patches. For example, the distance of the red patch from the cyan patch as shown in Figure 2.3a is 4. In this experiment, we employed our sampled hues from the HSL space, starting from red at 0 deg, separated 30 degrees, resulting in 12 stimulus hues. Similar to the ordering assigned to stimulus hues employed in [6], we assigned values in the range $[0, 5.5]$ at 0.5 steps starting with 0 for magenta. The plot in Figure 2.12 demonstrates our model patch distances as a function of hue distances with a clear correlation ($r = 0.93, p = 3.92 \times 10^{-30}$). Simply put, similar to the biological V4 cells, the pattern of responses in our mV4 neurons is highly correlated with the ordering of hues in the HSL space.

2.4.7 Hue reconstruction

Li *et al.* [6] observed that 1-4 patches are needed to represent any hue in the visual field and that different hues were encoded with different multi-patch patterns. Then, they suggested that a combination of these activated patches can form a representation for the much larger space of physical colors.

Along this line, we show, through a few examples, that for a given hue, a linear combination of mV4 neurons can be learned and used for representing that particular hue. Although it would be impossible to learn weights for the infinitely many possible physical hues, we show a few examples as an instance of the possible mechanism for color representation suggested by Li *et al.* [6].

In this experiment, for a given hue value, we uniformly and independently sampled the saturation and lightness dimensions at 500 points resulting in 500 colors of the same hue. The goal is to compute a linear combination of mV4 neurons, which can reconstruct the groundtruth hue. The hues in this experiment were represented as a number in the $(0, 2\pi]$ range. We performed an L1-regularized least square minimization [152].

Table 2.1 shows some of the results for this experiment. Interestingly, in all cases, no more than four neuron types have large weights compared to the rest of the neurons, in agreement with findings of [6]. Specifically, in the case of red and yellow hues, about 99% of the contribution is from

Table 2.1: The choice of weights for mV4 cells used for hue reconstruction in a few example hues.

Groundtruth hue (deg)	mV4 neuron					
	Red	Yellow	Green	Cyan	Blue	Magenta
red (360)	0.9934	0.0034	0.0006	0.0007	0.0006	0.0014
Yellow (60)	0.0010	0.9960	0.0008	0.0010	0.0004	0.0007
lavender (270)	0.0008	0.0019	0.0674	0.2946	0.0003	0.6351

only a single cell, red and yellow neurons respectively. The last row in Table 2.1 is most insightful. It presents the weights for a lavender hue in equal distance from blue (240 deg) and magenta (300 deg). The weights for this example seem counter-intuitive as they include cyan and magenta with positive contributions, and blue is absent. However, careful scrutiny of peak angles for mV4 hues reveals that lavender hue at 270 deg is somewhere between the peaks for mV4 cyan (at 193 deg) and magenta (at 300 deg), and closer to magenta. This hue is mainly reconstructed from magenta, with more than 63% contribution, while the absence of blue is compensated with cyan, shifting the reconstruction from magenta toward lavender.

Again, it must be stressed that this experiment was performed to examine the possibility of combinatorial representation mechanisms and a thorough investigation of this mechanism in the computational sense is left for future work. The examples shown here attest to the fact that intermediary hues encoded by mV4 neurons can indeed span the massive space of physical hues and are enough for reconstructing any arbitrary hue from this space.

2.4.8 V4 Representations Visualization

With observations from hue correlation and hue reconstruction experiments, it is of interest to visualize V4 representations in the final layer of the hue network. Following the hue correlation experiment results, the expectation is for V4 representations of dissimilar hues to be mapped to distinguishable clusters in the V4 space with samples at zero saturation and lightness clamped together. To visualize V4 representations, similar to the hue reconstruction experiment, we uniformly and independently sampled the saturation dimensions at 500 points for hues at the following hue angles: $\{0, 60, 270, 300\}$ deg, corresponding to red, yellow, lavender and magenta respectively. We applied Principal Component Analysis (PCA) to the six dimensional V4 representations and plotted the first three components as depicted in Figure 2.13. Each dot in this plot represents a single

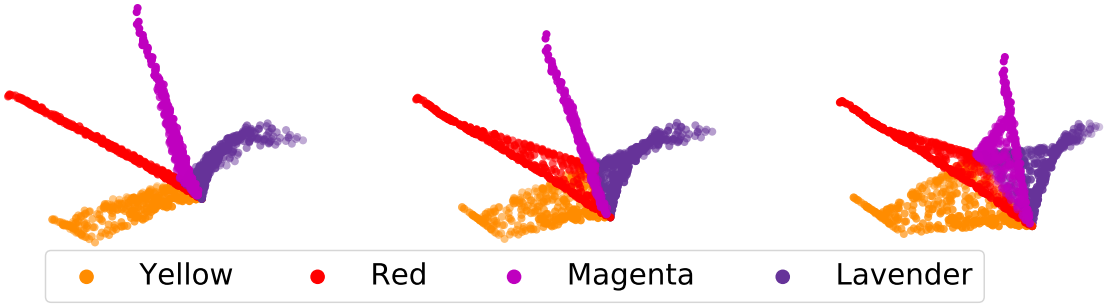


Figure 2.13: V4 representations visualization. The first three principal components of V4 representations plotted in 3D and viewed from different angles. Each dot represents a single sample with varying saturation and lightness and yellow, red, magenta and lavender hues. This figure demonstrate distinct clustering of samples according to their hues with the ordering of hues from yellow to red, magenta and lavender preserved. Note that samples with zero saturation and lightness look identical regardless of hue and are clamped together at the base of these four clusters.

sample and the color indicates the hue of the sample point. This figure demonstrates that the V4 representations are well-separated into distinct clusters for each hue. Moreover, the ordering of the hues is preserved in this space, confirming our observations from the hue correlation experiment. Finally, although modeling color constancy was not the focus of our work and computational color constancy was applied to our stimuli, results from the visualized representations in Figure 2.13 suggest that mV4 hue encoding are preserved even with changes in saturation and lightness.

2.4.9 Image Classification

We asked the question of how our mV4 hue representations perform in practical applications. To this end, we chose an image classification task in categorizing 103 classes of flowers from the dataset introduced by Nilsback and Zisserman [9]. A few examples of flowers in this dataset are shown in Figure 2.14. In their work, Nilsback and Zisserman combined shape and color feature kernels for classification of flowers using SVM. Nonetheless, our focus is on color and hence, we limit ourselves to color features. In [9], HSV channels were employed as color features and we remain with that choice for comparison with mV4 representations. For classification, we followed the steps explained in [9] by first performing a Kmeans clustering. Then, the normalized frequency histograms given the Kmeans clusters are computed followed by SVM classification. They found an optimum

Figure 2.14: Examples of flowers in the dataset employed for image classification and color image segmentation experiments described in sections 2.4.9 and 2.4.10. This dataset consists of 103 flower classes [9].

number of 1000 Kmeans clusters with the validation set and reported the recognition accuracy of 43%. We replicated their procedure and found 1400 Kmeans clusters as the optimum value with 42% recognition accuracy. Prior to these steps, however, all images are cropped according to the bounding box coordinates provided for each flower image in the dataset, and then each cropped image is resized to 256×256 pixels to make the image size compatible with our network input size. Resizing the images morphs flower shapes but the color features remain unchanged. Since our classification experiment is solely based on color, we do not expect resizing to affect the classification performance.

To evaluate the performance of our mV4 hue representations on flower classification, we simply used the six mV4 channels in the same manner that the three HSV channels were employed for classification. We found a significantly smaller number of Kmeans clusters to be enough for classification using V4 features. Our V4 features achieved 29% recognition performance with only 600 Kmeans clusters. Although we observed a drop of about 13% in performance compared to HSV features, the number of Kmeans clusters are significantly smaller, 600 for V4 vs. 1400 for HSV. We believe that V4 features, in a higher-dimensional space, have the potential to map dissimilar hues to relatively more distant points in the 6 dimensional mV4 feature space while keeping the similar ones in close proximity. As a result, data points are placed in distinctly well-separated groups, and a smaller number of Kmeans centers will be required for clustering the data into similarity groups. The reported number of clusters and accuracies in this experiment were obtained with five-fold cross-validation. Changes in the hyper-parameters of the classifier had little effect on the optimum

Feature type	# of Kmeans clusters	Recognition rate
HSV	1400	42%
V4	600	29%
V4SV	600	35%
$H_{smooth}SV$	1400	24%

Table 2.2: Flower classification performance based on color features.

number of clusters. Specifically, with V4 features, for a different setting of SVM parameters, the optimum number of clusters increased from 600 to 800, and not to a significantly larger number. This observation suggests that the difference between the number of clusters with V4 and HSV features is not due to differences in the classifier parameters.

The decrease in recognition performance from HSV features to mV4 features could be due to two differences between the two representations: first, unlike the HSV channels, our mV4 representation encodes the hue component of color and not saturation and value. Second, our mV4 neurons have relatively large receptively fields, while HSV data is at pixel-level. The higher resolution might put the HSV representation at an advantage. To overcome these differences between HSV and mV4 features in favor of a fair comparison, we tested with two additional sets of features: first, we concatenated our mV4 representation with the saturation and value channels of HSV. We called this representation V4SV. Second, we smoothed the hue channel of HSV, leaving saturation and value intact, with a Gaussian kernel with approximately the same extent as our mV4 receptive fields. This feature was called $H_{smooth}SV$.

The summary of classification performances is provided in Table 2.2. As the recognition rate of these new features shows in Table 2.2, adding saturation and value to mV4 representations in V4SV increases the performance by about 6% to 35%, while keeping the number of clusters at 600. This increase in performance suggests that flower classification is not solely based on hue and that saturation and value seem to add further information valuable for this task. In case of the $H_{smooth}SV$ feature, the number of clusters is the same as that of pixel-level HSV, at 1400 Kmeans cluster. However, the performance drops to 24%, which is even 5% lower than that of our mV4 representations with no saturation and value components.

As a conclusion, when we modified HSV and mV4 representations to make these features alike, our mV4 hue representations outperform the HSV channels. This, we believe, is due to better

separation of hues in our model hue encoding, which also results in less number of Kmeans clusters for classification purposes.

We acknowledge that neither our mV4 representation nor V4SV features achieved the 42% classification accuracy of HSV channels at full resolution. However, we hypothesize that should we increase the resolution of mV4 features in a future work, the classification accuracy of our model will match or even exceed that of HSV channels, just in the same way our V4SV features outperform $H_{smooth}SV$ channels at equal resolution. This extension is left for future investigations.

2.4.10 Color Image Segmentation

As another test for our model hue representation performance in practical applications, we performed a color image segmentation experiment. For this purpose, we employed HSV and V4SV features, described in the previous section. We randomly selected 12 images from the dataset and performed Kmeans on the color features. The number of clusters in this experiment was manually set, subjectively based on the image, to a number between 2 to 4. This number was fixed for both HSV and V4SV features applied to each image. We evaluated segmentation performance based on Intersection-over-Union (IoU) with respect to the provided groundtruth segmentation for each image in the dataset. The IoU results for the flower images employed in this experiment are shown in Figure 2.15 and the images along with HSV and V4SV segmentation are in Figure 2.16.

The IoU bars in Figure 2.15 clearly demonstrate that V4SV features outperform HSV segmentations in all but one of the examples. Additionally, as depicted in Figure 2.16, V4SV segmentations consist of contiguous regions that are also respective to boundaries, rather than small isolated segments observed in HSV segmentations.

2.5 Discussion

Our goal was to introduce an analytically-defined hierarchical neural network that incorporates discoveries of color processing the visual cortex. By means of such a model and adhering to available color processing knowledge in the brain, we also promoted our understanding of the color processing mechanisms in the brain and began to assign color representational roles to specific brain areas. We investigated the contributions of each visual area LGN, V1, V2, and V4 in hue representation by

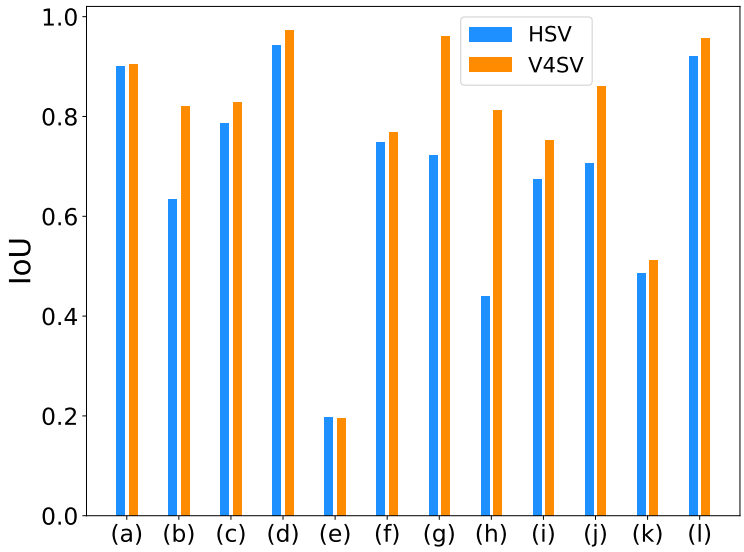


Figure 2.15: Intersection-over-Union (IoU) measure for the flower images shown in Figure 2.16. Each pair of IoU bars are labeled according to the caption index of images in Figure 2.16. In all the examples, except for one (flower image with label (e)), V4SV outperforms HSV.

proposing a mechanistic computational model inspired by neural mechanisms in the visual system. Through a gradual increase in nonlinearity in terms of cone inputs, we observed a steady decrease in tuning bandwidths with a gradual shift in peak selectivities toward intermediate hue directions. Although one might be able to obtain the end result with a mathematical model in a single-layer fashion, such models do not lend insight to the neuronal mechanisms of color processing in the brain. In contrast, not only do our model neurons in each individual layer exhibit behavior similar to those of biological cells, but also at the system level, our hierarchical model as a whole provides a plausible process for the progression of hue representation in the brain. The main difference in terms of potential insight provided by a single-layer mathematical model and our work is that our model can make predictions about real neurons that can be tested. A model whose contributions are of the input-output behavior kind cannot make predictions about component processes between input and output (see also [49]).

Multiplicative modulations in our model V2 layer increased nonlinearity in the hierarchy, narrowed tunings, and shifted tuning peaks toward intermediate directions. Our work suggests that the brain's neurons also utilize such multiplicative mechanisms, which result in nonlinearity in their tunings and shifting of tuning peaks to intermediate directions, consistent with the report from Kiper *et al.* [7]. Such nonlinearities result in representations that span the color space instead

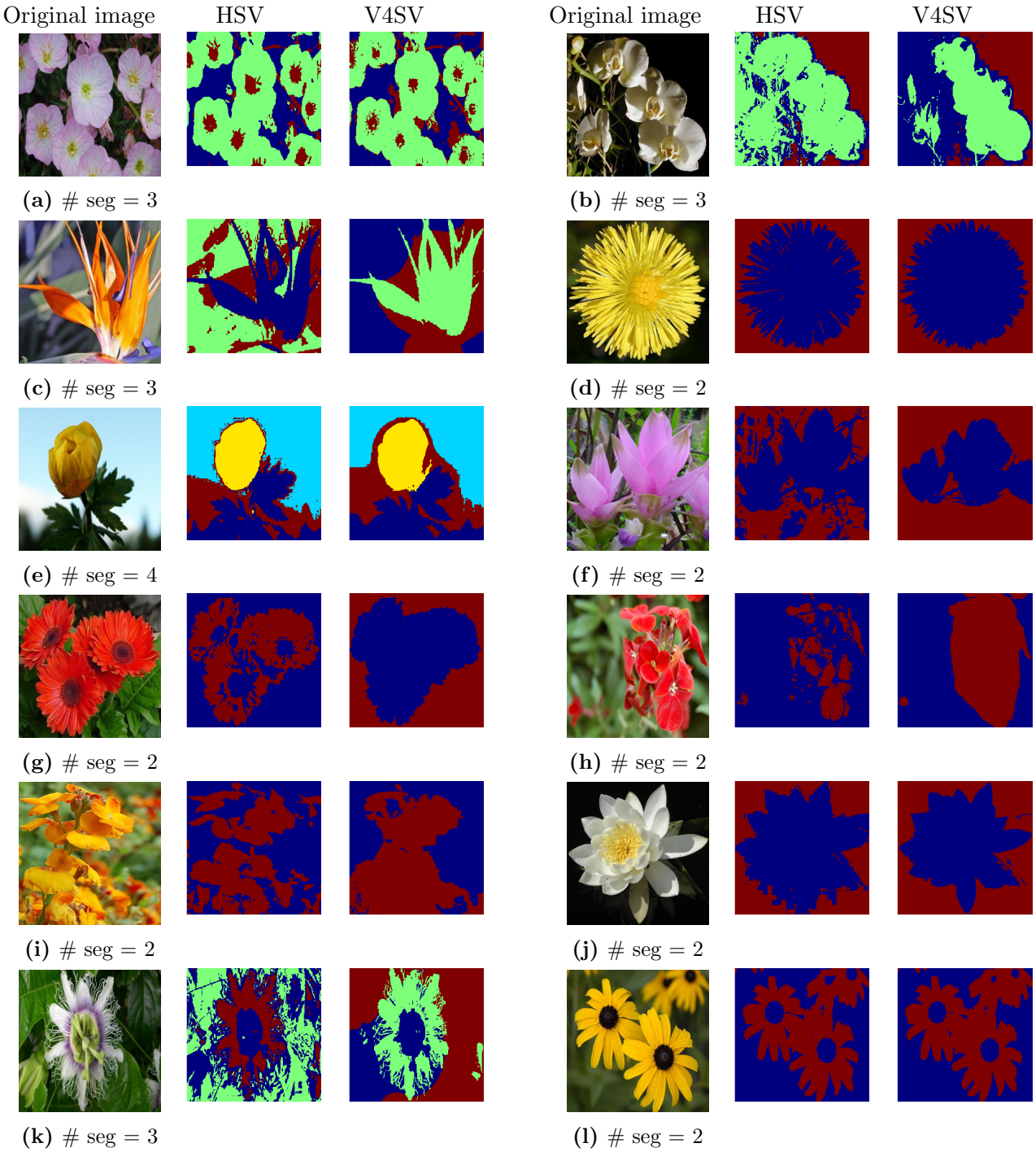


Figure 2.16: Color image segmentation result on randomly selected flower images from the dataset. In each set of three figures, original image, HSV and V4SV segmentations are shown from left to right. Note that the color assigned to each segment in some of the examples might be different from HSV to V4SV, due to the differences in Kmeans cluster indexing. A close look at the segmentations show a cleaner and more accurate segmentation with V4SV features. See, for example, segmentations in (g), (h), (j).

of reducing it to only "skeletons" of cardinal axes [125, 150]. They also enable discrimination of slightly different hues, such as red and orange, by increasing the representational distances due to narrow tunings.

Our experimental results demonstrated that hue selectivity for mV4 neurons similar to that of neurons in area V4 of the monkey visual system was successfully achieved. Additionally, our observations from the hue reconstruction experiment confirmed the possibility of reconstructing the whole hue space using a combination of the hue-selective neurons in the mV4 layer. How this is achieved in the brain, for the infinitely many possible hues, remains to be investigated.

Our hierarchical model provides an important suggestion about unique hue representations. Specifically, our computational experiments showed that as the visual signal moves through the hierarchy, responses with peaks close to unique hues start to develop; for example, for unique red as early as mLGN and for unique green delayed to mV2. Putting these together, we believe the answer to the question "which region in the brain represents unique hues?" is not limited to a single brain area, which in turn could be the source of disagreement among color vision researchers. Instead, our findings suggest that this question must be asked for each individual unique hue. Then, putting answers for all unique hues together will result in a set of brain regions. In our model, $\{\text{mLGN}, \text{mV2}, \text{mV4}\}$ is the answer. Although our model allowed us to make predictions about unique hues, our model neurons are not tuned for unique hues and the argument can be generalized for any arbitrary hue in intermediate directions.

While most of the color research is focused on single-opponent mechanisms and beyond those much is unknown, the hue network suggests that the observed responses in hue-selective neurons in higher ventral areas cannot be explained by simply combining single-opponent cell responses. Instead, the hue network provides an example that other types of mechanisms, multiplication in this case, could better explain cell responses. Comparing multiplicative modulations with pooling demonstrated a disparity in the effectiveness of the two types of modulation. These results suggest that more research with careful stimuli design that could identify color processing beyond single-opponency in higher ventral areas is required.

In our model, adding a variety of neurons such as concentric and elongated double-opponent color cells, similar to the examples demonstrated in Figure 2.17, would result in a more comprehensive system. However, we did not intend to make predictions about all aspects of color

processing but only hue encoding mechanisms. We found that concentric double-opponent color cells, for example, have tuning bandwidth distribution similar to single-opponent neurons and tuning peaks along cone-opponent axes. This finding suggests that the contributions of concentric double-opponent cells are comparable to those of single-opponent neurons for hue representation, but we did not investigate those contributions in other color representations. Whereas our network models single-opponent neurons and further extends this channel with multiplicative modulations, another hierarchical network introduced by Zhang *et al.* [10] modeled elongated single-opponent neurons and extended those with summation. Their proposed network implementing form processing along color borders demonstrated improved performance in tasks such as object recognition and color image segmentation compared to grayscale or shape-based descriptors. The color hierarchy proposed by Zhang *et al.* [10] is complementary to our hue network and combining the two hierarchies enables modeling of a wider variety of hue-selective neurons.

Our proposed hierarchy yields hue representations that could achieve comparable performance to that of the HSV when utilized in flower classification tasks. Although the model did not match the performance of HSV features, due to lack of S and V channels and large receptive fields of V4 neurons, we expect that overcoming these differences in the model, by implementing S and V channels and mechanisms that address the decrease in resolution in V4 neurons, the classification accuracies of the hue network will be improved. Performance improvement by concatenating V4 representations with S and V channels of HSV suggests that both saturation and value play a role in the classification task and that upon modeling saturation and value, a boost in performance can be expected. While the network did not match the classification accuracies of HSV channels, in the segmentation task, the hue network combined with value and saturation channels exhibited higher IoUs. Results from these two experiments were part of a pilot study and were not reported in the published form of this chapter. Therefore, these results are not final and further improvement of the model for better classification performance is left for a future work.

As a computer vision system, our hierarchical model can be further extended to encode saturation and lightness. Also, we would like to address the problem of learning weights in the network for which current biological findings have no suggestion, for example, the set of weights from mV2 to mV4. Lastly, in order to keep our model simple and avoid second-order equations, we skipped lateral connections between neuron types. However, these are part of the future development of a

second-order model for our network.

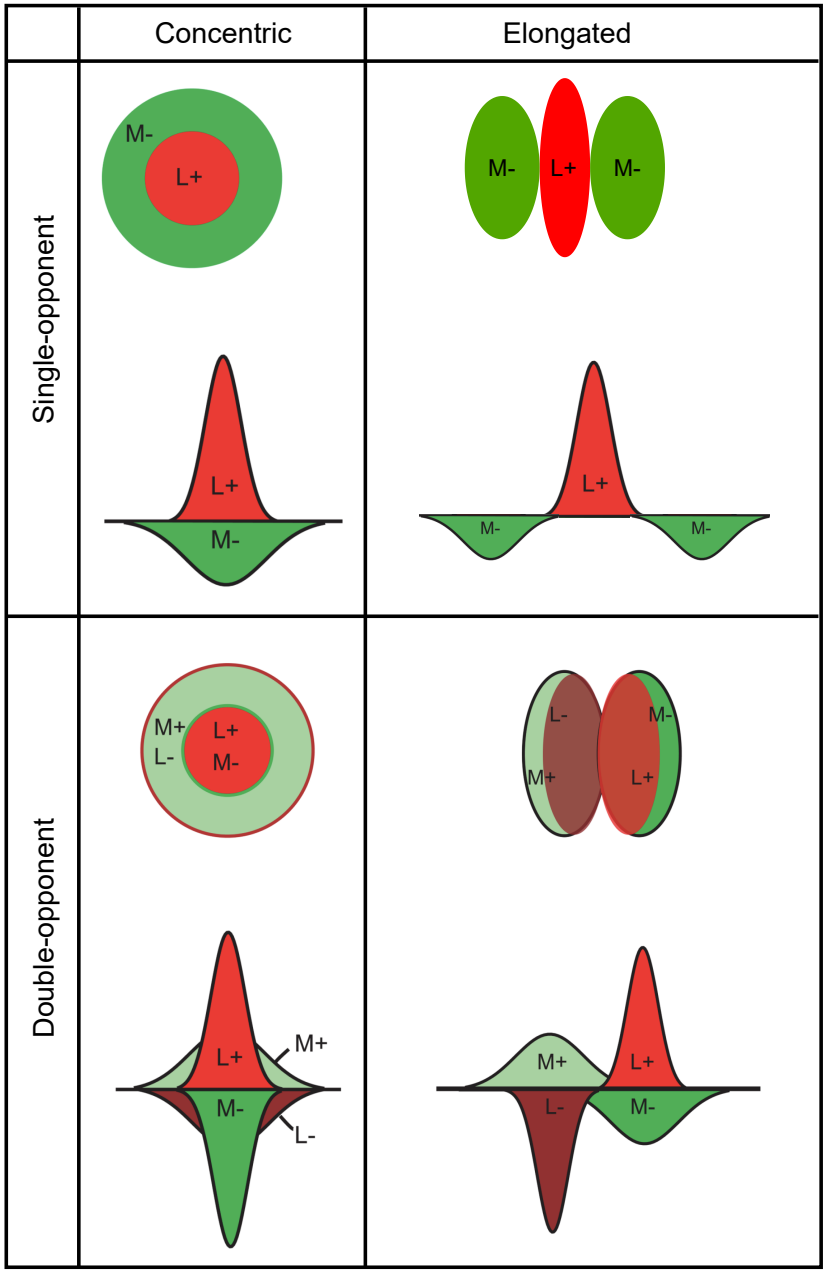


Figure 2.17: Examples of models of hue-selective neurons. Four models of single- and double-opponent neurons are depicted. Note that possibilities of receptive field profiles of hue-selective neurons are not restricted to these four models. This figure, however, demonstrates four models with concentric/elongated and single-/double-opponent receptive fields to show a few possibilities. Whereas in single-opponent cells each cone input has either an excitatory or inhibitory effect, in double-opponent neurons, input from a cone type can be both excitatory and inhibitory. In neurons categorized as "Elongated", cone inputs are pooled over an elongated summation subregions. These neurons are excited in presence of color bars and borders. In concentric cells, strongest cone inputs are to the RF center with symmetric subregions. The proposed Hue Network implements concentric single-opponent neurons and further extends this channel with multiplicative modulations. Our experiments with concentric double-opponent neurons did not suggest narrower tuning bandwidths in these cells compared to single-opponent concentric cells. The color model proposed by Zhang *et al.* [10] implements elongated single-opponent cells and extends this channel with summation.

Chapter 3

Figure Border Ownership: An Early Recurrence Model

The work in this chapter has been published previously as the following:

Paria Mehrani and John K. Tsotsos "Early recurrence enables figure border ownership.", in *Vision research*, 186, 23-33, [2021]

and has been extended here for the purposes of this document. Specifically, the literature review in the present chapter is an extended version of the published work.

3.1 Motivation

Objects in an image introduce occlusion borders between the object closer to the viewer and either the ground or other objects. Ownership at these borders belongs to the closer occluding one. Ownership information at occlusion boundaries plays a role for figure-ground assignment and perceptual organization. For example, in Figure 3.2, an image and its corresponding hand-labeled ownership assignment along some occlusion boundaries are shown. This ownership labeling provides a clear figure-ground organization of the observed scene, here, for example, the baby deer being in front of its mother.

Although localizing the occlusion boundaries is a necessary step for figure-ground organization processing, determining ownership at these boundaries requires more than edge detection and local feature extraction. In a neurophysiological study, Zhou *et al.* [11] showed that neurons exist in V1, V2 and V4, whose responses are context-dependent. Given the evidence of border ownership encoding in the visual cortex, this chapter is dedicated to proposing a hierarchical network that represents ownership at occlusion boundaries. It goes without saying that the introduced model is expected to satisfy known biological visual processing properties.

3.2 Background

3.2.1 Border Ownership Assignment in the Visual Cortex

Findings of Zhou *et al.* [11] suggest that the primate visual system assigns ownership along occlusion borders as an early step in perceptual organization and figure-ground assignment. Zhou *et al.* [11] found that some neurons in V1, V2, and V4 encode ownership along figure outlines by exhibiting a difference in responses to figure on either side of the border even with identical local features within their classical receptive fields. As an example, consider the responses of a neuron from this study shown in Figure 3.1a. In this figure, the little black ellipse in the middle of each display represents the classical receptive field of the neuron, and a close look reveals the same local features in pairs of stimuli A and B in each column. However, the context between the pairs differs in the sense that in displays of row A, the figure is on the left side of the border, while it resides in the opposite side for all displays in row B. The bars at the bottom of each column show the corresponding responses.

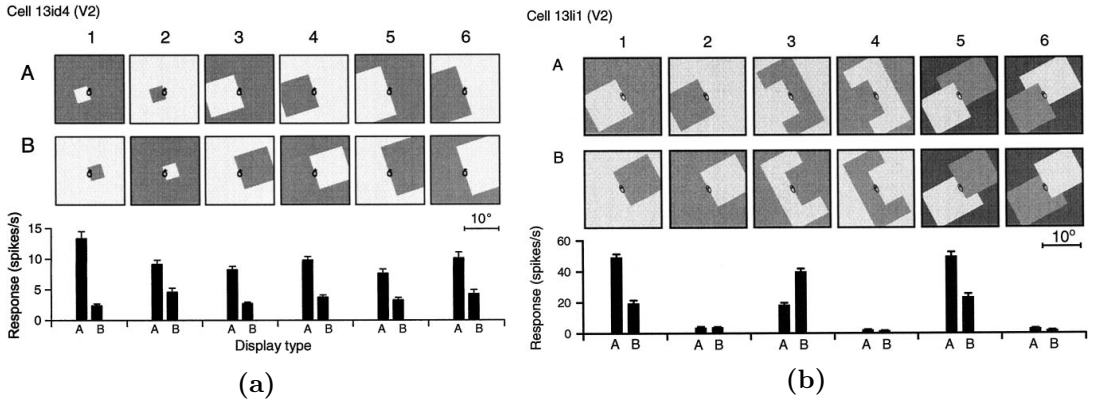


Figure 3.1: Responses of type 1 (a) and type 3 (b) neurons. A black ellipse shows the classical receptive field of the recorded neuron in each display. The local features within the classical receptive field of each neuron are identical in each column of displays. The responses for both neurons, however, show a preference to stimuli with the figure on the left, while the type 3 neuron also exhibits selectivity to contrast polarity at the border (compare responses to displays in columns 1 and 2 of (b)). Although the large figures in (a) (columns 3-6) are not complete, according to the Gestalt principle of closure, these figures are to be perceived as square figures rather than holes with the rest of the image as the ground. Plots adapted from [11].

For the neuron shown in this figure, the responses are stronger to stimuli in row A with the figure on the left and weaker to stimuli in row B with the figure on the right, for all pairs of displays presented. In other words, this neuron shows a preference to stimuli with the right side of the figure within its receptive field and a much weaker response to those with the left side of the figure in its receptive field. Zhou *et al*. [11] called this type of selectivity the side-of-figure preference and found that the context-dependent responses of these neurons encoded ownership information at occlusion boundaries and called them border ownership (BO) cells. Having observed this behavior from BO cells, Zhou *et al*. [11] concluded that these cells integrate global context from beyond their classical RF to encode such side-of-figure preference.

Zhou *et al*. [11] classified two types of BO cells, called type 1 and 3. Type 1 BO cells are selective to side-of-figure and not contrast polarity in their classical RF. Type 3 BO neurons are selective to side-of-figure and contrast polarity features. Type 2 and type 0 exhibit no side-of-figure preferences, regardless of their local feature selectivity. Responses of a type 1 neuron are shown in Figure 3.1a. Figure 3.1b demonstrates responses of a type 3 cell, where there are no responses to displays in even-numbered columns with local features not matching the selectivity of this neuron. The responses to stimuli in odd columns, however, show the preference for the figure to be on the

(a) **(b)**

Figure 3.2: (a) An example of an image from the Berkeley dataset for ownership assignment at occlusion boundaries in natural images [12, 13], (b) the corresponding human-labeled ownership assignment along some of the occlusion boundaries. In (b), white represents the side that owns the border, *i.e.*, side of the occluding object, and black is for the occluded region.

left side of the border. Zhou *et al.* [11] also found border ownership neurons maintain the difference of responses to figures of various sizes and changes in position of the border within their receptive fields. When presented with solid and outlined squares, border ownership responses were consistent across stimuli. As shown in the stimuli set for the type 3 neuron in Figure 3.1b, the difference of responses was maintained for simple as well as more complicated stimuli, such as a C-shaped figure or overlapping squares.

Intriguing about BO cells is their short latencies, *i.e.*, the time the difference of responses to figure on either side of the border achieves its half-peak point. BO cells showed consistently short latencies of 70ms to both synthetic [11] and complex natural [153, 154] stimuli. With such short latencies, and despite two decades of research, the source of global context that affects BO cells is still unknown. Not only do BO cells have surprisingly short delays of 10-25ms for the differentiation of responses, but also as Zhou *et al.* [11] noted, their "responses to the preferred and non-preferred sides diverged almost from the beginning". This early divergence is shown in Figures 3.1a and 3.1b (For more examples, see [11], Figures 4, 7, and 20, and [155], Figures 2A, 4A-D, and 5).

In pursuit of understanding the global context integration in BO cells, Qiu *et al.* [156] observed that attention is not required for border ownership assignment and that border ownership was assigned for any figure in the display, both attended and ignored. However, they found that attentional enhancements were stronger for the figure on the preferred side of the border. In an interesting work, Zhang and von der Heydt [14] investigated the structure of the surround that

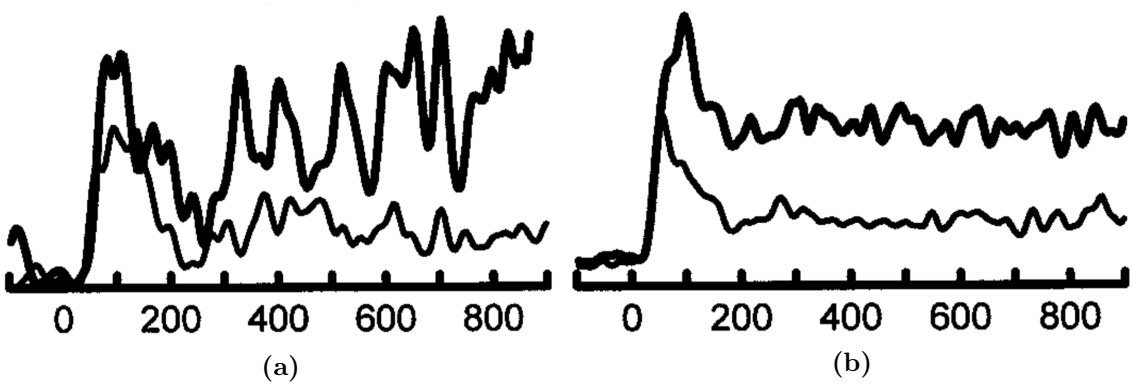


Figure 3.3: Responses of BO cells over time. (a), (b) The average response of all BO neurons in V1 and V2 as a function of time after the stimulus onset, reported by [11]. These plots show the divergence of responses to preferred (thick lines) and non-preferred (thin lines) sides from almost the beginning. Plots are adapted from [11].

provides contextual information to border ownership neurons. Their stimuli set contained displays of fragmented Cornsweet squares[1] such as those in Figure 3.4. While one edge was centered on the classical receptive field, some of the rest of the fragments were occluded, and the effect of the presence or absence of each fragment was studied. They found all fragments have a modulatory effect on the response of the neuron in a uniform fashion, with positive and negative modulations from fragments of squares on preferred and non-preferred sides respectively. Furthermore, they obtained weaker border ownership signals for scrambled fragmented figures such as those presented in panel C of Figure 3.4.

Perhaps a prominent aspect of the study by Zhang *et al*. [14] is the evaluation of the plausibility of a feedforward model with two "hot-spots"[2] providing context to border ownership cells, similar to the one suggested by Sakai and Nishimura [158]. They found that the population data does not confirm the assumption of such models and that the surround influence on BO responses is "broadly distributed on both sides" of the RF. They examined the plausibility of models with only lateral connections[3], such as that of Zhaoping's [159]. They discovered that lateral connections could not

[1]The contours of Cornsweet figures are step edges with exponential decays on either side such that the inside of the figure has the same intensity as the background [157]. Cornsweet stimuli allow for seamless occlusion of contour fragments.

[2]In their proposal, Sakai and Nishimura [158] suggested employing modulatory signals from a facilitatory region and a suppressive region located asymmetrically on either side of an occlusion boundary for border ownership assignment. Zhang *et al*. [14] called this proposal the two hot-spots (facilitatory and suppressive) hypothesis.

[3]In Zhang's work [14], the term "lateral connections" stands for interareal interactions between neurons with receptive fields in neighboring regions of the visual field. In computational models such as convolutional neural networks (CNNs), lateral connections are implemented using inter- and intra-feature-map interactions of neurons within a single layer with neighboring receptive fields.

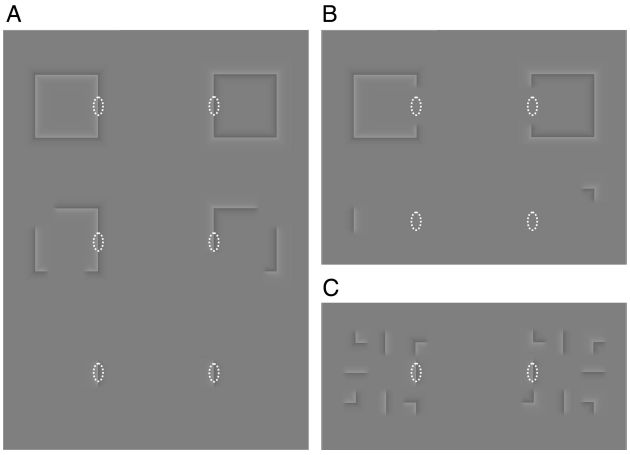


Figure 3.4: Displays of fragmented Cornsweet figures for testing the effect of surround providing contextual information for border ownership assignments (adapted from [14]). The white dashed ellipse shows the receptive field of the recorded neuron. Squares on either side of the receptive field with both contrast polarities were tested. Each square is divided into eight fragments, including center and corner fragments of equal length. The center fragment on one side of the figure is placed on the receptive field. In each display, some of the remaining seven fragments are occluded, as some examples are shown in panel A. In a control study, the fragment centered on the receptive field is occluded, with some of the other fragments present (panel B). They also recorded the responses where the seven fragments of the square are scrambled (panel C).

be the only source of contextual information for border ownership assignment and a model with both types of connections, feedback and lateral, could explain the data. Specifically, having observed an earlier arrival of the signal from distant shape parts to BO neurons compared to closer fragments, Zhang *et al.* [14] suggested lateral modulation could not explain this observation. Earlier, Craft *et al.* [18] had put forth a similar suggestion by providing a detailed discussion on the time course of border ownership cells from the study by Zhou *et al.* [11] and speed of conduction for lateral connections. This suggestion was also confirmed in a later neurophysiological study by Sugihara *et al.* [160], who tested border ownership neurons with displays of small and large figures. Sugihara *et al.* [160] found that even though the latencies of border ownership cells increase for larger figures, responses in these cells are much faster to be affected by only lateral signal propagation. They argued that the latencies of border ownership neurons in their experiments did not rule out the possibility of a combination of feedback and lateral modulations.

Investigating the border ownership responses to occlusion boundaries in natural images, two separate studies reported similar short latencies [153, 154]. Additionally, Williford and von der Heydt [153] examined the possibility of feedback from IT considering the time course of IT cells

reported by Brincat and Connor [161] and concluded that modulations from IT are not biologically plausible for BO assignment.

To summarize, the answer to the question, "which area(s) in the brain provides context to border ownership neurons?" is unclear. Instead, it is clear that this fundamental problem is solved in the biological visual system and very early on, in V1 as an early stage in the processing pipeline, which speaks to its importance for forming more abstract representations in later processing stages. Computational modeling of border ownership assignment was addressed in both computer vision and computational neuroscience communities. Next, the relevant literature is reviewed.

3.2.2 Computational Models for Border Ownership Assignment

The development of a computational model capable of determining border ownership at occlusion boundaries has been addressed by both computer vision scientists and computational neuroscientists, but often with different objectives. In computer vision, the goal is to have as accurate a border ownership label at occlusion boundaries as possible, by means of computational algorithms with no constraints on biological-plausibility. The performance in such models is usually measured against a manually-labeled dataset. In contrast, the latter approaches aim at designing models that meet a set of known biological constraints while the simulated neurons replicate the behavior of the biological ones, hoping to suggest a mechanism for the emergence of the observed signal in the brain. Often, the performance of these models is judged on how well the simulated neurons replicate the behavior of their biological counterparts on a similar set of stimuli. Sometimes, the plausibility of these models is examined with further neurophysiological studies. In this section, both type of approaches are reviewed.

3.2.2.1 Computer Vision Models

In computer vision, some methods rely on motion and depth information for boundary assignments [162, 163]. Inferring figure-ground organization in 2D stationary images was either addressed by limiting the input to line drawings [164, 165], or by extracting local features for natural images and applying machine learning techniques [166, 13, 12]. Usually, approaches in the latter category formulate the problem in a conditional random field (CRF) framework to incorporate context into their model. A number of these models, such as [166], employ shape-related features like convexity-

concavity of the contour segment or local geometric class like sky, ground, or vertical. A CNN-based model for figure-ground organization in a single 2D image, called DOC, was introduced in [167]. Their network consists of two streams, one detecting edges and the other specifying occlusion orientation, solving a joint labeling problem. They also introduced a semi-automatic method to provide figure-ground labeling for 20k images from the PASCAL dataset and showed superior performance over both PASCAL and figure-ground BSD [168] dataset.

Except for methods based on low-level cues, which are not among state-of-the-art, computer vision approaches generally attempt to solve the figure-ground organization problem by providing high-level features such as object class. For example, Hoiem *et al.* [166] first classified each superpixel in the image as one of vertical, sky, or ground, which provided strong priors over the object classes. In other words, these methods solve an object categorization problem prior to addressing figure-ground organization. Even in the case of DOC, although the network is trained on edge and occlusion boundary labels, there is no mechanism to ensure no implicit representation of object category is learned. Given an object category, these models mainly leverage learned statistics of how likely an object is to be the figure or ground. For example, the most likely event for the occlusion boundaries between any vertical object and sky is to be owned by the former category.

On a more quantified note, in a recent study, Hu *et al.* [42] probed three CNNs, including DOC, Mask R-CNN [169], and ResNet50 [170], to examine border ownership encoding in these models. They employed simple experimental stimuli such as those utilized by Zhou *et al.* [11] as well as natural images. They found that earlier layers showed contrast tuning while border ownership encoding increased in deeper layers. Moreover, the DOC network did not generalize to simple experimental stimuli. These findings depict obvious differences between CNNs and the brain, namely:

1. In the primate brain, border ownership encoding appears very early in the visual hierarchy, as early as V1 and V2, whereas in CNNs, BO encoding is delayed to deeper layers,

2. In the primate brain, BO cell responses were stronger to simple experimental stimuli than to natural one [153].

Hu's findings next to those of Bau *et al.* [91] who reported appearance of object representations in CNNs from early layers strengthen the hypothesis that instead of solving the border ownership

assignment before object recognition, these methods solve the latter first and infer border ownership based on learned statistics. With the mentioned ordering of visual processing in these models, these approaches detour addressing the border ownership assignment problem and are in contrast with findings of border ownership assignment. In particular, the short latencies of border ownership cells suggest that ownership assignment happens well before object recognition in the ventral stream [11, 171] as we will discuss in detail in Section 3.2.3, which leaves us with an unanswered question: how to achieve border ownership assignment without first solving object classification?

3.2.2.2 Biologically-inspired Models

In the years following the border ownership study by Zhou *et al.* [11], there were a number of endeavors to suggest a mechanism for border ownership assignment in the brain. For example, Zhaoping [159] proposed a model of neurons in V1 and V2. Contextual information is provided by means of lateral connections in V2 and imposing Gestalt grouping principles such as continuity and convexity as well as T-junction priors. Sakai and Nishimura [158] employed a vast pool of suppression and facilitation surround neurons for this purpose. In subsequent work, Sakai *et al.* [172], studied the responses of border ownership cells in this model to the complexity of shapes and found a decrease in responses with an increase in the complexity of shapes. Later, the plausibility of these models was ruled out by the neurophysiological findings of Zhang *et al.* [14]. Specifically, Zhang *et al.* [14] reported that in contrast to the proposal of Sakai and Nishimura [158], context from both sides of the RF affects BO responses and that the effect is not localized. Additionally, Zhang *et al.* [14] reported that the signal from distant shape parts arrived earlier than that of closer fragments, an observation that cannot be explained by the lateral model suggested by Zhaoping [159]. In the same year as Zhang's work, Super *et al.* [173] suggest another feedforward model based on surround suppression. An odd notion in their model is that feature maps for the figure and background are fed as input to their two-stream network: in a sense, they provide the "answer" in advance.

The majority of computational models rely on feedback from higher layers. For example, Jehee *et al.* [174] introduced a recurrent network with five areas V1, V2, V4, TEO and TE for contour extraction and border ownership computation. Similarly, border ownership assignment in models of Layton *et al.* [175] and Tschechne *et al.* [176] is based on feedback from V4 and IT. Sugihara *et*

al. [160] as well as Williford *et al*. [153], however, found that the time course of border ownership neurons does not support feedback from areas IT and beyond.

Feedback from V4 in some of the biologically-inspired models is based on responses of grouping neurons. The earliest model, introduced by Craft *et al*. [18], assigns border ownership using feedback from contour grouping cells in V4 and relies on convexity and proximity priors. Russell *et al*. [177] incorporated this model in a larger network and demonstrated improvements in saliency detection. Another extension of Craft's model was introduced by Layton *et al*. [178] with an additional layer of neurons, named R cells, with larger receptive fields than the grouping cells in V4. Then, border ownership is a result of feedback modulations from both grouping and R neurons, as well as imposed priors like convexity and closure. Despite the fact that no neurophysiological evidence against these models has been found, these approaches are dependent on shape priors such as convexity and proximity. Moreover, they rely on feedback from grouping cells in V4 to provide the required contextual information. Nonetheless, as we discuss in detail in Section 3.2.3, the time course of V4 neuron responses does not support the suggestion of border ownership modulations by feedback from these cells.

As a summary, the existing biologically-inspired models based on lateral connections or ventral feedback, though replicating the behavior of biological border ownership cells, cannot provide a timely signal carrying contextual information to border ownership neurons.

3.2.3 Border Ownership Assignment: Two Hypotheses

Studying context integration time course, Sugihara *et al*. [160] reported mean BO latency of about 70ms in both V1 and V2 with the edge onset of responses at about 42ms and 48ms respectively. These latencies together with early divergence of responses imply the existence of a contextual signal in V1 as early as 42ms post-stimulus onset. If feedback is the source of such signal, a key question is which area in the brain is fast enough to integrate context and provide it to V1 in 42ms. Given the BO signal time course and that the difference of responses are maintained for figures as large as 10deg [11], neurons in the area(s) providing context to BO neurons are expected to have two properties: first, faster than 42ms latency, including feedback conduction time for their signal to be present in V1 at edge onset, and second, extending the surround in V1 neurons to as large as 10 deg [11]). With these two properties in mind, here, we explore two possible hypotheses: one,

signals from higher ventral areas, and two, signals from higher dorsal areas; in both cases feedback signals would impact responses[4].

The ventral pathway has projections mainly from parvocellular cells in the LGN, whereas the dorsal stream contains projections from magnocellular cells. Parvocellular cells have small receptive fields with medium conduction velocity. In contrast, magnocellular neurons are known for their large receptive fields and fast conduction. It is thus not surprising that the dorsal stream with magnocellular projections is reported to exhibit similar properties [1, 179], which motivated the dorsal hypothesis here. For BO cells with short latencies, the timing of global context arrival is an essential component. In object categorization, a visual task with reported short latencies at 100ms, evidence for the influence of the magnocellular pathway in both monkeys [180] and humans [181] was reported. Here, we explore the possibility that BO assignment follows suit in the form of dorsal modulations.

Ventral feedback from IT with latencies at 80ms [182] was ruled out in [153]. Another possibility in the ventral stream is V4 that we examine here. We consider MT as a candidate dorsal area that provides context to BO cells in V1.

Angelucci *et al*. [183] studied the extent of feedback connections to V1 neurons and suggested that feedback from MT extends the surround to as large as 25×RF size, an area larger than 10degs for V1 neurons with 1deg RF size. At the time of writing this dissertation, we did not find a study that had directly investigated the extent of feedback from V4 to V1, similar to the study conducted for MT, V2, and V3 feedback to V1 by Angelucci *et al*. [183, 102]. Therefore, we assume an extent of feedback for V4 neurons similar to that of V3 cells [5] and at 10×RF size, nearly covering a 10deg area. With MT and V4 providing context from distant locations in the visual field, both areas could provide global context to BO cells in V1. In other words, both MT and V4 realize one of the two properties mentioned above, leaving the latencies to be examined.

Comparing response latencies in different visual areas is challenging when those time courses are reported in different individuals and with various experimental settings and stimuli. Despite the challenge, short latencies of some cortical areas such as MT, MST, and FEF were confirmed

[4]Given the experimental settings that were employed in [11, 14, 160] and that our focus was on explaining the observations of these studies, we restricted the possibilities to feedback from V4 and early dorsal recurrence from MT. In other experimental settings, other possibilities such as feedback from FEF could be explored.

[5]V3 is hierarchically between V2 and V4 [15] and therefore, we are assuming that the extent of V4 cells is at least as large as that of V3 neurons.

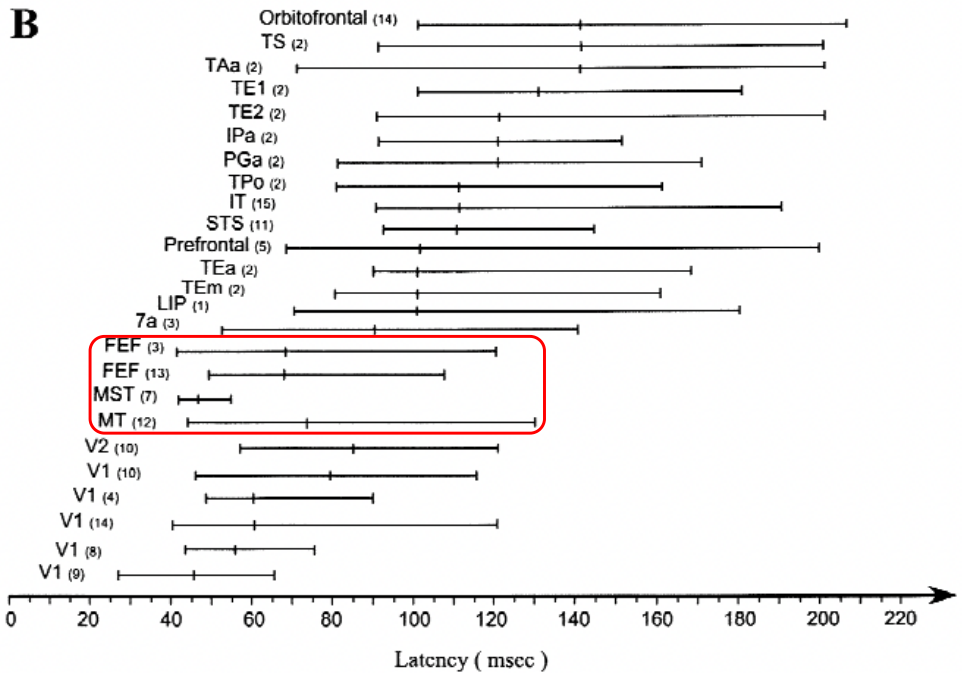


Figure 3.5: Response latencies reported in various visual areas. Latencies of areas MT, MST and FEF (two different studies) are highlighted with a red square. The reported latencies for these areas are shorter than 50ms. Figure adapted from [15].

in various studies [15, 2, 184]. Figure 3.5 depicts the latencies of responses in various visual areas adapted from [15], demonstrating latencies shorter than 50ms for these areas. Bullier [15] called these areas the "fast brain" and stated that they "belong to the dorsal stream or its continuation in the frontal cortex" with activations "sufficiently early to influence neurons in areas V1 and V2", supporting our dorsal modulations hypothesis.

In examining the latencies of MT and V4 neurons, two studies [1, 2] with a direct comparison of neuron responses in these areas are helpful. Maunsell [1] investigated the properties of neurons in V1, V4, and MT and suggested distinct parvocellular and magnocellular pathways across the visual cortex. The plots in Figure 3.6 demonstrating population responses of V1, MT, and V4 neurons highlight the latencies in these areas. The dotted line in these plots at 28ms indicates the earliest V1 response for easier comparison. These plots demonstrate shorter latencies of MT neurons than V4 cells (39ms versus 50ms for the beginning of response after stimulus onset). A similar time course in MT neurons with mean latency at 40ms was reported in [179].

Another similar study by Schmolesky *et al.* [2] compared latencies in several visual areas in

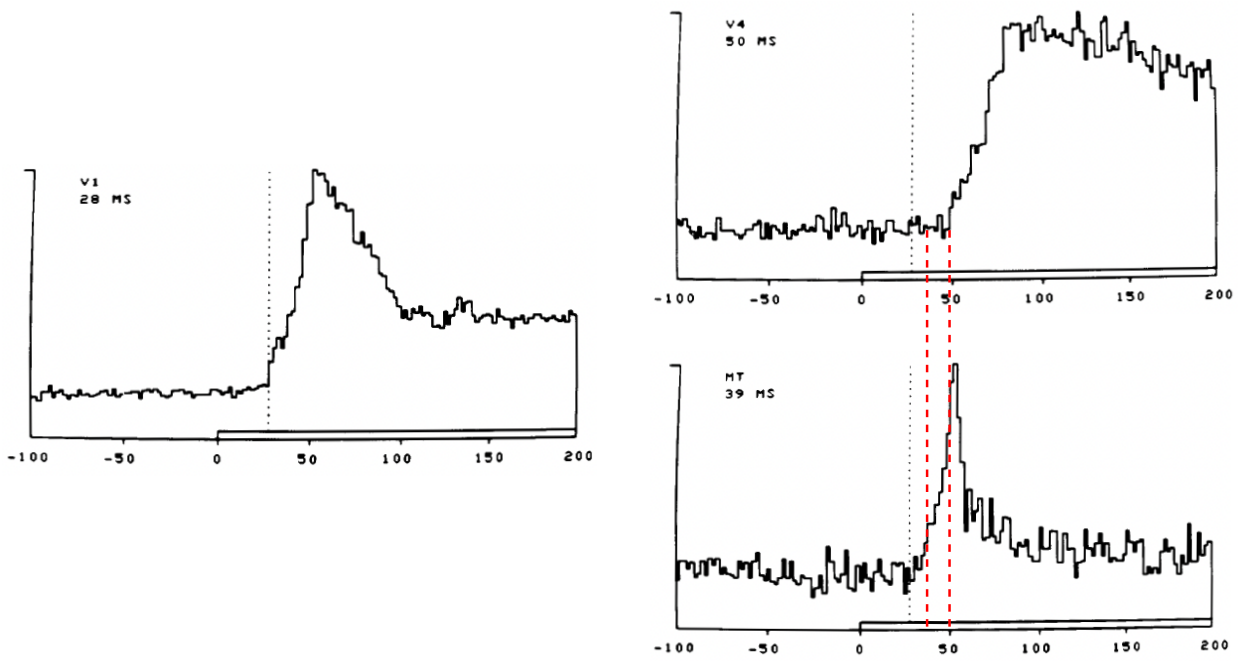


Figure 3.6: Response latencies in V1, V4 and MT adapted from [1]. The dotted line in each plot shows the early V1 response at 28ms for easy comparison. In the two plots on the right for V4 and MT latencies, the early MT and V4 responses at 39ms and 50ms respectively are highlighted with dashed red lines. These plots show shorter latencies of MT neurons compared to V4 cells.

anesthetized monkeys, "often in the same monkey". Schall [184] provided an update to the latencies reported in [2] with some of the responses measured during visual task performance. The latencies in our areas of interest, namely V1, V4, and MT, were in anesthetized monkeys in both [2] and [184]. Therefore, we re-plotted the latencies diagram from [2] in Figure 3.7. In this figure, we removed the latencies from all other areas reported in [2] and retained those of V1, V4, and MT. This figure demonstrates that the earliest V4 response emerges after more than 75% of V1 neurons have responded at about 72ms post-stimulus onset. In contrast, more than 50% of MT units respond at about 72ms, and their latencies more closely follow that of V1 cells. As a summary, shorter latencies of MT neurons compared to V4 cells in in [2] and [1] provides support in favor of the dorsal hypothesis.

The reported latencies in MT and V4, as summarized in Table 3.1 and can be seen from Figures 3.6 and 3.7, differ between the two studies [2] and [1]. Schmolesky *et al.* [2] suggested that these differences could be due to "lack of anesthesia and/or difference in stimulus presentation and data analysis". The point here is not the difference in latencies between these two studies but how

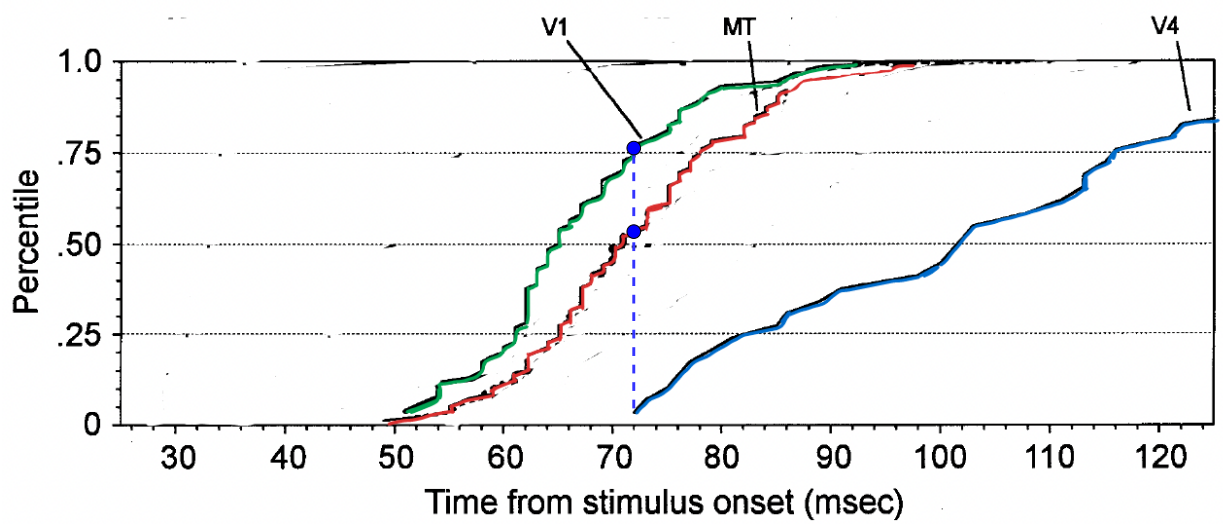


Figure 3.7: Response latencies of neurons in V1, V4, and MT adapted from [2]. We removed latencies of all other areas and retained those of V1, V4 and MT as these areas are of interest in the present work. These latencies in anesthetized monkeys are longer than those reported in [1]. However, similar to those reported by Maunsell [1], MT neurons have much shorter latencies compared to V4 with latencies comparable to V1 cell. Note that the dashed blue line highlights the earliest reported latency in V4 at about 72ms. More than 75% of V1 and 50% of MT neurons respond earlier than 72ms.

Study	V1 (mean)	MT (mean)	V4 (mean)	MT (mean) - V4 (mean)
Maunsell [1]	28ms	39ms	50ms	11ms
Schmolesky *et al.* [2]	66ms	72ms	104ms	32ms

Table 3.1: Response latencies reported by Maunsell [1] and Schmolesky *et al.* [2]. The last column indicates the difference of mean latencies between MT and V4. According to this column, MT latencies are at least 11ms shorter than V4.

latencies in V4 and MT compare to that of V1 and if neurons in these areas can give a timely feedback to BO cells in V1. Latencies from both studies suggest MT cells are fast with at least 11ms shorter latencies compare to V4 (rightmost column in Table 3.1). Moreover, MT neurons have incredibly short MT-to-V1 conduction of 1.3ms. Therefore, considering the mean latency of 40ms in responses, feedback from MT to V1 arrives at 41.3ms, suggesting that these neurons could be the cause of early divergence in BO responses.

Even considering the earliest reported response that we could find in V4 studies at 40ms [185] and assuming V4-to-V1 conduction time to be the same as V4-to-V2 time at 6ms [160] result in arrival of feedback from V4 at 46ms, well past the 42ms of edge onset in BO neurons. Note that the V4 latencies reported in [185] were for neurons exhibiting selectivity to convexities and concavities

of figure parts. This kind of selectivity on its own requires context or similarly, knowledge of figure side. From where in the brain such contextual information is provided to V4 shape-selective neurons at such extremely short latencies, even shorter than V1 edge response onset, is a question for further investigation and is not the focus of this work. Here, we remain with the suggestion that V4 feedback is not fast enough to provide *the context that causes the early divergence of BO responses*.

To test the plausibility of MT initiating the contextual divergence of BO responses in V1, we propose the Recurrent Border Ownership (RBO) model based on early recurrence from the dorsal stream combined with lateral modulations within the ventral pathway. Our hierarchical network of neurons is explained in detail in the next section.

3.3 The Recurrent Border Ownership Network

The RBO network, whose architecture is demonstrated in Figure 3.8 meets known biological properties with its parameters set according to neurophysiological findings. Although our model, due to its hierarchical architecture and convolutional processing, might look similar to a CNN [186], the similarities end there; there is no learning.

In this chapter, the following abbreviations are employed: mvSimple and mvComplex for model ventral simple and complex cells respectively, and mdSimple for model dorsal simple cells.

As Figure 3.8 displays, the RBO model is a two-stream network, implementing the ventral and dorsal pathways in the brain, with their signals meeting at border ownership cells. Each layer in the model simulate neurons at four scales. The model ventral stream implements simple, complex and BO cells in V1. In this network, mvSimple neurons are selective to borders and edges, comprising four different selectivity combinations shown in Figure 3.9a, and complex cells integrate responses of simple cells [50, 187, 188, 189]. The final population of border ownership neurons in the RBO will consists of types 1 and 3 BO cells as classified by Zhou *et al.* [11]. That is, neurons selective to side-of-figure and not contrast polarity (type 1) and those selective to both features (type 3).

The dorsal pathway implements model dorsal simple cells in V1 and MT neurons. The dorsal pathway is well-known to be selective to spatiotemporal features. However, a number of neurophysiological studies provided evidence that MT also responds to stationary stimuli [93, 94, 95, 17].

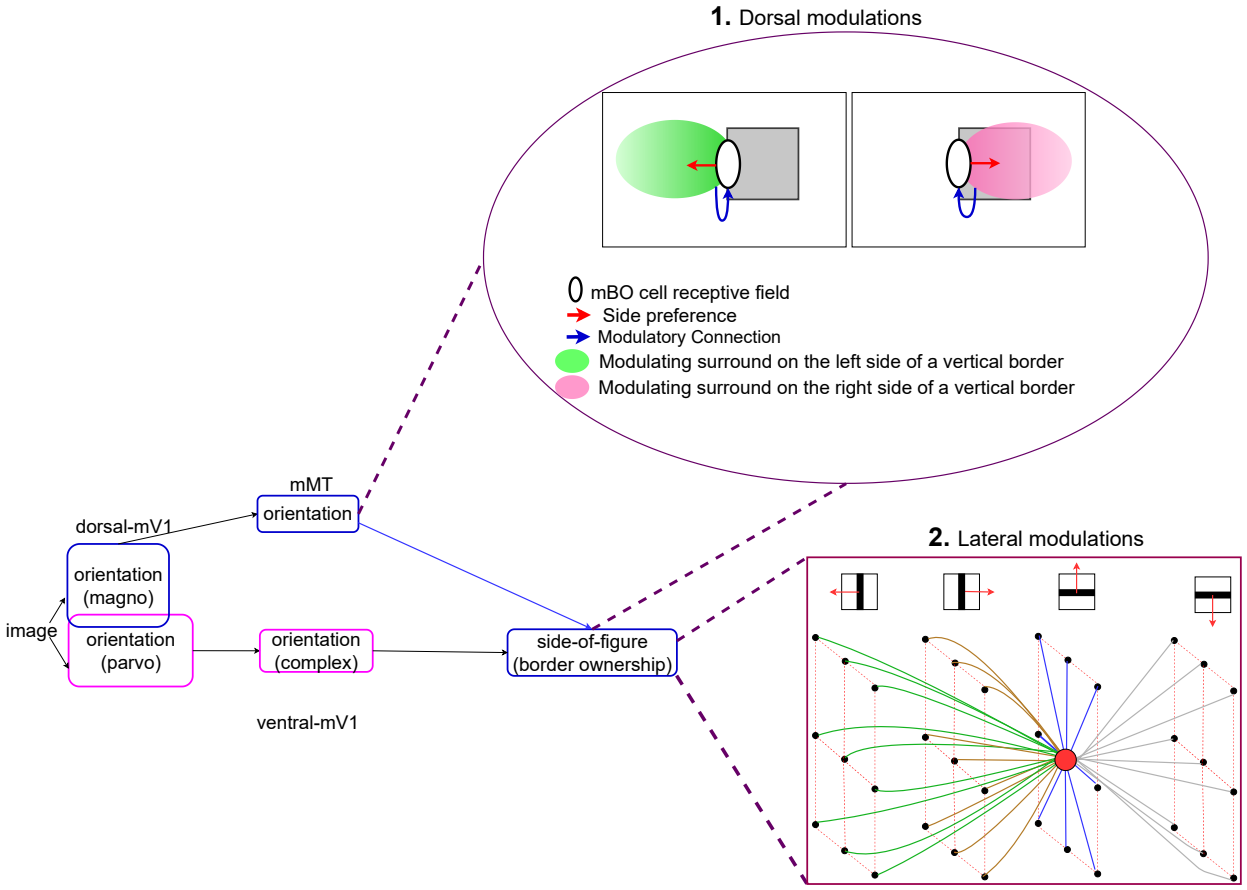


Figure 3.8: Border ownership network architecture. Magno and parvo orientation-selective neurons in the RBO distinguish the two magnocellular and parvocellular channels in this architecture. RBO neurons are implemented at four scales. Each scale defines an independent path from that of other scales. Border ownership neurons receive a feedforward signal from complex cells, and then, mBO responses are computed in two steps: 1. Dorsal modulations from mMT neurons (blue arrows), 2. Lateral modulations within the ventral stream. Dorsal modulations initiate the difference of responses to figure on either side of the border. The big ellipse shows a large view of dorsal modulations. The same stimulus (a single dark square) is shown to mBO neurons with opposite side-of-figure preferences. Red arrows indicate the side preference for each mBO cell. The neurons with BO preference to figure on the left of a vertical border are modulated by mMT cells whose RFs span the green area on the left of the border. In contrast, mBO neurons with preference to figure on the right of a vertical border are modulated by mMT neurons whose RFs span the pink surround area. Facilitatory dorsal modulations are computed as a Gaussian weighted sum of mMT responses, indicated using a gradient filling in each surround area. This is the first step in BO response computation (indicated by number "1" in the figure.) With scale selection performed in the last layer, lateral modulations begin between spatially neighboring neurons within a map as well as those across maps of similar local feature selectivities as the example in the rectangular box shows. Lateral modulations are both facilitatory and suppressive on either side of the border for each mBO cell and comprise the second step in BO response computation (indicated by number "2" in the figure.)

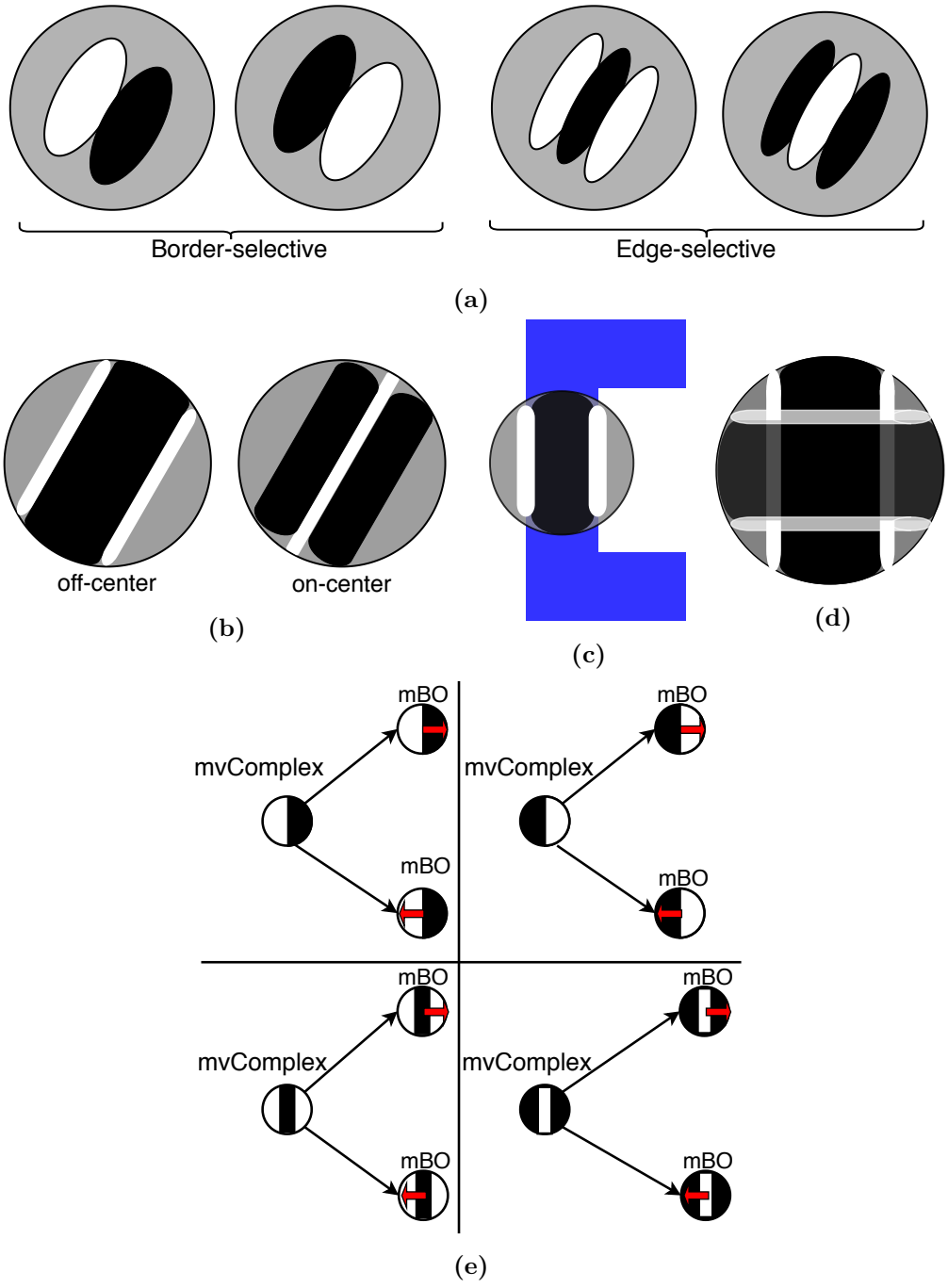


Figure 3.9: Examples of selectivities in model neurons. (a) Two types of selectivities to borders between solid dark-light regions and edges in mvSimple neurons are implemented. These selectivities define four combinations of selectivities in mvSimple cells. The border-selective mvSimple cells are implemented using Gabor filters and responses of edge-selective mvSimple neurons are obtained by difference of Gaussians filters following [16]. (b) Off- and on-center mMT selectivities. Following [17], mMT cells are selective to narrow bars that span their RFs. (c) An example of an off-center mMT on a C-shape stimulus. The off-center mMT cell is strongly activated by the pair of borders from the vertical part of the C shape.

Figure 3.9: (Previous page.) (d) Superimposed RFs of two off-center mMT neurons with horizontal and vertical orientations. A strong sum of responses of these neurons signals the existence of a closed shape. (e) Examples showing each mvComplex selectivity combined with dorsal modulations yield two sets of mBO selectivities. Four local feature selectivities in mvComplex cells to vertical borders and edges are shown. Each mvComplex cell sends a feedforward signal to two mBO cells passing its local feature selectivities to the mBO neurons. Dorsal modulations result in an initial side-of-figure preference in each mBO, which is shown by a red arrow.

As such, we skip the temporal aspect of dorsal responses as the RBO inputs are stationary images. Similar to mvSimple neurons, our mdSimple cells are selective to borders and edges with larger RF sizes than those of the ventral stream [190]. The mdSimple neurons send a feedforward signal to "on-center" and "off-center" mMT neurons with elongated excitatory and inhibitory regions following [17]. Figure 3.9b depicts examples of mMT RFs. While the on-center mMT responses signify the existence of edges, off-center mMT neurons indicate a pair of C-shape edges on either side of their RFs. Figure 3.9c illustrates an example of an off-center mMT cell with a pair of edges on either side of its RF, which result in strong activation of this neuron. Consequently, the sum of off-center mMT neurons at all orientations will be large for closed shapes, a key feature for border ownership assignment (See Figure 3.9d for an example).

In the RBO, mBO responses are computed in two steps:

1. Each mBO neuron receives a feedforward signal from mvComplex cells and is dorsally modulated by mMT cells on its preferred ownership side;

2. Scale selection [191] on mBO neurons with same selectivities is performed in the last layer of the RBO and lateral modulations in local neighborhoods is performed.

Dorsal modulations from one side of the border infuse side-of-figure preference to mBO cells, resulting in opposite mBO selectivities, as illustrated in Figure 3.9e. A larger view of dorsal modulations is shown in Figure 3.8, where mMT cells on the preferred side signal the existence of figure-like features, with modulatory strength proportional to their RF distances to that of the mBO cell. One can view such dorsal modulations as a recurrent loop between mV1 and mMT, albeit with feedback projection to the ventral mV1 rather than dorsal.

After dorsal modulations, scale selection [191] is performed in the last layer of the RBO, combining ownership signals from mBO neurons with identical selectivities across scales. Finally, mBO

neurons in local neighborhoods perform lateral modulations, as shown in the larger view rectangle in Figure 3.8. Specifically, each mBO neuron's response is iteratively updated based on its neighboring mBO activations. These facilitatory/suppressive lateral updates achieve a consistent mBO response pattern in local neighborhoods. Moreover, lateral modulations provide context from outside the classical RF, extending it to a slightly larger area called the near surround [190].

In the RBO, model neurons are implemented at horizontal and vertical orientations. In addition, all model receptive fields are square-shaped, even though round shapes might be used in some of the figures in this document in a figurative manner. Below, the computational details are explained.

3.3.1 Model ventral simple and complex cells

Following Rodriguez-Sanchez *et al.* [16], responses of our edge-selective mvSimple cells are computed using difference of Gaussians (DoG) with the same parameter settings. In particular, they used:

$$\begin{aligned} g(x,y) &= \frac{1}{2\pi\sigma_{x_1}\sigma_y}\exp\left(-\frac{1}{2}(\left(\frac{x'}{\sigma_{x_1}}\right)^2+\left(\frac{y'}{\sigma_y}\right))\right) \\ &\quad -\frac{1}{2\pi\sigma_{x_2}\sigma_y}\exp\left(-\frac{1}{2}(\left(\frac{x'}{\sigma_{x_2}}\right)^2+\left(\frac{y'}{\sigma_y}\right))\right) \\ x' &= x\cos(\theta)+y\sin(\theta), \qquad (3.1) \\ y' &= -x\sin(\theta)+y\cos(\theta) \qquad (3.2) \end{aligned}$$

where σ_y is the height and $\sigma_{x_1}, \sigma_{x_2}$ are the width of the Gaussians and θ is their orientations. Border-selective mvSimple cell responses were obtained using Gabor filters,

$$\begin{aligned} g(x,y;\lambda,\theta,\psi,\sigma,r) &= \exp\left(-\frac{x'^2+r^2y'^2}{2\sigma^2}\right) \\ &\quad \times\cos\left(2\pi\frac{x'}{\lambda}+\psi\right), \qquad (3.3) \\ x' &= x\cos(\theta)+y\sin(\theta), \qquad (3.4) \\ y' &= -x\sin(\theta)+y\cos(\theta) \qquad (3.5) \end{aligned}$$

with $r=0.5, \lambda=0.2$, $\psi=\pm\frac{\pi}{2}$ for two settings of contrast polarity, and $\sigma=\frac{\mathrm{RF}}{4}$. We set the mvSimple cell receptive fields at four scales to $\{0.4^\circ, 0.6^\circ, 0.8^\circ, 1^\circ\}$ according to [190, 192]. Model complex cells were computed by a Gaussian-weighted sum of mvSimple cells [16]. Both types of

cells are half-wave rectified.

3.3.2 Model dorsal simple cells

Our mdSimple cell RF sizes, set as $\{0.9°, 1.33°, 1.76°, 2.2°\}$, are larger than those of mvSimple cells [190]. In our implementation, edge-selective mdSimple neuron parameters are: $\sigma_y = \text{RF}, WR = 2.5, AR = WR * [10, 9, 8, 7]$, where WR is the width ratio and AR the aspect ratio of the DoG kernel (See [16] for details). For border-selective mdSimple cells, the difference of two Gaussians with $\sigma_y = \text{RF}, \sigma_x = \frac{\sigma_y}{AR}$ were employed.

3.3.3 Model MT cells

We set mMT neuron RF sizes based on the findings of Fiorani *et al.* [193], to $\{2.5°, 3.26°, 4.02°, 4.78°\}$. Felleman and Kass [17] found that MT cells, with on and off excitatory regions, showed an increase of responses to bars with the length up to the receptive field size. However, the effective bar width is $\frac{1}{10}$-th of the receptive field size. In fact, for MT neurons with 4° receptive fields, the most effective bar width was about 0.25°. In other words, MT cells are selective to long narrow bars, with length matching the receptive field size. The receptive field of an example model MT cell with such selectivities and an excitatory region in the middle is depicted in Figure 3.9b. Such a receptive field profile is not unusual in the visual system as a similar one was suggested in cat simple cells by Hubel [194] (their "SIMPLE CORTICAL CELLS" figure, subfigures a and b). We called these MT cells with the on area in the middle as "on-center" MT neurons and implemented these cells using the Difference of Gaussians formulation. We also implemented MT cells with an inhibitory region in the middle and two excitatory areas on the sides, similar to Hubel's subfigure f in his "SIMPLE CORTICAL CELLS" figure. An example of a model off-center MT neuron receptive field is shown in Figure 3.9b. Note that these neurons exhibit selectivity to long and narrow bars on either side of the receptive field with strong activations when two bars are present on both excitatory areas. We refer to these neurons as "off-center" MT cells. In the RBO, DoGs with $\sigma_y = \text{RF}$, $AR = [33, 52, 80, 126.6]$, and $WR = [3.3, 5.2, 8, 12.6]$ for the four scales of RFs implement on- and off-center mMT cells.

Various studies of the Magnocellular pathway [195, 196], and MT recordings of macaque [197] and humans [198], suggested that these neurons are highly sensitive to contrast, whereas this

sensitivity in V1 neurons is much lower. Moreover, MT responses are saturated at contrasts higher than 1.6% [198]. Therefore, dorsal responses, R are rectified with:

$$\Phi = \frac{1 - e^{-R/\rho}}{1 + \frac{1}{\Gamma}e^{-R/\rho}}. \tag{3.6}$$

We set $\Gamma = 0.001, \rho = 0.02$ to enforce low contrast sensitivity.

3.3.4 Model border ownership cells

3.3.4.1 Dorsal modulations

As shown in Figure 3.9e, each mvComplex neuron sends signal to two set of mBO cells. Let mvC(x, y, θ) denote a mvComplex neuron with its RF centered at (x, y) and selective to borders/edges at orientation θ. Then, we indicate the two mBO cells with opposite side-of-figure preferences that receive a feedforward signal from mvC(x, y, θ) as mBO$(x, y, \theta + \frac{\pi}{2})$ and mBO$(x, y, \theta - \frac{\pi}{2})$. The two parameters $\theta + \frac{\pi}{2}$ and $\theta - \frac{\pi}{2}$ denote the direction of ownership for a border with orientation at θ angle (the direction of the arrows in Figure 3.9e).

The initial border ownership response is determined by:

$$\begin{aligned} &R_{\text{mBO}}(x, y, \theta \pm \frac{\pi}{2}, t = 0) = R_{\text{mvC}}(x, y, \theta) \times \\ &[\sum_{d_1, d_2} w(\pm d_1, \pm d_2)(\sum_{\phi} R_{\text{mMT\_ON}}(x, y, \phi, \pm d_1) \\ &+ \sum_{\phi} R_{\text{mMT\_OFF}}(x, y, \phi, \pm d_2))] \end{aligned} \tag{3.7}$$

where t, set to 0 in this equation corresponding to edge onset latency, represents the time, x, y specify the RF center, R denotes neuron responses with the subscript indicating the model neuron type. In this equation, mvC responses are multiplicatively modulated by on-center and off-center mMT cells, represented as mMT_ON and mMT_OFF respectively. The responses of orientation-selective mMT cells are summed to account for all the possible orientations, ϕ, of the figure contour segment opposite to the occlusion boundary. This summation will peak in the presence of a closed figure on the preferred side, hence, resulting in a strong modulatory signal to mBO cells. In Equation 3.7, d_1, d_2 determine the extent of surround from mMT cells providing context to mBO

neurons, their sign specifying the modulatory surround direction. The linear weighting function $w(\cdot,\cdot)$ assigns larger weights for mMT cells closer to mBO neuron. Note that each modulation term obtained by expanding the outer summation in Equation 3.7 ($\sum_{d_1,d_2}$) is a manifestation of the isotropic inhibition proposed by [199] since each term is a weighted multiplicative process with weights set to 1 for all orientations.

In Equation 3.7, d_1 and d_2 determine the extent of surround for mBO neurons. Shushruth *et al*. [200] discovered that the far surround in V1 and V2 could exceed 12.5°, and [102] observed that long-distance interactions provided by feedback from MT to V1 could extend the surround to 13 times the size of V1 RFs. According to [200, 102], we set the extent of surround for mBO neurons provided by mMT cells to a maximum of 9°, 13 times the average mV1 receptive field size.

After dorsal modulations, in the output layer of the RBO and at each visual field location, the network has a total of 16 mBO neurons (2 (side-of-figure preferences) × 4 (local feature selectivities) × 2 (orientations)).

3.3.4.2 Lateral modulations

Lateral modulations in our model was achieved by a few iterations of relaxation labeling (RL) [201, 202], which provide local context and ensure smoothness in responses. At each visual field location, responses of the 16 mBO neurons determine the border ownership. However, inconsistencies may arise in local neighborhoods due to uncertainty or noise. RL is a framework that deals with uncertainties by iterative interactions that take place between neighboring neurons in a cooperative or competing fashion. Specifically, through RL iterations, compatible ownership responses strengthen one another, and weaken incompatible ones in their neighboring neurons. In our implementation, the different selectivities of mBO cells define the set of labels for RL, and the normalized mBO responses after dorsal modulations from Equation 3.7 are assigned as the initial probability for each label at each visual field location. The connections between neighboring neurons define the local neighborhood providing context to each mBO cell. A larger view of a local neighborhood example in our model is demonstrated in Figure 3.8. In RL, the compatibility function between the set of labels defines the influence of neighboring neurons on a label at each location. For example, matching labels have facilitatory effect on one another while opposing labels are suppressive. In the RBO network, the facilitatory effect is imposed using a Gaussian function with a sharp fall

to highly reward matching labels. The penalty for opposing labels was determined using a linear function.

At each RL iteration, the border ownership responses are updated as below:

$$\begin{aligned} R_{\mathrm{mBO}}(x, y, \theta+\beta, t+1) = \quad & (1 + P(x, y, \theta+\beta, t)) \\ & \times R_{\mathrm{mBO}}(x, y, \theta+\beta, t), \quad \text{for } t = 1, \ldots, 10 \end{aligned} \tag{3.8}$$

where $R_{\mathrm{mBO}}(.,.,.,t)$ is the response at time t, $\theta+\beta$ with $\beta \in \{+\frac{\pi}{2}, -\frac{\pi}{2}\}$ denotes the ownership direction, and x, y signify the RF center, P in the range $[-0.5, 0.5]$ represents the label confidence obtained by integrating compatibilities with immediate neighbors as explained above at time t. In particular, P is computed as:

$$P(x, y, \theta+\beta, t) = \sum_{(a,b)\in \mathrm{neighbor}(x,y)} \sum_{\alpha \in A} (c(\theta+\beta, \alpha) P(a, b, \alpha, t)), \tag{3.9}$$

where neighbor(x, y) denotes the set of visual field locations in the local neighborhood of the cell at (x, y), A is the set of modeled BO preferences (the eight preferences demonstrated in Figure 3.9e), and $c(\theta+\beta, \alpha)$ denotes the compatibility function between BO preferences $\theta+\beta$ and α.

In the RBO, dorsal modulations occur at the onset of edge response followed by iterations of lateral modulations. To determine the number of iterations for RL, one must decide how far local context can travel within the visual field. Sugihara *et al.* [160] found that lateral propagation in BO neurons is not fast enough to travel from one side of a large figure all the way to its opposite side, which suggests that the impact is localized and not global. Therefore, we limit RL iterations in Equation 3.8 to 10 (equivalent to 10 synaptic connections) extending the near surround to about 2 times the RF size [183]. This choice was also made based on the time course of border ownership processing with respect to the onset of responses reported by [11]. A detailed analysis on the time course of lateral modulations in our model is provided in Section 3.4.4.

3.4 Evaluation

3.4.1 Stimuli

All the stimuli employed for comparison with BO cells shown in Figures 3.10 and 3.11 were generated according to the stimuli employed by Zhou *et al*. [11] with input set to 400×400 pixels. We measured 1° visual angle at 50cm to be 32×32 pixels for our stimuli size.

In generating the randomly overlapping polygons dataset (ROPD), we kept the input size at 400×400 pixels. While the stimuli in cell recordings are controlled in figure size, placement and intensities, in ROPD, these features are randomly set to add a complicated flavor to the stimuli. Additionally, we did not limit the number of figures in each image to one or two, but varying between 1 and 8 figures. Each set of stimuli in ROPD with a fixed number of figures contains 100 images, summing to a total of 800 images.

3.4.2 Quantification

For quantitative evaluations, we define normalized difference of responses (NDR) as

$$D = \frac{R_{\text{preferred}} - R_{\text{non\_preferred}}}{\max(R_{\text{preferred}}, R_{\text{non\_preferred}})}. \tag{3.10}$$

The normalized difference D, a value in the range $[-1, 1]$, gives a common ground to compare the strength of the BO signal differences to preferred and non-preferred stimuli. This measure can also be interpreted as the confidence of a mBO cell in its ownership assignment. The closer to 1 the NDR, the higher the confidence of the neuron in the ownership assignment. Therefore, this measure can be employed to compare the ownership assignment confidences in mBO cells versus that of BO neurons, and equivalently, compare the assignment confidence before and after RL. Zhou *et al*. [11] performed an analysis on the ratio of responses to non-preferred over preferred stimuli, where larger values signify uncertainty in the neuron's ownership assignment. Normalized differences are equal to (1 − response ratios).

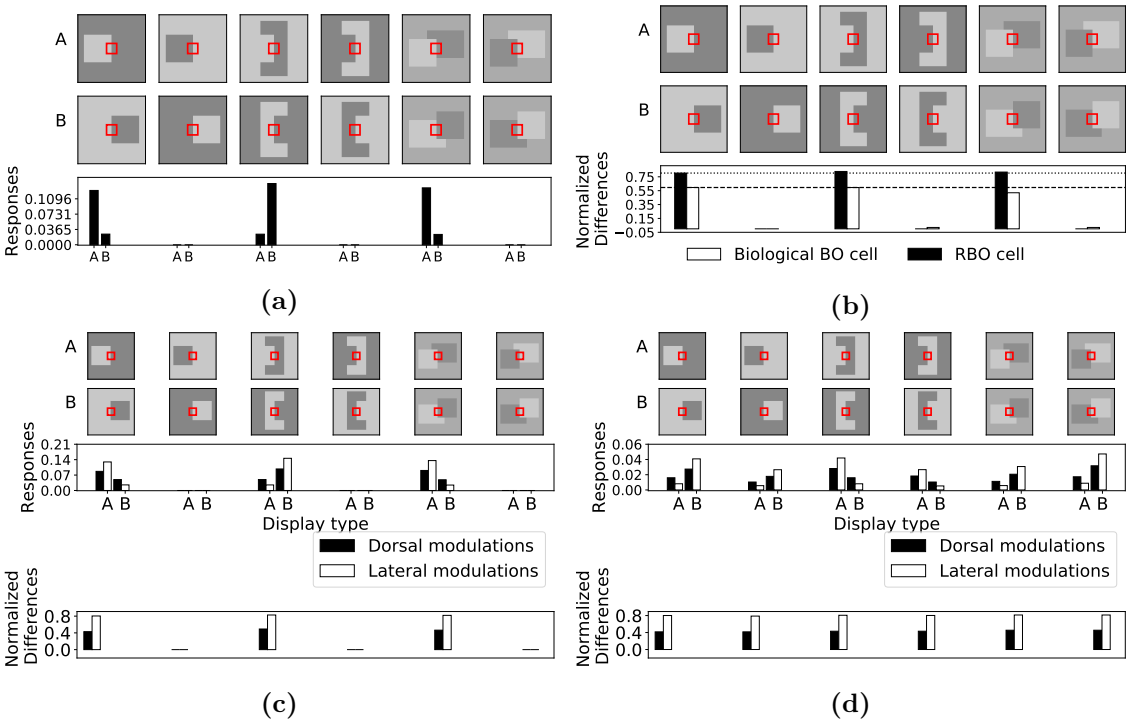


Figure 3.10: Example responses of a mBO cell responses and comparison with BO cell responses. (a) Example responses of a mBO cell with the red rectangle representing its classical RF. Comparing these responses with the responses of the BO cell shown in Figure 3.1b reveals a similar pattern in exhibiting the difference of responses. (b) Normalized difference of responses for the mBO cell which responses were depicted in Figure 3.10a and the BO cell with responses illustrated in Figure 3.1b show that both mBO and BO cells consistently maintain the difference of responses across all stimuli. (c), (d) Examples of mBO responses for before and after lateral modulations. The bottom plot in each sub-figure represents the normalized differences for before and after lateral modulations.

3.4.3 Model Cell Activations Resemble Biological Responses

3.4.3.1 Difference of Responses

Presenting the stimuli employed in the biological study by Zhou *et al.* [11] to the RBO network, we recorded the responses of mBO neurons. The goal was twofold: first, to test whether dorsally-modulated neurons could exhibit the difference of responses similar to BO cells; and, second, to evaluate the role of lateral modulation in BO assignment. An example of responses of a mBO cell is shown in Figure 3.10a. As the plot shows, the responses of this mBO cell to figure on either side of the border differ in strength, with no responses when the local feature within the RF does not match the selectivity of the neuron. Comparing the responses of this cell to that of the BO neuron shown in Figure 3.1b demonstrates a similar pattern in ownership encoding. As a quantitative

comparison between the mBO and the BO responses, we computed the NDR to preferred and non-preferred stimuli for both cells, depicted in Figure 3.10b. We found that both the mBO and BO cells maintain the NDR, across simple and complicated overlapping figures.

Asking how much of response differences could be attributed to the early recurrence versus lateral modulations, we plotted mBO responses before and after lateral modulation. In the top row of Figure 3.10c, responses of the same neuron as in Figure 3.10a are depicted separately for the dorsally-modulated component and after lateral modulations. This plot clearly demonstrates that lateral modulations enhance the difference of responses. Another example is shown in the top row of Figure 3.10d, demonstrating the same pattern of enhancement in response differences due to lateral modulations (See the Appendix A for responses of all mBO cell types in the model). To quantify the effectiveness of lateral modulations, we computed the NDR before and after lateral modulations, shown in the bottom rows of Figure 3.10c and 3.10d. As these plots confirm, lateral modulations nearly double the response difference by facilitating the strength of initial dorsally-modulated responses to preferred stimuli while suppressing those to non-preferred stimuli.

3.4.3.2 Invariance

Examining invariance in mBO responses, we recorded the responses of mBO cells to changes in position, size and solid/outlined figures. Also, for quantitative comparisons, we computed the NDR in mBO and biological cells reported by Zhou *et al.* [11]. Invariance in mBO responses and comparison with BO cells are shown in top and bottom rows of panels a-c in Figure 3.11. Figure 3.11a illustrates position invariance in an mBO cell, with the bottom plot demonstrating that both mBO and BO cells remain invariant to border position up to a distance from the RF center. The difference in the biological cell is expanded to almost 3deg, whereas for the mBO neuron it is up to 1deg. This is not surprising since the available data was for a V2 BO cell, with possibly a larger RF than our mBO cell in V1. Size invariance in mBO neurons is shown in Figure 3.11b, with the NDR in the mBO cell closely following that of a BO cell reported in [11], shown in the bottom plot. Invariance to outlined/solid figures in a mBO cell is depicted in Figure 3.11c. Both mBO and BO cells preserved the NDR in the bottom plot.

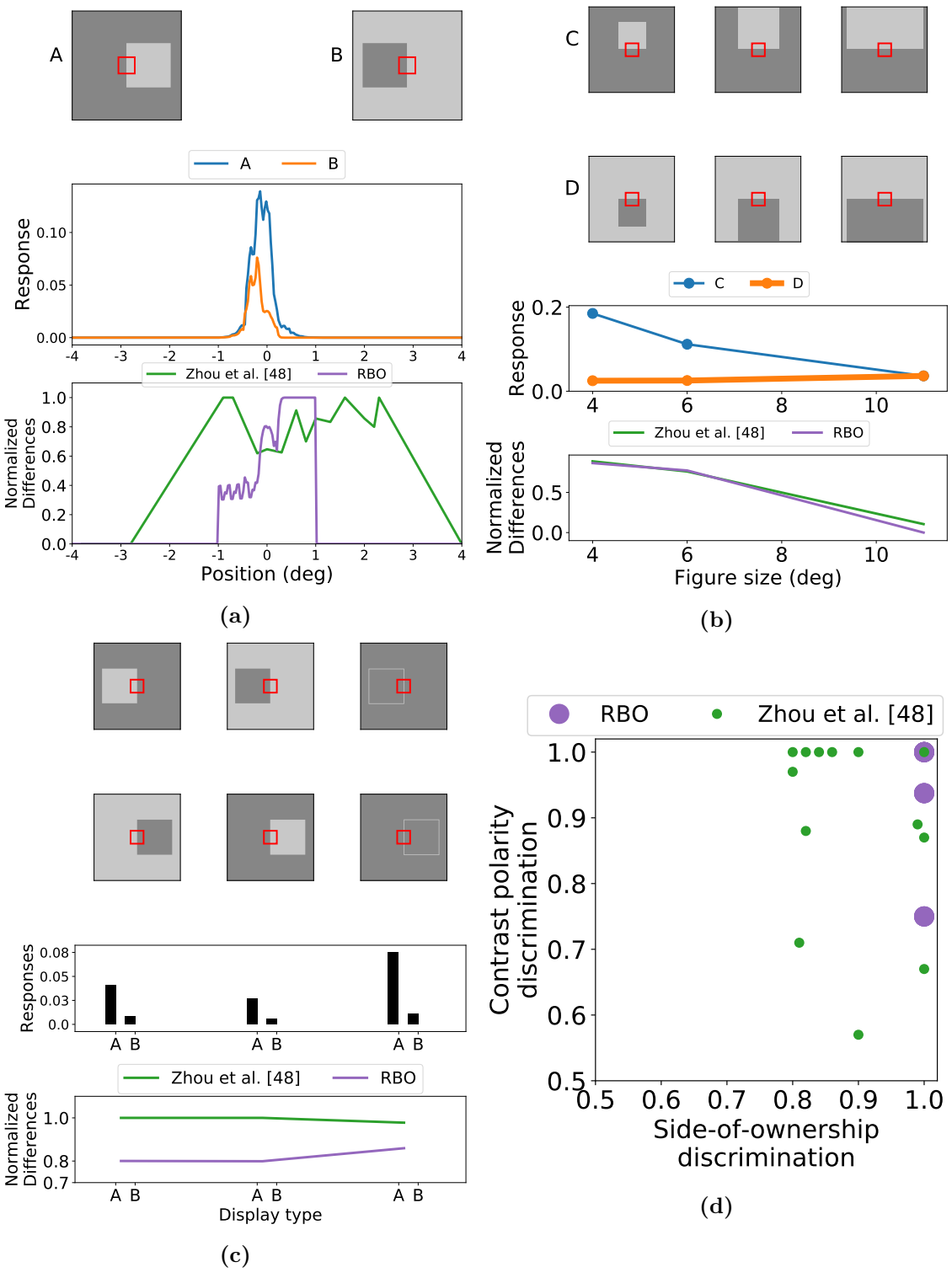


Figure 3.11: RBO response comparison with biological BO cells. In panels (a-c), showing invariance to position, size, and solid/outlined figures respectively, the top plot demonstrates mBO responses and the bottom plot compares the normalized difference of responses of the RBO and BO cells reported by Zhou *et al*. [11]. (d) Ownership/contrast accuracies compared between RBO and biological cells with greater than 80% accuracy.

3.4.3.3 Reliability

To measure the reliability of neural responses in signaling ownership/contrast polarity, Zhou *et al.* [11] introduced a decision model based on the responses of a BO cell to four stimuli combining ownership and contrast polarity possibilities. Then, from the first 1s of responses for each neuron, they computed the ownership/contrast decision by subtracting responses to stimuli with opposing features. The accuracy, computed based on a "correct" decision, represents the probability of a correct decision in the first 1s of responses. Similarly, we computed the accuracies of RBO responses based on responses before and after lateral modulations. Figure 3.11d illustrates that similar to BO neurons with 100% ownership accuracy where contrast accuracy varies between 65% and 100%, in the RBO, ownership accuracy is at 100% with contrast accuracy varying between 75% and 100%.

These results demonstrate that the RBO cells mirror the responses of BO neurons, suggesting that dorsal modulations could be the source providing context that cause the early response divergence in BO cells. Moreover, the initial divergence is further enhanced locally by means of lateral modulations.

3.4.4 Lateral Modulations Time Course

Craft *et al.* [18] provided a detailed review of conduction velocity and horizontal signal propagation latency. Following Craft *et al.*, we assumed $200\frac{\text{mm}}{\text{s}}$ conduction velocity and estimated the near surround at 2.5mm in radius. Therefor, the latency for lateral propagation from one side of the near surround to the RF center would be about $\frac{2.5}{200}$ ≈13ms. That is, after the early divergence induced by dorsal modulations, lateral modulations take about 13ms to enhance the BO signal, which is within the 10-25ms latency reported by Zhou *et al.* [11] for BO signal after the onset of responses. The temporal course of lateral modulations in our model is plotted in Figure 3.12 for a few stimuli. The green-shaded area represents the 13ms time interval for lateral modulations. This figure shows that the mBO signal becomes stable faster for smaller squares than for the larger ones (Compare red and brown curves in Figure 3.12). This behavior was also observed in biological BO neurons [14, 160]. Interestingly, this plot also reveals that the enhancing effect of lateral modulations are more significant for larger figures. In other words, while dorsal modulations initialize the side-of-figure preference to large figures, the difference in responses is established for

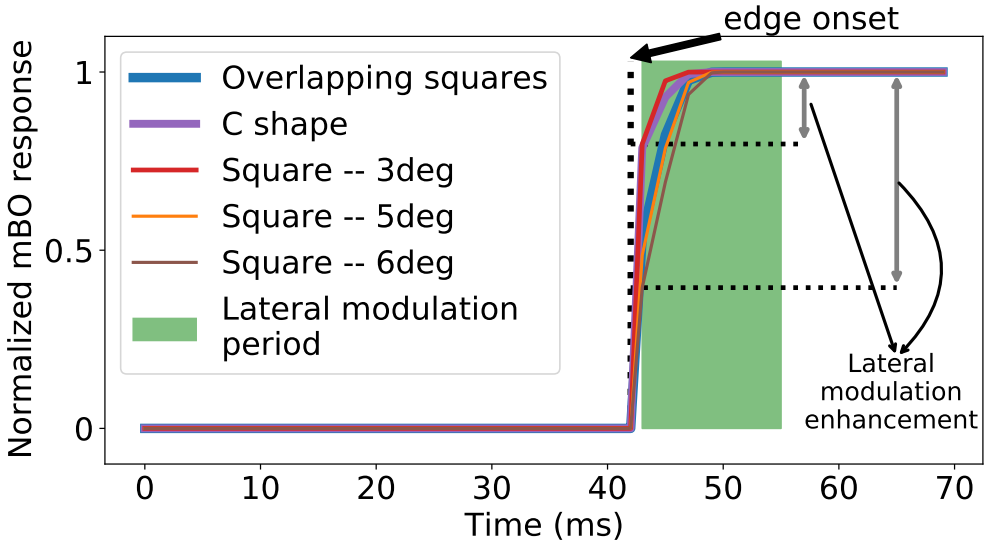


Figure 3.12: Lateral modulations time course. Dorsal modulations occur at edge onset and immediately after that, lateral modulation iterations begin for about 13ms (shown as the green-shaded area), extending the RF to two times the classical RF size. Note that in square examples of various sizes, lateral modulation enhancement gets stable faster for smaller squares than for larger squares. This behavior was also reported in biological BO neurons (See [14], their Figure 14).

mBO cells when a consensus in neighboring neurons is observed.

3.4.5 The Kanizsa's Square Explained

We tested the RBO performance on a few examples of overlapping shapes other than those employed in [11]. Moreover, we asked how the RBO perceives stimuli with illusory contours. If BO neurons assign ownership to occluding illusory shapes, then perhaps it could explain the perception of illusory contours in the brain as ownership assignment occurs very early in the ventral stream and such ownership signal is carried higher in the visual hierarchy and affects the processing in later areas.

Figures 3.13a and 3.13b show examples of RBO results on overlapping shapes with the vectorial modulation index (VMI) [18] along contours, where the direction and length of the vector represent the side and strength of ownership. These examples show that even with no priors on shape or T-junctions that were imposed in the model by Craft *et al.* [18], the RBO responses are in agreement with the perception of two overlapping shapes.

Examining the RBO perception of illusory contours, we presented the network with one, two and four Pac-Man figures gradually forming a Kanizsa's square. As the VMI vectors in Figures

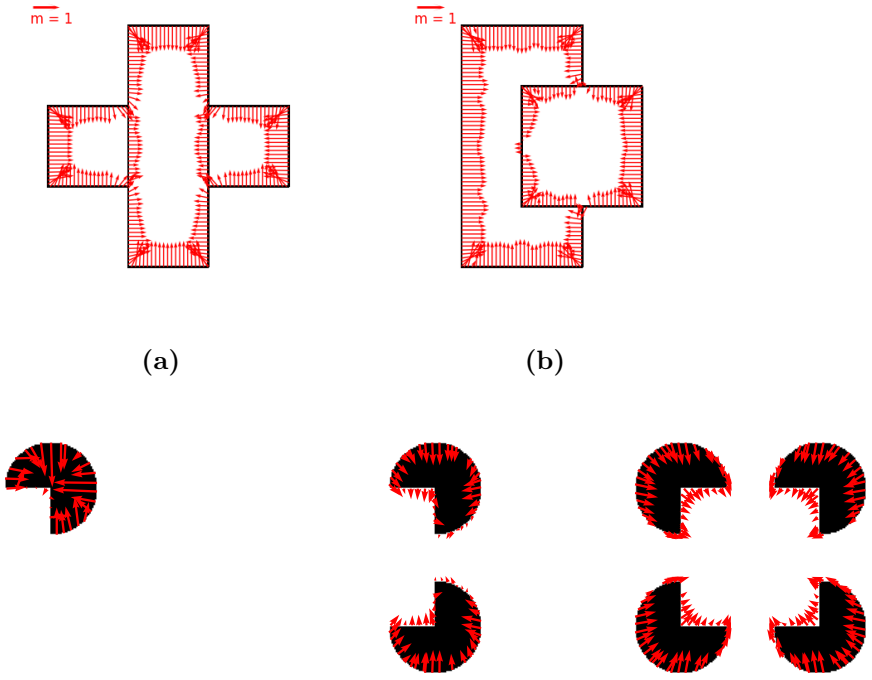


Figure 3.13: RBO responses to simple stimuli of overlapping shapes and illusory contours. (a), (b) Examples of the RBO performance on simple overlapping shapes with the VMI vectors along occluding contours. (c), (d), (e) Examples presenting the RBO performance to stimuli with one, two and four Pac-Mans. In (c), the ownership directions indicate the presence of a single figure, with a gradual shift to signaling a light bar and a light square occluding dark circular ground figures in (d) and (e).

3.13c, 3.13d, and 3.13e indicate, mBO cells signal the existence of a bar and a square in the middle and right examples as occluding figures. That is, the RBO network explains the illusory shape. Similar observations in BO cells were reported with ambiguous figure contours [154] or missing figure fragments [14], however, with stimuli that contained single figures. In our case, there are competing features along each border for the isolated Pac-Mans, suggesting that even in the presence of competing figures, mBO cells assign ownership to illusory occluding shapes; a suggestion that needs to be further investigated in BO neurons.

3.4.6 The Randomly Overlapping Polygons Dataset

We evaluated our model performance in terms of its accuracy in border ownership assignment on stimuli more complicated than the controlled synthetic stimuli often used in cell recordings. Therefore, we introduced the ROPD dataset, with randomly placed polygons of random sizes and intensities. The number of polygons in each image varies between 1 and 8, with 100 images for each fixed number of polygons. We also compared the accuracy and robustness of the RBO assignments against that of the model suggested by Craft *et al.* [18] (from here on called "Craft's" model). Models introduced in [177] and [178] are extensions of Craft's network and Craft's suggested mechanism is a benchmark model for BO assignment.

Craft's network takes binary edge and T-junction maps as input. Since this assumption does not hold in the biological visual system and for a fair comparison between the two models, we extended the implementation of our model simple and complex layers to Craft's model. We also implemented a layer of T-junction maps that fed to Craft's network. Craft *et al.* [18] set their mBO and grouping cell RFs to 0.5° and between 1° and 9°. Although 9° for RF size of V4 neurons, compared to the largest RF scale for our mMT cells at 4.78°, is very large, we kept their initial setting.

Figure 3.14a demonstrates examples from the ROPD along with the VMI vectors for both Craft's and RBO models. As the figure shows, the VMI vectors from Craft's model appear to be random in their direction and not collinear even along the same edge of a polygon. In contrast, the RBO vector directions match the ownership direction very well. Note that even without any priors such as those imposed in Craft's model signaling T-junctions, convexity or proximity, the overall perception of the RBO for the scene yields an accurate map of occlusion boundaries along

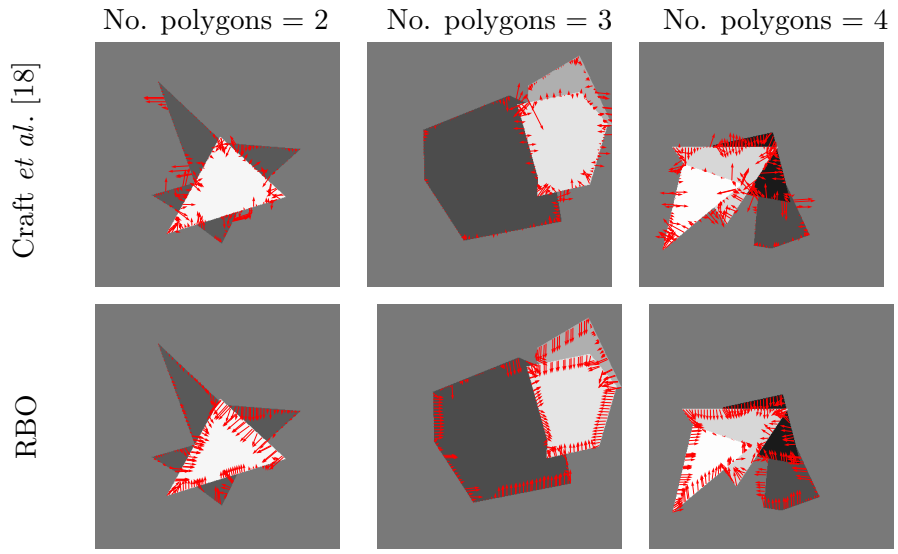
junctions and non-junction borders. We believe two key features of the RBO model yield its superior performance. First, the RBO VMI vectors are collinear along the occluding contours mainly due to lateral modulations in our model, which are lacking in Craft's model. Second, dorsal neurons are highly sensitive to contrast and can accurately signal the existence of important features even with low-contrast borders. This property especially helps with overlapping polygons with similar intensities.

For a quantitative performance comparison, we computed the average ownership assignment accuracies for each sets of images with a fixed number of polygons. The average accuracies of both models as a function of the number of polygons is illustrated in Figure 3.14b. The RBO model outperforms that of Craft's on all sets of images, with the overall average accuracy at 78% versus 70%, suggesting that the RBO network is both more accurate and more robust in border ownership assignment for simple and complicated scenes. Both models, however, experience a drop in performance with an increase in scene complexity. This drop in performance is not surprising in presence of competing features on both sides of a border resulting in smaller difference of responses. Similarly, Williford *et al.* [153] reported weaker BO signals to more complex stimuli of natural scenes.

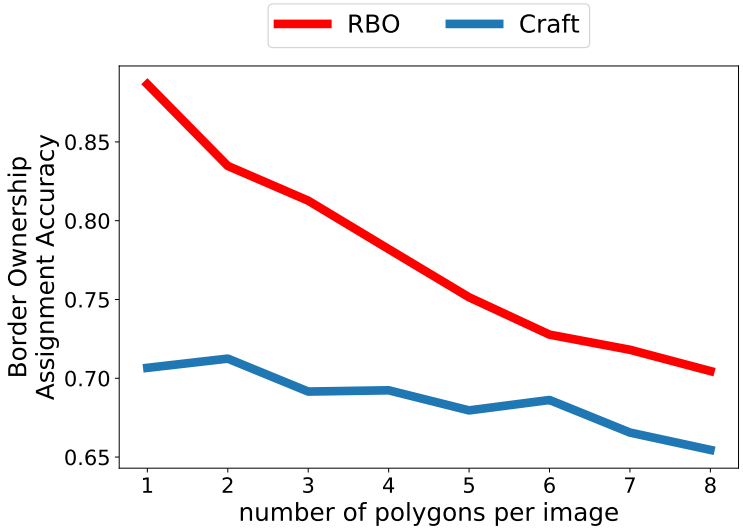
3.5 Discussion

Our goal was to suggest a mechanism for BO assignment that could explain the early response divergence in BO neurons, which happens before object classification. Consequently, we designed a hierarchical model inspired by the findings of the brain and set the properties of each set of neurons in the model accordingly. Our model maintains the ownership assignment confidence across simple and complicated stimuli similar to BO cells. Additionally, our experiment with illusory contours suggest that the perception of illusory contours in the brain could be explained by BO assignment along these contours. Although our experimental results attest to the plausibility of our hierarchical network for BO assignment, the RBO model lends detailed insight to the role each of dorsal and lateral modulations play in a quantified way.

We proposed dorsal modulations since fast MT cells with large RFs fit all the criteria to initiate the early response differences, while V4 latencies do not support their role in early response



(a)



(b)

Figure 3.14: Model performance comparison. (a) Examples of images from the ROPD with the VMI vectors obtained from the model suggested by [18] (top row) and the RBO (bottom row) on stimuli with 2, 3 and 4 overlapping polygons. BO assignment by Craft's model in the top row appear somewhat random compared to the collinear and more accurate RBO assignments. (b) Accuracies in border ownership assignment from both [18] and RBO model.

divergence. Feedback from V4, however, can further enhance the difference of responses in BO neurons. We acknowledge any doubts about our suggestion of V4 feedback to BO cells on the basis of comparing latencies from various studies. Our study puts forth additional possibilities and our findings suggest the need for further investigations with direct comparisons of latencies in MT, V4 and BO neurons in V1 to elucidate the cause of early divergence in BO responses.

Our dorsal-to-ventral modulations is not the first of such models. Previously, early recurrence showed improvements in low-level feature representation [203] and edge detection [199]. In a similar attempt, we demonstrated the role of such modulations in providing global information essential for even higher levels of abstraction, *i.e.*, BO assignment. Our experimental results suggest that although the dorsal stream is best known for motion, the coarse selectivity of its neurons to spatial features within their spatiotemporal selectivity [17] provides context to BO cells and contributes to shape processing in the ventral stream.

Our mMT neurons are modeled based on various studies of the Magnocellular pathway and area MT that found these neurons to be highly sensitive to contrast [195, 196, 197, 198], and their responses not so robust with iso-luminant color stimuli [204, 205]. These properties of MT cells open a venue for investigating MT contributions in ownership assignment. Namely, if MT cells provide context to BO neurons, then BO latencies are expected to increase with iso-luminant stimuli in absence of robust MT signals while such delays are not expected to appear with low-contrast stimuli.

Findings of Hupe *et al.* [100] could help design another testing paradigm. Specifically, they showed that inactivating MT would affect neurons in V1, where some cells were even entirely silenced. MT also significantly affected V1 responses *early on* in their time course [101]. The dorsal modulation hypothesis can be evaluated in a similar fashion by inactivating area MT and studying the latencies of BO cells. If MT provides context to BO responses, one would expect that lack of MT feedback would delay the divergence, increase the latencies, and perhaps decrease the strength of the BO signal.

Our model is the first to combine global and local context, which allowed us to investigate contributions of dorsal and lateral modulations in ownership assignment. Although dorsal modulations result in BO assignments, they do not warrant consistent assignments among neighboring cells. In contrast, lateral modulations ensure consistent labeling in local regions. Our results demonstrate

an enhancement in the mBO signal after lateral modulations; an effect that was consistent among all mBO cells. Lateral modulations provide local support with a lag to BO neurons, due to the time required for the signal to travel in a local neighborhood. In fact, our model of early recurrence followed by lateral modulations demystifies the roles of feedback and lateral modulations in BO encoding and explains previous observations that the signal from far fragments of a square figure arrived earlier than the signal from near corners of the square [14].

Our model might appear similar to the surround modulation mechanisms suggested by Sakai *et al*. [158], who showed that asymmetric surround regions could explain the BO responses. Despite similarities, the following differences are notable:

1. Sakai *et al*. focus on the organization of facilitatory/suppressive surround for BO cells while in the RBO, our focus was on the source of the nonclassical modulation,

2. In [158], surround regions are stochastically selected for each neuron; the RBO is deterministic,

3. The RBO distinguishes the contributions of each of far and near surround. Sakai *et al*. do not differentiate between those,

4. Whereas Sakai *et al*. found robust BO signals from neurons with iso-orientation suppression and cross-orientation facilitation regions, we did not restrict suppression/facilitation to certain orientations. Instead, all orientations modeled in the RBO contribute to both dorsal and lateral modulatory signals.

With lateral modulations and in the absence of feedback, hard-coded priors are required, as suggested by Zhaoping [159]. Otherwise, the neurons will face a chance ownership assignment without any priors. Here, we successfully exhibited that with no priors enforced, dorsal feedback gives a prior for lateral communication in local neighborhoods that form the near surround for mBO cells [183]. In short, dorsal feedback both provides global context and eliminates the need for any built-in priors for BO assignment.

We introduced the ROPD dataset as a means to evaluate the performance of our model and the model by Craft *et al*. [18] in complex scenes. Our results show that the RBO network assignments are more accurate than those of Craft's model in all sets with small to large number of polygons, suggesting that the RBO network is more robust to complexities and irregularities that may arise

in a visual scene. One source of complexity in this dataset is the randomness in the intensities of the polygons in the scene, which was well-handled in RBO due to its robustness to low-contrast borders. Bullier [15] suggested that robustness at low contrast in the dorsal stream - because of better low-contrast sensitivity of Magnocellular cells - could be injected to the ventral stream and so, in the BO assignment case, yield a more robust encoding.

As a final remark, evidence from border ownership neurons portray the importance of global context in early visual processing without any need for solving a more complicated problem such as object classification. The RBO network is a realization for how such encodings can be achieved in a computational model without object class information.

Our current model is limited to horizontal and vertical orientations and adding a variety of orientation selectivities is left for future work.

Chapter 4

Combining Analytics and Learning: Learning Shape Representations

4.1 Motivation

The hue network introduced in Chapter 2 and the RBO network described in Chapter 3 were examples of how recent neuroscience discoveries can be incorporated into biologically-inspired computer vision models. These models were inspired by and analytically defined according to some known neural mechanisms in the primate visual system that deal with passive observers and static images. For example, the model V1 neurons in the RBO network were analytically defined with DoG and Gabor filters to model responses of V1 neurons to oriented edges and borders [206, 207]. Similarly, in the hue network, model single-opponent V1 cells were defined according to the findings of hue-selective neurons in V1 [114]. Although this approach is alluring, models defined solely based on the discoveries about visual processing in the primate brain will be limited to the those visual mechanisms.

Despite significant advances in neuroscience, one cannot help noticing that early visual processing aspects have been relatively well-explored, but the same cannot be said about intermediate and higher visual areas. Even with many discoveries about higher processing stages, there is still much to uncover. Consequently, models truthful to the revealed processing aspects of the brain can only go so far in modeling higher visual areas, resulting in missing complex visual features corresponding to unknowns in those areas. Absent complex visual representations in a model will likely diminish its performance.

At first glance, the situation might look bleak: with a lot that remains unknown, it looks impossible to implement a model according to known aspects yet with the capacity to represent complex visual features pertaining to still-unknown mechanisms. To address this problem, we propose a hybrid architecture of analytically-defined and learned parameters. In particular, the proposed architecture pairs analytically-defined neural networks with learning algorithms to model the known aspects and tune the unknown parameters with available data. This approach is in contrast to end-to-end learning where all the model parameters are learned. Our suggestion is to follow a simple rule: there is no need to re-learn the parameters of which we have a reasonable estimate; we only learn model parameters for which we have no or little knowledge.

To demonstrate how a hybrid architecture of analytically-defined neural network combined with learned layers can be employed in a computational vision system, the specific problem of

shape processing in V4 is studied in this chapter. V4, as an intermediate processing stage in the ventral stream (see Figure 4.1) and with extensive connections to many visual areas, is believed to play a role in various visual tasks such as object recognition and visual attention (See [208, 23] for reviews). In the shape representation domain, V4 transforms low-level visual representations of V1 (oriented edges) [50, 194] and V2 (corners and junctions) [209, 210] to more abstract representations in IT (objects, faces, etc.) [97, 98]. As such, for a biologically-inspired model as defined in Chapter 1 and geared toward object recognition, modeling V4 shape representations as those of intermediary nature appears to be essential.

Various studies on V4 found evidence of selectivities to convex and concave shape parts in this visual area [19, 20, 28]. Still, with decades of research on V4 and major discoveries of shape processing in this visual area, the mechanisms by which the observed representations in V4 are achieved is a mystery. In particular, it is yet unclear how V4 neurons map responses of curvature-selective V2 cells into their observed part-based shape selectivities. For a biologically-inspired shape model that includes both V2 and V4, this unrevealed processing hinders modeling. In order to account for these unknown processing aspects, in this chapter, an analytically-defined shape processing network is introduced which explicitly models neurons in V1, V2 and V4. With model V4 neurons as the output layer of this network, the weights in the final layer are learned in a supervised fashion. In order to learn the weights in the last layer, responses of biological V4 cells to a shape stimuli set is provided to the network. With the parameters learned from biological data, the learned weights suggest mechanisms with which V4 neurons integrate responses of V2 cells across their receptive field, improving our understanding of shape processing in this visual area.

Proposing a hybrid architecture to learn how V4 neurons combine curvature-selective V2 responses is one contribution in this chapter. Another important contribution entails suggesting a mechanism by which convex and concave representations of curved segments could be realized in the ventral stream. As will be discussed in detail throughout this chapter, a curved segment could be a convex or a concave part of a shape, depending on its position with the rest of the shape. How this type of representation is achieved in the ventral stream is still unknown. The proposed hybrid architecture in this chapter addresses this problem in an analytical way. In particular, the proposed approach identifies previously discovered V1 and V2 representations that are necessary

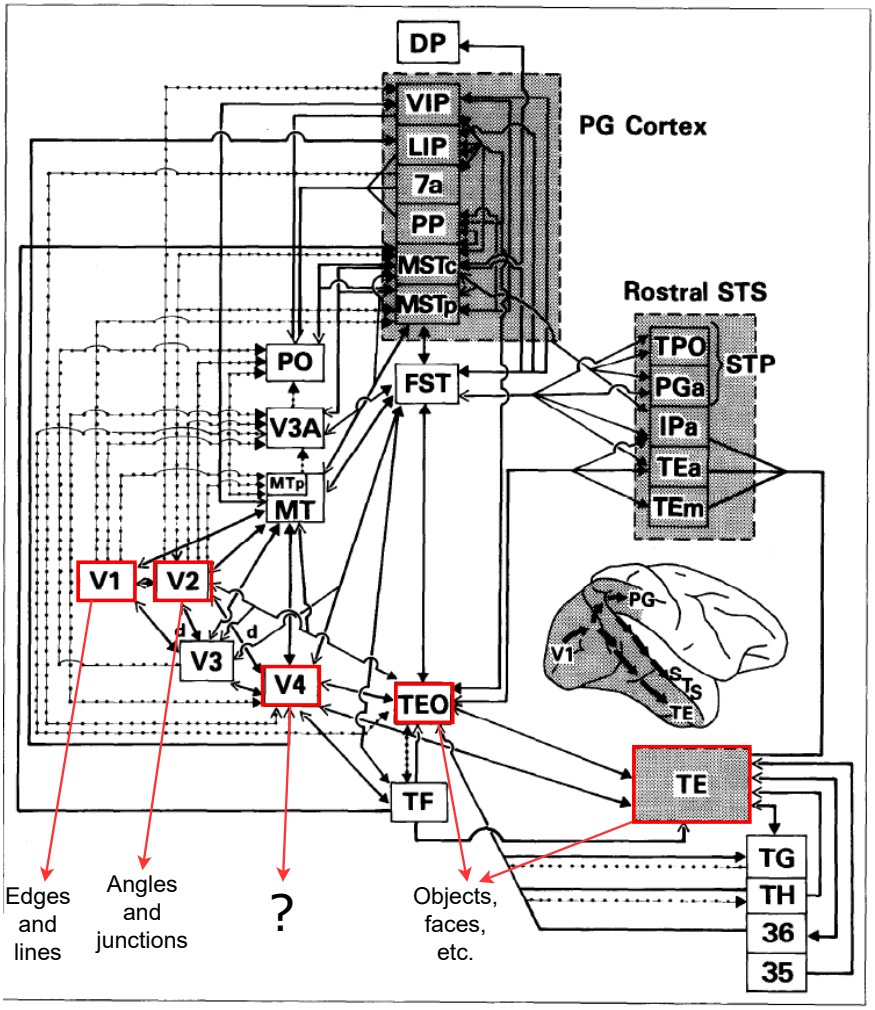


Figure 4.1: The visual hierarchy and their connections. Some of the areas involved in shape processing in the ventral stream are highlighted with red boxes. Visual processing in these areas is not restricted to form analysis. However, the focus in this chapter is on shape processing and therefore, we limit discussion to the processing of shape. In the visual hierarchy, in V1 and V2, orientation and junction selectivities respectively are transformed to more abstract representations of objects and faces in IT (TE and TEO) via the intermediate processing stage of V4. Recovering processing mechanisms in V4 enhance our understanding of this transformation.

components to distinguish convex and concave shape parts.

4.2 Background

The proposed hybrid model in this chapter consists of an analytically-defined architecture and a learning component. Therefore, this section is divided into two parts: one reviewing the shape representation background relevant to the analytically-defined network and one providing a brief

overview of learning a sparse code pertaining to the learning step employed in the proposed model.

4.2.1 Shape Representation

This section starts with a review of neuroscientific findings of shape processing in V4. With the unveiled shape representations in V4, a short section explains how such shape properties are defined in 2D geometry. Then, a review of existing biologically-inspired models mimicking V4 shape responses is followed by a summary on shape representation.

4.2.1.1 The Intermediate Processing Area V4

As an intermediate processing stage and with connections to and from various visual areas (See Pasupathy *et al.* [23] for a review), V4 is believed to play a significant role in visual processing. Shape processing studies on V4 suggest selectivity to complex shape features [211] and on the role of this visual area in shape discrimination tasks [212]. But perhaps the series of research by Pasupathy and Connor [19, 20, 21] were most influential in our understanding of shape processing in V4. Whereas previous studies suggested selectivity to features other than straight lines [211] and were vague in terms of V4 representations, Pasupathy and Conner's studies demonstrated selectivity to convex and concave shape parts, evidence that advocates for an object-centered representation in V4.

The first study in the trio [19] consisted of V4 recordings to contour features and provided evidence that most V4 neurons better responded to curved segments as part of a shape contour than to isolated curved segments. Their stimuli set consisted of convex and concave contour segments as part of a larger shape that passed beyond the RF as well as outline curved segments. For example, the three sharp segments in Figure 4.2a depict sharp convex, sharp outline and sharp concave segments. In all three stimuli, identical edges form the curved segment. But the placement of the curved segment with respect to the rest of the shape in the convex and concave stimuli determines the stimulus type.

The stimulus set from this study illustrated in Figure 4.2b shows part of the presented shapes that coincides with V4 RFs. Example responses of one of the studied V4 cells is shown in Figure 4.2c demonstrating selectivity to convex contour segments pointing toward the left and no or weak responses to corresponding outline and concave features. This observation suggests that simply the

straight edges forming the curved segment were not the only features contributing to the neuron responses as in if it was, the neuron would have responded with equal strength to the corresponding outline and concave stimuli. Pasupathy and Connor reported that most of the studied V4 cells exhibited a similar pattern of responses, *i.e.*, selectivity to a particular shape feature.

Following their observation, they designed a stimuli set of 366 shapes that systematically combined convex and concave parts into closed shapes, as demonstrated in Figure 4.3. They studied responses of 109 V4 neurons to the closed shapes in the set in the following way: first, they decomposed each shape into convex/concave parts and represented each component according to its curvature and angular position with respect to the object center. An example of this representation is illustrated in Figure 4.4a. This way, each shape is represented as a set of points in the two-dimensional space of angular position $\times$ curvature (APC). To recover each neuron's tuning, they fitted a Gaussian to its responses in the APC space. For example, the fitted Gaussian in Figure 4.4b shows that the neuron has a peak to acute convex parts at about 180 deg in angular position. Pasupathy and Connor found many cells in their study were selective to complex boundary configurations, for example, a convexity at a particular position and adjacent to a broad concavity. Examining the population coding in V4 neurons, they could reconstruct a complex shape from the population responses and concluded that a population coding could represent the complete shape in this area [21].

Following this, many others examined responses of V4 neurons to curvature features reporting a progression of curvature preferences from V1 to V4 [213], curvature-polarity selectivity in V4 cells [214], curvature information forming in V4 [171], scale and position invariance [29, 215, 216], and also joint encoding of shape and texture [185], blur [217], and outlined/filled figures [218]. Carlson *et al.* [28] found that a simulated population of neurons with sparse constraints account for the recorded cell responses. Studying V4 responses to occluded shapes, Bushnell *et al.* [219], Kosai *et al.* [220] and Fyall *et al.* [221] found sensitivity to occlusion in these neurons. Anatomically, two recent studies [222, 223] found functional domains in V4 including curvature and corner areas with neighboring regions spanning a diverse set of shape selectivities.

Despite all advances revealing representations in this visual area, how V4 neurons achieve their selectivities is unknown. In particular, while we know V1 neurons are selective to orientation, and V2 cells represent corners with endstopping, achieving an object-centered representation, *i.e.*, selec-

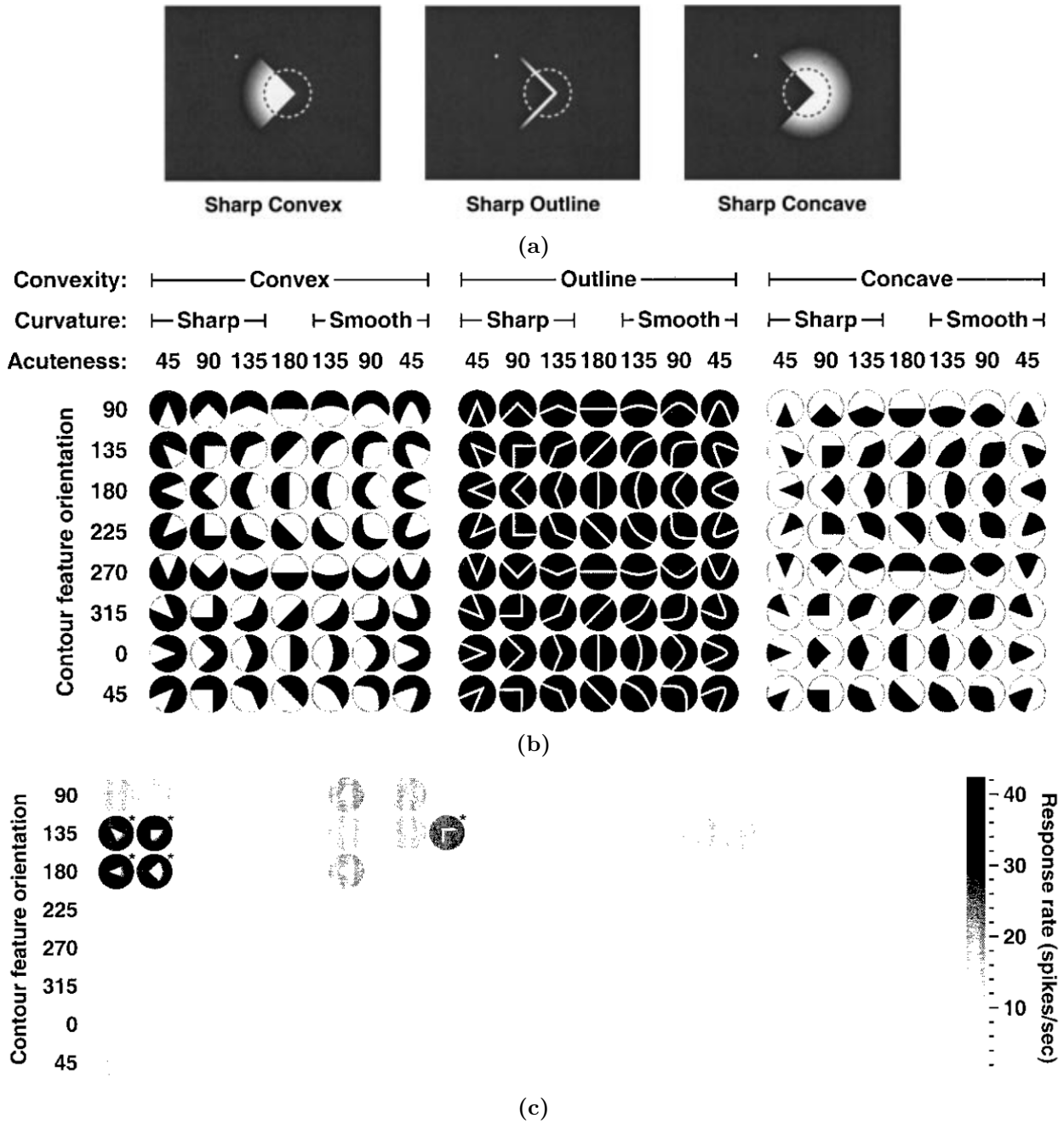


Figure 4.2: Stimuli set from the study by Pasupathy and Connor [19] to investigate responses of V4 neurons to convex/concave shape features. (a) Example stimuli with 90° sharp angle pointing toward the right projected as a convex, outline and concave segments from left to right respectively. Note that in the sharp convex and sharp concave panels, the segment is part of a bigger closed shape defining convexity and concavity. The sharp outline panel shows an isolated contour segment. (b) The complete stimuli set tested in the study by Pasupathy and Connor [19]. Each stimulus is presented as a white icon within a black circle corresponding to the RF. (c) Example neuron responses to the stimuli shown in panel (b). This neuron shows selectivity to sharp convexities pointing toward the left. In their study, Pasupathy and Connor reported that most V4 cells had exhibited similar selectivities to a particular contour feature. All panels are adapted from [19].

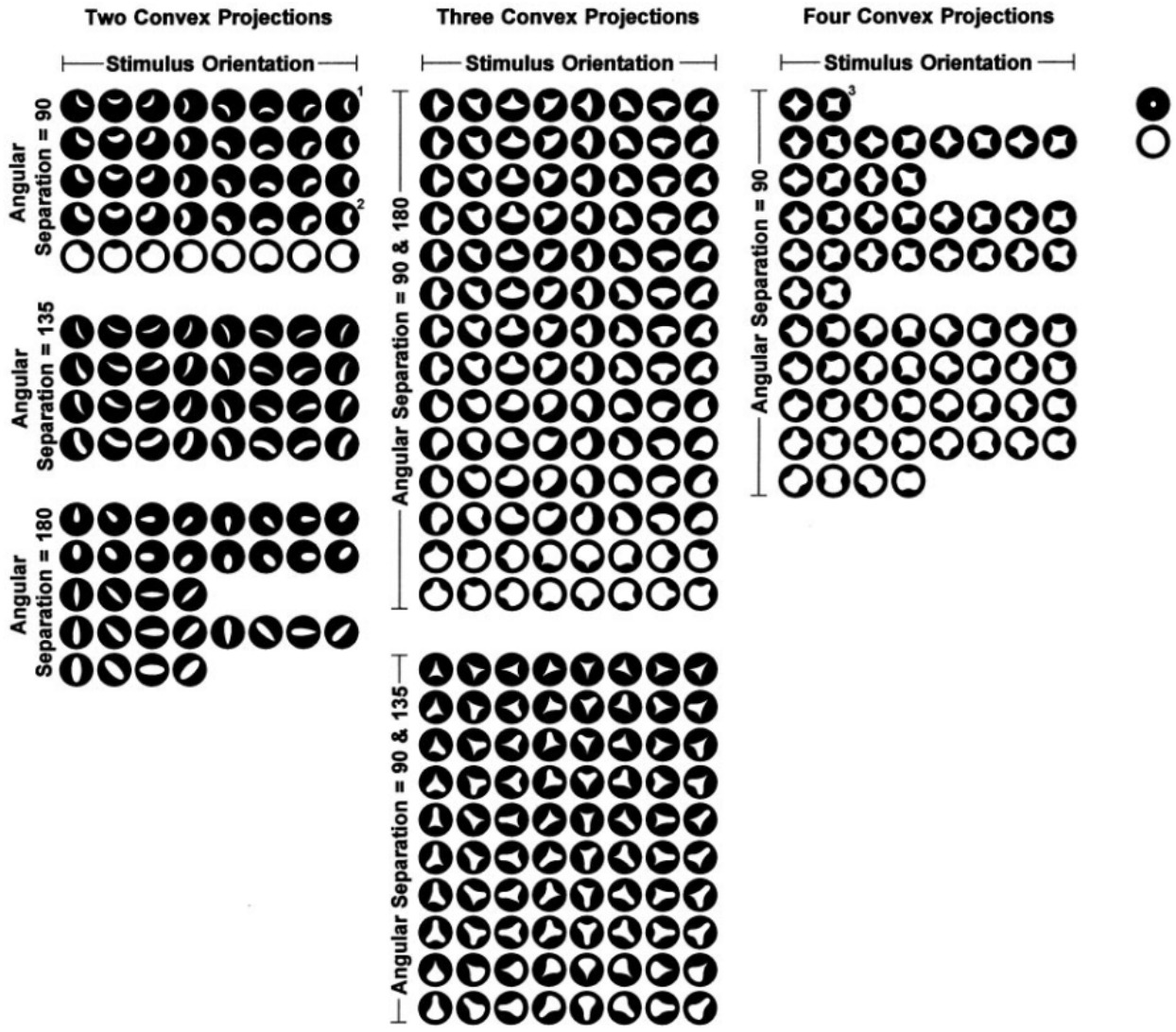
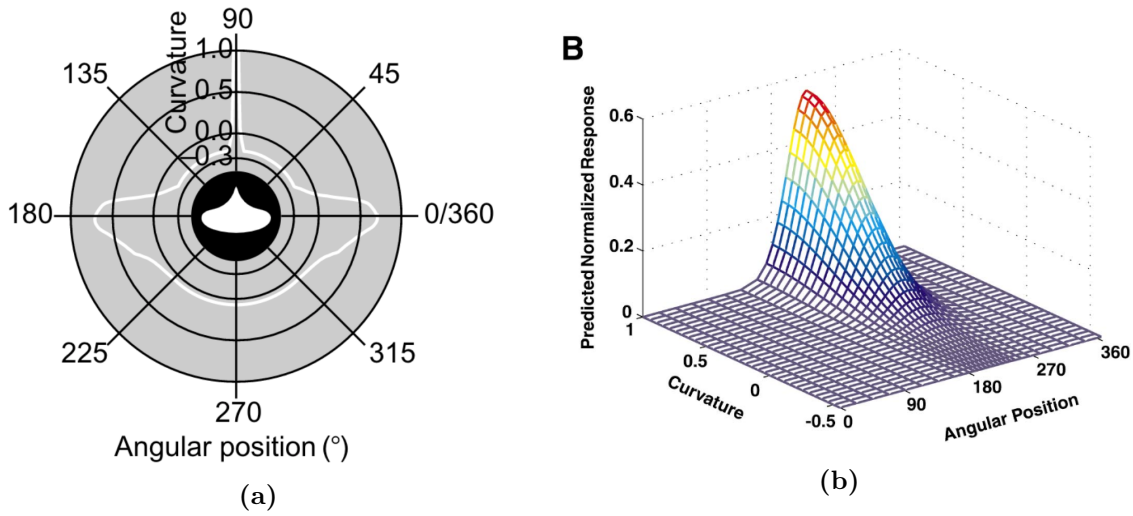


Figure 4.3: The stimuli set employed by Pasupathy and Connor [20] to study responses of V4 neurons to shape parts. Each stimulus is shown as a white icon within the RF (black circles). They systematically combined convex/concave components into simple closed shapes (Adapted from [20]).

tivity to convexities and concavities, is a leap yet to be discovered. In the next section, definitions of convexity and concavity in 2D geometry are reviewed in order to pinpoint the components that yield such shape representations.

4.2.1.2 Object-centered Encoding: A Computational Perspective

In two-dimensional geometry, curvature refers to the amount that a geometrical shape deviates from a straight line at each point, often defined as the reciprocal of the radius of the osculating



(a) (b)

Figure 4.4: In their studies [20, 21], Pasupathy and Connor represented each shape in the Angular Position and Curvature (APC) space. (a) shows the angular position × curvature space where the radial direction shows curvatures from -0.3 to +1.0. In this polar plot, a shape is shown at the center of the plot with its APC representation denoted by a white curve (adapted from [21]). (b) A two-dimensional Gaussian fitted to responses of a V4 neuron in the APC space encoding its tuning function. This neuron's tuning has a peak for acute convexities at around 180deg (adapted from [20]).

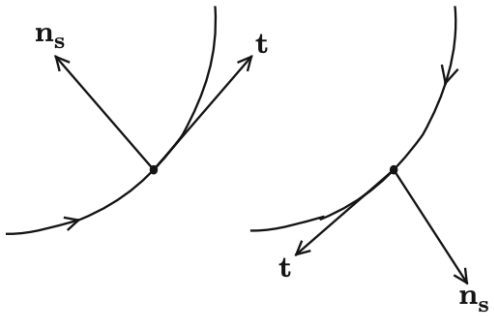


Figure 4.5: Identical contour segments have opposite sign of curvature as a function of curve parametrization (Adapted from Pressley [22]).

circle at each point on the planar curve, measuring the magnitude of *signed* curvature. To be more exact, consider the two planar curves in Figure 4.5. The arrowhead on each curve represents the direction of increase in the curve parameter s. The two vectors $\mathbf{t}$ are tangents to the curves at a single point, and the unit normals $\mathbf{n_s}$ are obtained by rotating the tangents in the counterclockwise

direction by $\frac{\pi}{2}$. Then, the signed curvature, κ_s is defined as the scalar such that

$$\frac{d\mathbf{t}}{ds} = \kappa_s \mathbf{n_s}. \tag{4.1}$$

By this definition, curvature for 2D curves has two components: magnitude and sign. Generally, the mention of curvature refers to the magnitude of signed curvature. We distinguish between these two components and refer to those as curvature magnitude and curvature sign.

An important implication from the definition above is that curvature sign for a 2D curve is determined by its parametrization, s, as demonstrated in Figure 4.5. For closed simple 2D shapes, curvature sign is defined according to the convention of counterclockwise parametrization, which essentially identifies interior versus exterior for the closed contour. In other words, what we need for the computation of signed curvature for a closed contour is to first identify inside vs. outside for the shape, or in vision terminology, we need to identify figure vs. ground along the contour.

Given the signed curvature definition for simple 2D shapes, a consequential question is: are V4 neurons selective to curvature magnitude, curvature sign or both? In a recent publication, Pasupathy *et al.* [23] suggested that V4 responses show selectivity to both curvature magnitude and sign. They argued that if V4 responses were "based on a simple preference for feature conjunctions", or equivalently curvature magnitude, then, V4 cells would have shown similar preference to the two shapes shown in Figure 4.6 (panel b). To reject this possibility, Pasupathy *et al.* provided evidence (Figure 4.6 (panel c)) from V4 cells exhibiting an object-centered tuning to contour conformations[1], implying that a notion of figure-ground is passed to these cells.

Interestingly, in differential geometry, inside-outside information for simple closed curves is studied as a *global* property of curves. Similarly, in the visual system, as was shown in Chapter 3, side-of-figure preference depends on *global* context. Together, available evidence from V4 neurons suggests that both side-of-figure information and curvature magnitude features form the building blocks that have an impact on V4 responses. As such, a model of shape selectivity mimicking V4 responses requires to develop both feature types in its formulation. Next, formulation of both curvature components in existing computational models that address V4 shape processing mechanisms are reviewed.

[1] arrangement of parts into a whole (adapted and modified from the Merriam-Webster Dictionary [224]).

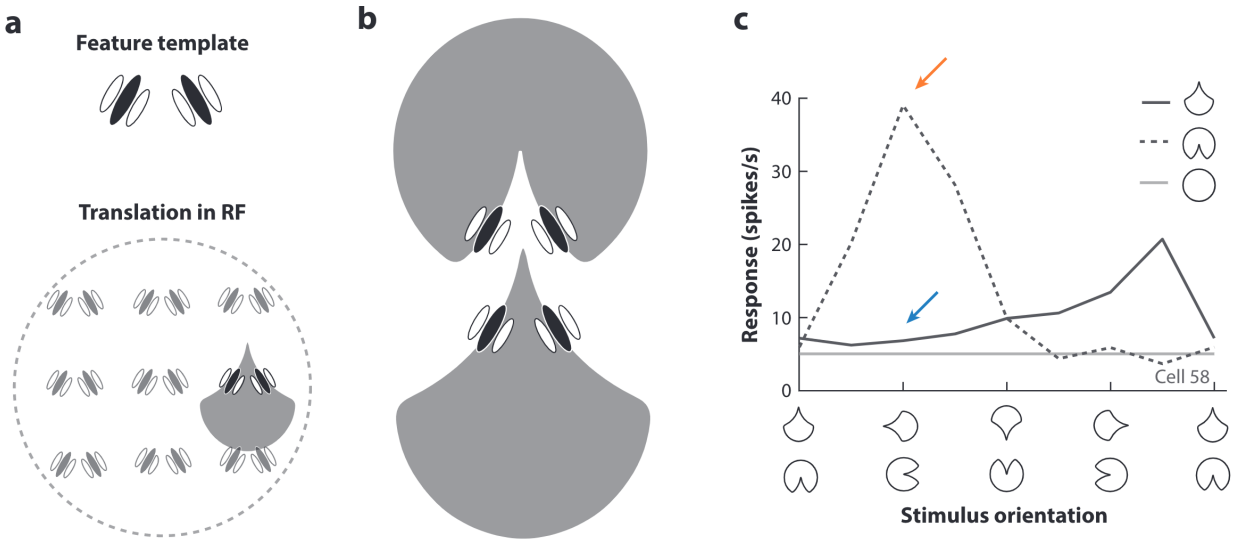


Figure 4.6: Object-centered representation in V4 neuron. In panel (a), a combination of orientations could encode contour conformations with sharp curvature. When pooled over multiple locations within the RF, translation-invariance is achieved. Panel (b) shows how the same combination of orientations could encode both a sharp convexity and concavity. A feature-conjunction model predicts similar responses for both shape conformations. However, as shown in panel (c), V4 neurons exhibit a difference of responses to these curvature components with opposite signs, suggesting an object-center encoding in V4. Figure adapted from [23].

4.2.1.3 Computational Models for 2D Shape Encoding

The majority of existing models are hierarchical, with the exception of the model proposed by David *et al*. [225] whose suggestion was rejected in a later neurophysiological study by Oleskiw *et al*. [226].

The hierarchical models suggesting plausible mechanism for shape processing in V4 consist of a biologically-inspired architecture of neurons with simulated V4 cells at the top [227, 228, 229, 230, 231, 232, 233, 24, 16]. These models are feedforward networks that combine earlier representations to make more complex features in their higher layers in a purely bottom-up fashion and without recurrence or feedback. Whereas providing global context requires integrating representations from far visual locations, a simple feedforward model lacks the capacity to provide context to neurons in its earlier layers. Therefore, it is impossible for such architectures to inject global information to neurons in layers prior to their model V4 cells and thus to encode signed curvature similar to V4 neurons.

A few CNN architectures were probed in some recent studies and the selectivities of artificial

neurons in CNNs were compared to V4 responses [56, 234]. It was argued at length in earlier chapters that even though these findings could potentially be significant, V4-like representations in CNNs do not lead to insight about biological shape processing mechanisms. Regardless, the CNNs that were investigated in these studies were feedforward networks to which the same argument about global context and signed curvature encoding applies.

Among the proposed models for shape selectivity in V4, to the best of our knowledge, HMAX [24] and 2DSIL [16] are the only two biologically-inspired hierarchical models that directly compare their shape representations to biological V4 data. Both models are, similar to the rest of hierarchical models, feedforward networks and therefore, lack mechanisms required for a signed curvature encoding. However, HMAX [24] and 2DSIL [16] are two models closest to the one proposed in this chapter. As a results, the architecture and computations of these models are reviewed separately in detail below.

4.2.1.3.1 HMAX

The model for V4 shape processing, proposed by Cadieu *et al.* [26, 24], consisted of a subset of the HMAX network layers [25, 235, 236]. The HMAX hierarchy coupled layers of S and C cells performing feature selection (S layers) followed by a max operation (C layers). Part of the HMAX architecture that was employed in modeling V4 by Cadieu *et al.* is depicted in Figure 4.7. Although C1 neurons were claimed to correspond to V2 cells, due to the max operation, these cells only ensure translation invariance in orientation selectivity, similar to complex cells in V1. As such, it is not unreasonable to suggest that the HMAX architecture skips representing angles and junctions reported in V2 neurons [209, 210].

Cadieu *et al.* [26, 24] employed S2 and C2 layers to model shape selectivity and position invariance in V4 neurons respectively. All the weights in this network are fixed according to experimental data except for the C1-S2 weights which are learned given the data from the V4 study by Pasupathy and Connor [20]. Specifically, they introduced a heuristic greedy algorithm to select between 2 and 25 C1 units, along with their weights, feeding to S2. Figure 4.8 illustrates the learned S2 configurations corresponding to 109 V4 neurons from the study by Pasupathy and Connor [20]. As the figure shows, with no explicit curvature encoding, the learning step is pushed to compensate for curvature representations but does not necessarily result in curvature-like configurations. Finally,

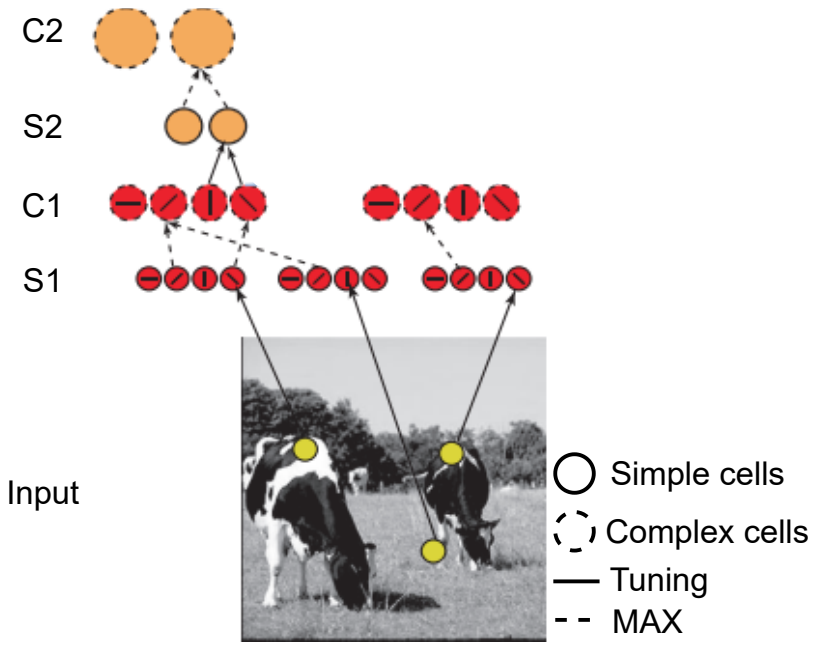


Figure 4.7: The HMAX model. HMAX consists of a hierarchy of feature selection layers followed by a max operation. To model V4 responses, Cadieu *et al*. [24] utilized the first four layers of the original HMAX network, corresponding to V1 and V4 areas in the brain. Figure adapted from [25].

HMAX does not encode curvature sign.

As an example, consider a learned RF with a clear curvature-like representation from HMAX shown in Figure 4.9, panel a (the RF identified by a blue rectangle in Figure 4.8). The three C1 components with strongest contributions to responses of this neuron are isolated in the right square in Figure 4.9, panel a. Panels b, c, and d in this figure can be compared to similar ones in Figure 4.2a, where matching convex, outline and concave curved segments were shown. Similar to the stimuli in Figure 4.2a, the mild convex and mild concave cases are part of a larger shape and hence a notion of inside-outside determines the curvature sign in these cases. This HMAX neuron, however, will exhibit strong responses to all of the convex, concave, and outline curved segments. In other words, HMAX, unlike V4 cells [19], does not encode curvature sign.

4.2.1.3.2 2DSIL

The model proposed in this chapter is an extension of the 2DSIL network, whose architecture is depicted in Figure 4.10. Rodriguez-Sanchez *et al*. [16] hypothesized that intermediate processing stages play a key role in shape selectivity in higher layers; a missing feature in most computational

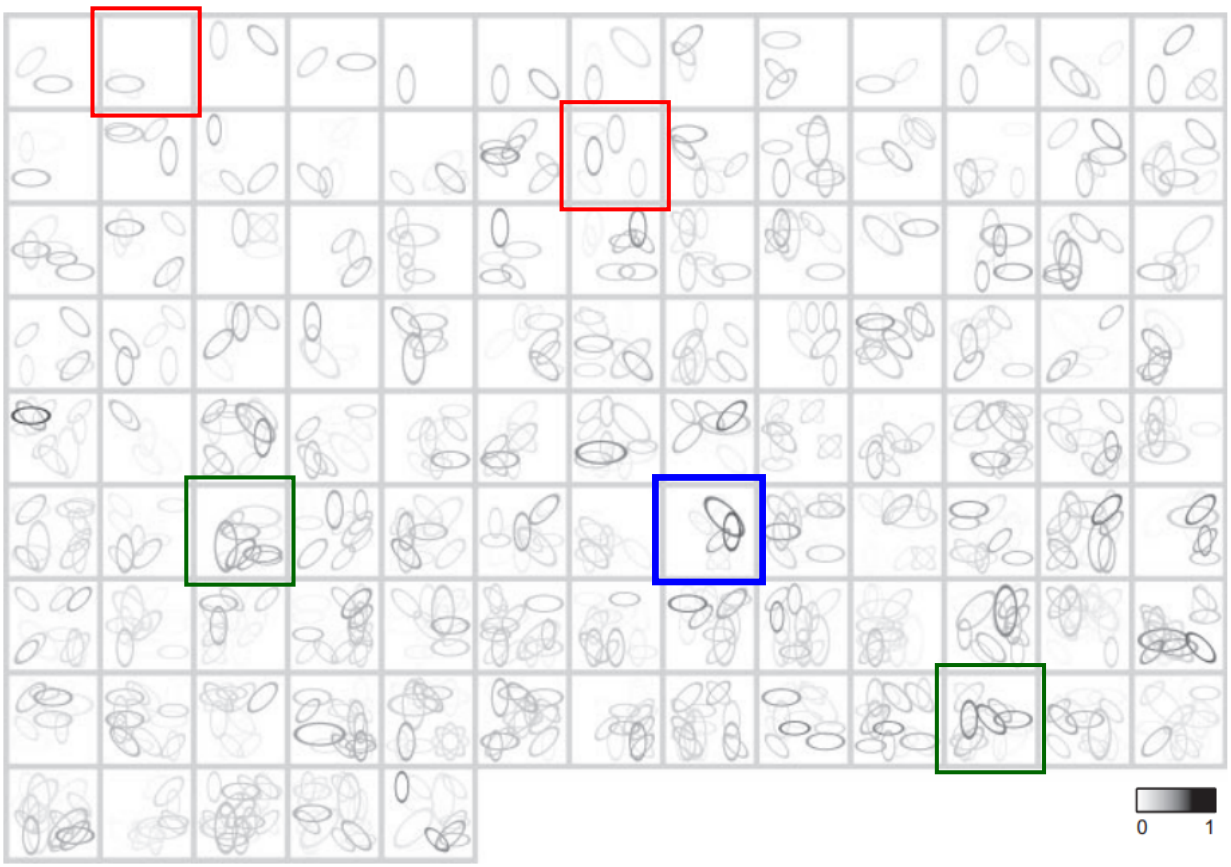


Figure 4.8: Learned S2 configurations in Cadieu *et al*. [26]. Cadieu *et al*. [26] employed a four-layer HMAX model combined with a greedy learning algorithm that selected between 2 and 25 C2 units to reproduce responses of 109 Macaque V4 cells. Each square in this figure represents the set of C2 units selected by their greedy algorithm to reproduce the responses of a single cell. Each ellipse in a single C2 unit corresponds to one contributing C1 unit, with the grayscale level of its boundary representing its weight (darker means a larger weight). The marked red squares show two examples of cells for which no curvature-like representation was learned, and the green squares are two examples with the learned patterns that are difficult to interpret. The neuron RF identified by a blue rectangle is employed in Figure 4.9 to demonstrate how the learned RFs by HMAX cannot represent shape parts. Figure adapted from [24].

models including the HMAX network. Specifically, whereas the human and primate brain performs processing through several layers of neuronal computations to represent complex features, all those intermediate layers between V1 and V4 are skipped in HMAX. To test the importance of intermediate representations in shape processing, Rodriguez-Sanchez *et al*. explicitly modeled endstopped cells encoding a class of curvature features in V2. Responses of these neurons are combined in the local curvature layer of 2DSIL, which sends a feedforward signal to shape-selective V4 neurons. To model individual biological V4 cells from the study by Pasupathy and Connor [20], they determined

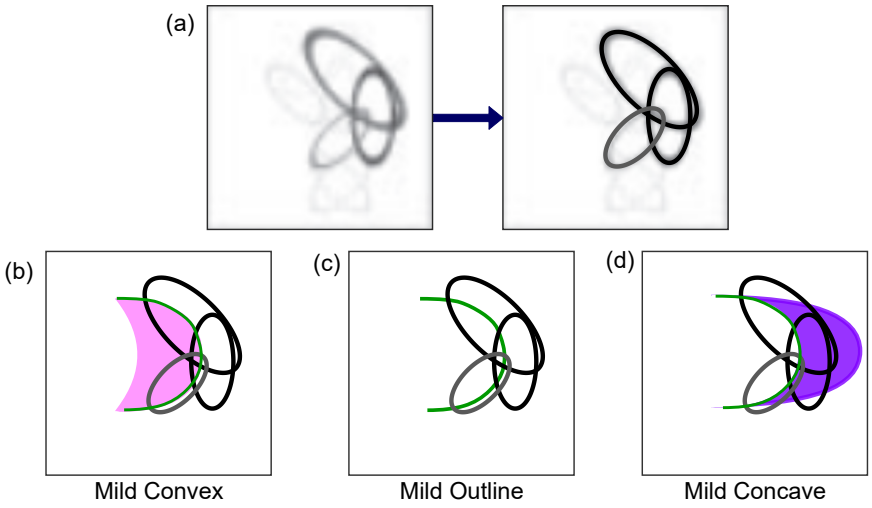


Figure 4.9: A learned RF by HMAX with clear curvature-like representation. (a) The left square shows the RF directly extracted from [24] and the right square shows the RF profile with the three C1 components with strongest contributions to responses of the neuron. Panels (b), (c), and (d) show the RF profile super-imposed with mild convex, mild outline and mild concave stimuli. This neuron will exhibit strong activations to all three stimuli as the highlighted green curved segment, a common feature between all three stimuli, overlaps with the three C1 components. This behavior is in contrast to the reported observations in V4 neurons.

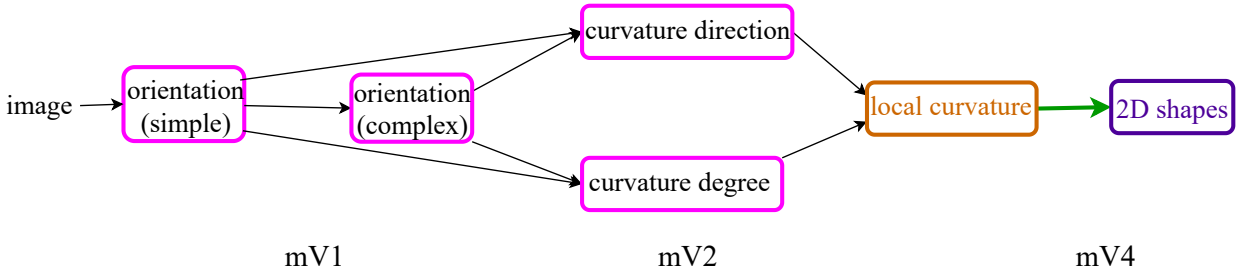
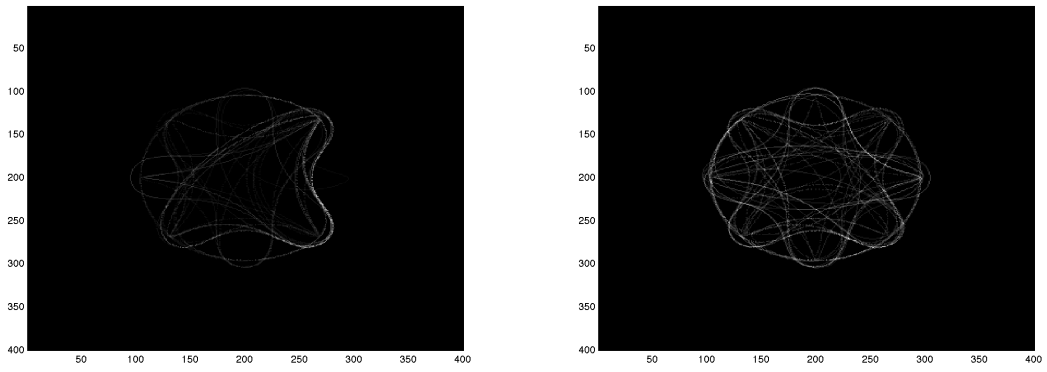


Figure 4.10: The architecture of the 2DSIL network. In 2DSIL, simple and complex cells are selective to oriented edges. Endstopped neuron in this network combine responses of simple and complex cells to achieve a set of curvature representations. Local curvature cells combine responses of endstopped neurons and send a feedforward signal to shape selective cells in V4 (adapted from [16].

the weights between local curvature cells and V4 neurons from the shape parts in common among the neuron's preferred stimuli. With this approach, having a single selectivity template capturing all contributing features to neuron responses is not guaranteed (as shown in the example in Figure 4.11b), limiting the ability to predict responses of those cells with no selectivity template.

Even though 2DSIL explicitly models curvature-selectivity, its implementation does not include a curvature sign encoding. In particular, in 2DSIL, two types of endstopping neuron represent curvature features: degree and direction. These two types of model neurons will be called mEsDeg



(a) **(b)**

Figure 4.11: Example of maps that helped determining neuron shape selectivity in 2DSIL. (a) The neuron showed selectivity to stimuli with a convexity in the lower right corner of the receptive field followed by a concavity in the counter clockwise direction. (b) An example that shows the selectivity of the neuron is not clear from its selectivity map. Figure adapted from [16].

and mEsDir for short. Both mEsDeg and mEsDir cells in 2DSIL combine a simple cell at the RF center and two displaced complex cells with facilitatory and inhibitory contributions respectively. The structure of endstopped-curvature degree neurons in 2DSIL shown in Figure 4.12a demonstrates that mEsDeg neurons are designed to signal deviations from straight lines, encoding curvature magnitude. Figure 4.12b illustrates the structure of mEsDir neurons. These cells in 2DSIL signal the direction towards which the contour curves with respect to the orientation selectivity of their facilitatory simple cell, and not curvature sign. The example in Figure 4.12c demonstrates two mEsDir neurons with identical selectivities that will be strongly activated with both convex and concave contour segments. To summarize, 2DSIL encodes curvature magnitude and curvature direction, but not curvature sign and thus, does not encode signed curvature.

4.2.1.4 Summary

Neurophysiological findings suggest that V4 responses encodes signed curvature. However, none of the existing computational models could explain how such a representation could be achieved in the ventral stream. Considering V4 models, the tuning fitting approach employed by Pasupathy and Connor [20], the APC model, can be seen as a shape model which maps curvature-position features to V4 responses. Although the APC model does not lend insight on how a signed curvature representation is achieved in the ventral stream, it represents shape parts in an object-centered

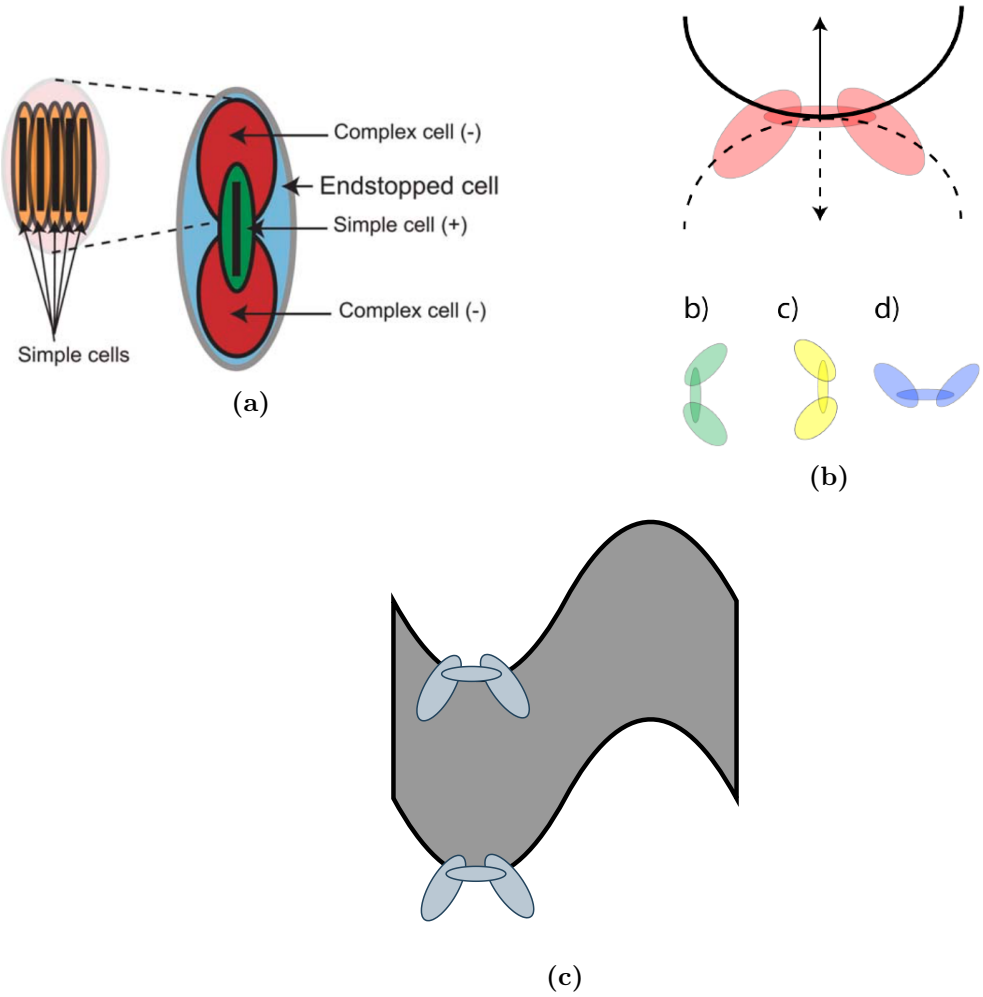


Figure 4.12: Model Endstopped neurons in 2DSIL. (a) Model Endstopped - curvature degree (mEsDeg) cells from 2DSIL combine responses of a simple cell (green ellipse in the B panel) centered within their receptive field with responses of two displaced complex neurons (brownish ellipses in panel B). (b) Examples of endstopping neurons encoding curvature direction introduced in [27] at horizontal and vertical orientations. Responses of endstopping neurons are determined from interactions of simple and complex cells shown as ellipses. (c) Model endstopped neurons in 2DSIL representing curvature direction (mEsDir) are agnostic to curvature sign. In this figure, the two mEsDir cells will have similar activations to the vally-like parts of the figure even though these segments correspond to parts with opposite curvature signs.

manner and suggests how V4 cells integrate shape components within their RF. As such, the APC model, along with HMAX and 2DSIL, will be considered as relevant work in the remainder of this chapter. The APC model, however, imposes a strong prior, *i.e.*, a single Gaussian to V4 responses, which impedes capturing complex and long-range interactions between shape parts within the receptive field. Additionally, all the three models limit shape part contributions to facilitatory ones. Whereas neurons in other brain areas exhibit a combination of excitatory and inhibitory

regions with their RFs, these models do not account for inhibitory shape parts, which leaves one with an incomplete understanding of V4 selectivities.

To conclude the shape representation background, two important questions with regards to V4 responses are still unanswered: first, how a signed curvature encoding is achieved in the ventral stream, and second, how V4 neurons integrate shape parts within their RF.

4.2.2 Learning a Sparse Code

Later in this chapter, we propose combining an analytically-defined neural network for V4 with a sparse coding learning step. Consequently, a brief review of sparse coding is sketched below.

In general, the term "sparse coding" can be described as explaining an observation using a *small* set of primary components.

4.2.2.1 Sparse coding in machine learning

In machine learning, sparse coding often falls under the category of unsupervised learning algorithms and its goal is to learn a set of primary features from a set of observations. These features form an over-complete set of atoms of a dictionary and are extracted by minimizing a reconstruction error term. Having learned a set of features, then, any given input can be represented by a sparse linear combinations of the dictionary atoms. Mathematically, given a training set of observed inputs $X_t, t \in \{1, \ldots, T\}$, an over-complete dictionary $D = \{d_1, \ldots, d_m\}$ is learned by solving the following optimization problem:

$$\min_{D} \sum_{t=1}^{\tau} \min_{\gamma_t} \|X_t - D\gamma_t\|_2 + \lambda \|\gamma_t\|_1, \tag{4.2}$$

where γ_t is the code or representation for the observation X_t. With X_t and D as $n \times 1$ and $n \times m$ matrices, the code γ_t will be a $m \times 1$ vector. The first term in this equation is called the reconstruction error and the second term ensures sparsity of the code vector. This optimization problem is tackled by alternatively solving for γ_t and D. Various algorithms are introduced for each of the two subproblems (See [237, 238, 239, 240] just as a few examples). Adding discriminatory terms to Equation 4.2, sparse coding can be employed in discriminative tasks [241].

Note that in Equation 4.2, one code or representation, γ_t, is found for each training data, while the dictionary is learned with respect to all the training examples. In dictionary learning, often,

the only imposed constraint is normal length for the dictionary atoms. That is, $\|d_i\|_2 = 1$ for all atoms of the dictionary. However, in a number of approaches dubbed as *parametric dictionary design* [242, 243], the structure of the dictionary atoms is assumed to be known a priori and only the parameters of the dictionary atoms are learned in the outer optimization problem. For example, assuming Gabor-like dictionary atoms, for each atom, Gabor parameters are learned from the training set. This family of approaches has the advantage of having far fewer learnable parameters compared to the general dictionary learning algorithms.

4.2.2.2 Sparse coding in Neuroscience

Sparse coding in the neuroscience community alone has different definitions and is a controversial topic. Moreover, in a recent survey, Beyeler *et al.* [244] lists studies that have found evidence of sparse coding in the brain, while citing studies that suggest against it. But perhaps the most well-known work in this area was by Olshausen and Field [239] in which they showed that by imposing sparsity constraints, a set of features that resemble V1 receptive fields would emerge from learning the statistics of natural images. These learned statistics were obtained with an unsupervised approach. Later approaches modified the work of Olshausen and Field to make their approach more biologically compatible [245, 246, 247].

Even though in the present work, our proposed approach incorporates sparse coding, we emphasize that the sparse coding step is employed in the machine learning context and our motivation for using a sparse coding formulation was purely computational. Therefore, we refrain from drawing any connections between sparse representations in V4 [28] and the learned ones in the sparse coding step of the proposed model.

4.3 The SparseShape Network

With the aforementioned state of affairs, in this work, we seek to answer the following two questions:

1. How is an object-centered representation achieved in the ventral stream?
2. Recovering neuron tunings with a single Gaussian is limiting in that contribution from only a single shape part can be recovered. Here we ask: what will the tunings reveal if we relax the fitting prior?

To answer the first question, we introduce the SparseShape network as an analytical neural network that suggests a mechanism for object-centered representations in the ventral stream. The SparseShape network extends the 2DSIL model to reach this goal.

Our proposal to the second question is to pair the suggested analytical neural network with a learning step to learn weights in the last layer of the network, accounting for unknown shape processing mechanisms in V4. More specifically, we seek to find how shape-selective neurons in V4 integrate responses of curvature-selective cells in V2 [209, 210] to achieve their reported responses. For this purpose, we relax the single-Gaussian prior and replace it with a sparsity one that would allow capturing shape part contributions across the receptive field in both facilitatory and inhibitory manners. This latter step demonstrates how combining biologically-inspired models with learning algorithms can improve the performance of these networks and recovering certain shape processing aspects of which little or no knowledge is in hand.

4.3.1 The Analytically-defined Architecture

The SparseShape network extends 2DSIL by combining it with the RBO network introduced in Chapter 3. As explained earlier, to model neurons selective to curvature sign, figure-ground information is required. As a result, the combined model will have the necessary features that can result in a signed curvature encoding. The architecture of SparseShape is depicted in Figure 4.13 with each of the various model neuron types shown with a single box. Similar to 2DSIL and RBO, simple and complex cells in SparseShape are selective to oriented edges. mBO neurons signal side-of-figure information while curvature magnitude and direction are encoded at the endstopped level in mEsDeg and mEsDir neurons. In SparseShape, new cells endcoding curvature sign are modeled by combining mBO and mEsDir signals. With the added curvature sign representation, the local curvature neurons from the original 2DSIL are remodeled by receiving their input from endstopped-curvature degree and endstopped-curvature sign representations. In other words, in SparseShape, the local curvature neurons combine curvature magnitude and sign and therefore, encode *signed curvature* (κ_s in Equation 4.1). Finally, the green arrow in Figure 4.13 represents the set of weights that are learned by employing a sparse coding algorithm. This final stage of the network is explained after the computational details of neurons in SparseShape are explained below.

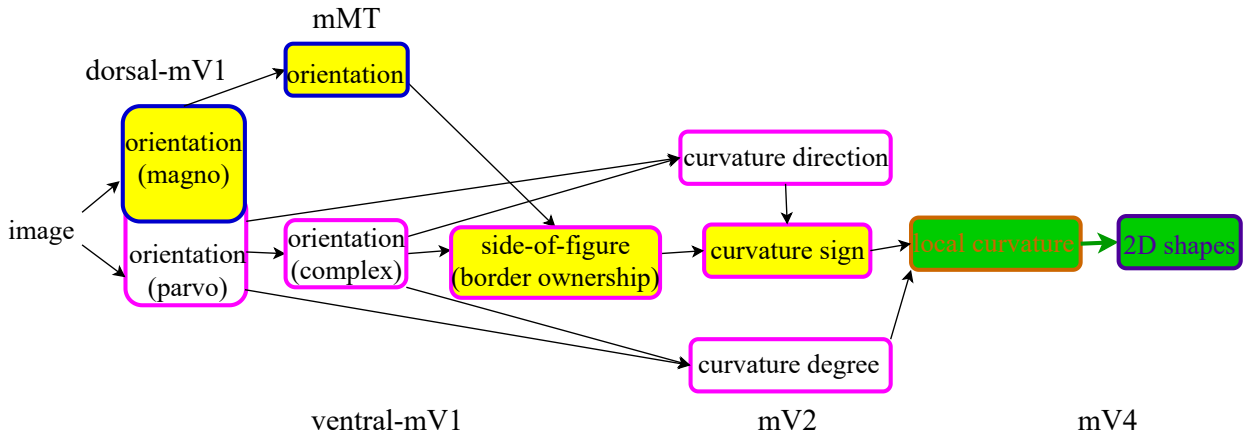


Figure 4.13: The SparseShape architecture. This hierarchy is designed by combining the RBO network introduced in Chapter 3 with that of the 2DSIL network. The blocks with yellow and green fill represent model neurons that are added to 2DSIL or re-modeled respectively. To obtain an object-centered encoding, additional mechanisms to represent inside vs. outside for shapes is required. These mechanisms are implemented through the RBO network with border ownership encoding. In SparseShape, curvature sign model cells are added to the network to achieve an object-centered shape tuning in mV4. Accordingly, the model local curvature cells are re-modeled with feedforward signals from curvature degree and sign cells. Similarly, 2D shape cells at the final layer of the network are re-modeled with the sparse coding step.

Similar to the RBO, neurons in SparseShape are implemented at four scales that result in eight combined curvature maps in the local curvature layer (four scales representing curvature magnitude $\times$ two signs). Unlike RBO that implemented horizontal and vertical orientations, SparseShape extends the RBO network to model 12 orientations in the $[0, \pi)$ range. In 2DSIL, only edge-selective neurons in model ventral simple cells are implemented while RBO includes both edge- and border-selectivity. To reconcile the selectivities in the two networks, responses of mBO neurons with edge and border local selectivity are combined. Moreover, the model ventral simple and complex cells sending their signal to endstopped-curvature degree and direction are those selective to edges. In SparseShape, neurons up to and including border ownership cells are the same as those in the RBO network. The computational details for these neurons can be found Chapter 3.

Names of some of the cells in SparseShape tend to get long and to avoid repeating long names, abbreviations are used in the remaining of this chapter. The list of these abbreviations can be found Table 4.1.

Neuron Name	Function	Abbreviations
Model ventral simple cell	orientation	mvSimple
Model ventral complex cell	orientation	mvComplex
Model border ownership	side-of-figure	mBO
Model endstopped curvature degree	curvature magnitude	mEsDeg
Model endstopped curvature direction	curvature direction	mEsDir
Model endstopped curvature sign	curvature sign	mEsSign
Model local curvature	signed curvature	mLocalCurv
Model V4	2D shape	mV4

Table 4.1: List of SparseShape neurons with their name abbreviations.

4.3.1.1 Model Endstopped - Curvature Degree

Endstopped-curvature degree neurons (mEsDeg) in 2DSIL, whose RF is depicted in Figure 4.12a, combine activations of a simple cell at their RF center with two displaced complex neuron as follows:

$$R_{\text{mEsDeg}}(x, y, \theta) = \Phi[c_c\phi(R_c(x, y, \theta)) - (c_{d1}\phi(R_{d1}(x, y, \theta)) + c_{d2}\phi(R_{d2}(x, y, \theta)))], \tag{4.3}$$

where c_c, c_{d1} and c_{d2} are the gains for the center and displaced neurons. $R_c(x, y, \theta)$, $R_{d1}(x, y, \theta)$, and $R_{d2}(x, y, \theta)$ are responses of the center and displaced cells with their RFs centered at (x, y) and selective to orientation θ. The function ϕ rectifies negative responses to zero. In this equation, Φ is the same as the one defined in Equation 3.6, which scales responses to highly intense stimuli. The gains and the parameters of the rectifier Φ from this equation in SparseShape match those employed in 2DSIL. It is worth noting that the four scales of mEsDeg encode selectivities to a variety of curvature magnitudes from acute to broad curvatures.

4.3.1.2 Model Endstopped - Curvature Direction

In 2DSIL, for each orientation in the model, two model endstopped-curvature direction cells (mEsDir) signaling opposite directions are implemented. These two types of neurons are implemented

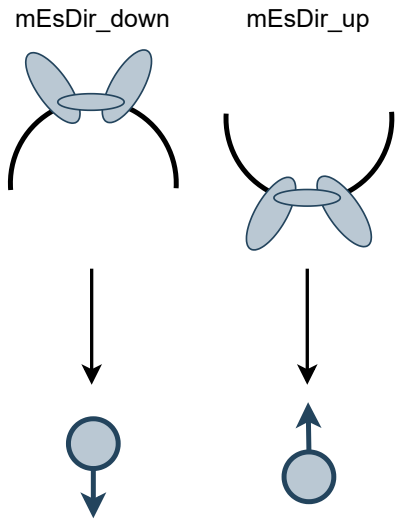


Figure 4.14: The convention to show model endstopped neurons encoding curvature direction (mEsDir). The mEsDir neuron with stronger responses when the contour segment curves downward denoted as mEsDir_down will be show by a circle with an attached arrow pointing downward. The mEsDir cell with opposite selectivity, shown as mEsDir_up, will be illustrated as a circle with an upward arrow attached to it. These neurons will be shown in *blue*.

as:

$$R_{\text{mEsDir}}(x,y,\theta-\frac{\pi}{2}) = \Phi[c_c\phi(R(x,y,\theta)) - \tag{4.4}$$

$$(c_{d1}\phi(R(x_{d1},y_{d1},\theta+\frac{\pi}{4})) + c_{d2}\phi(R(x_{d2},y_{d2},\theta+\frac{3\pi}{4})))], \tag{4.5}$$

$$R_{\text{mEsDir}}(x,y,\theta+\frac{\pi}{2}) = \Phi[c_c\phi(R(x,y,\theta)) - \tag{4.6}$$

$$(c_{d1}\phi(R(x_{d1},y_{d1},\theta+\frac{3\pi}{4})) + c_{d2}\phi(R(x_{d2},y_{d2},\theta+\frac{\pi}{4})))], \tag{4.7}$$

again with c_c, c_{d1} and c_{d2} as the gains for the center and displaced neurons, $d1, d2$ indicating displacements, and R representing responses of endstopped, center and displaced cells. As depicted in Figure 4.12b, the displaced neurons are at 45deg and 135deg angle with respect to the center cell. The two parameters $\theta+\frac{\pi}{2}$ and $\theta-\frac{\pi}{2}$ denote the direction of the curve relative to the central mvSimple cell orientation at θ angle (the direction of the arrows in Figure 4.14). Similar to Equation 4.3. the parameters in this equation match those of 2DSIL.

4.3.1.3 Model Endstopped - Curvature Sign

Intuitively, a contour segment of a simple closed curve is convex when the contour segment curves toward inside the figure. Geometrically, for a counterclockwise parametrization of the curve, convex segments are parts of the shape at which the direction of the tangent vector derivative and the normal vector are the same. Otherwise, the curve segment is concave. In SparseShape, this intuition and geometry can be implemented by combining responses of mEsDir and mBO cells, where mEsDir represent the tangent vector derivative direction and mBO neurons encode inside vs. outside (or the normal vector $\mathbf{n_s}$ direction) at each point. More specifically, a contour segment is convex when the direction of the winning mEsDir neuron and the winning mBO cell[2] are in agreement and concave otherwise. With winning mBO and mEsDir neurons at a given orientation, four scenarios might occur. These four scenarios are illustrated in Figure 4.15 for horizontal orientation. Scenarios 1 and 2, where the direction of winning mBO and mEsDir are the same, represent convexities, and scenarios 3 and 4 with opposite directions for the winning mEsDir and mBO neurons correspond to concavities.

In SparseShape, model endstopped-curvature sign (mEsSign) cells are implemented as:

$$\begin{aligned} R_{\text{mEsSign\_convex}}(x,y,\theta) &= \phi((R_{\text{mEsDir}}(x,y,\theta+\frac{\pi}{2}) - R_{\text{mEsDir}}(x,y,\theta-\frac{\pi}{2})) \times \\ &\quad (R_{\text{mBO}}(x,y,\theta+\frac{\pi}{2}) - R_{\text{mBO}}(x,y,\theta-\frac{\pi}{2}))), \\ R_{\text{mEsSign\_concave}}(x,y,\theta) &= \phi(-(R_{\text{mEsDir}}(x,y,\theta+\frac{\pi}{2}) - R_{\text{mEsDir}}(x,y,\theta-\frac{\pi}{2})) \times \\ &\quad (R_{\text{mBO}}(x,y,\theta+\frac{\pi}{2}) - R_{\text{mBO}}(x,y,\theta-\frac{\pi}{2}))), \end{aligned} \tag{4.8}$$

with parameters the same as those of mEsDir neurons. Note that in these equations, $(R_{\text{mEsDir}}(x,y,\theta+\frac{\pi}{2}) - R_{\text{mEsDir}}(x,y,\theta-\frac{\pi}{2}))$ and $(R_{\text{mBO}}(x,y,\theta+\frac{\pi}{2}) - R_{\text{mBO}}(x,y,\theta-\frac{\pi}{2}))$ denote the direction of winning mEsDir and mBO neurons and the multiplication tests their direction alignment.

[2] As a reminder, at each visual location, a pair of mBO neurons with identical local feature selectivity but opposite side-of-figure preference are modeled. Between the pair, the neuron with stronger responses is called the winning mBO cell. Similarly, a pair of mEsDir at each visual field location with opposite direction selectivities can be considered. The mEsDire between the pair with stronger responses is called the winning mEsDir cell.

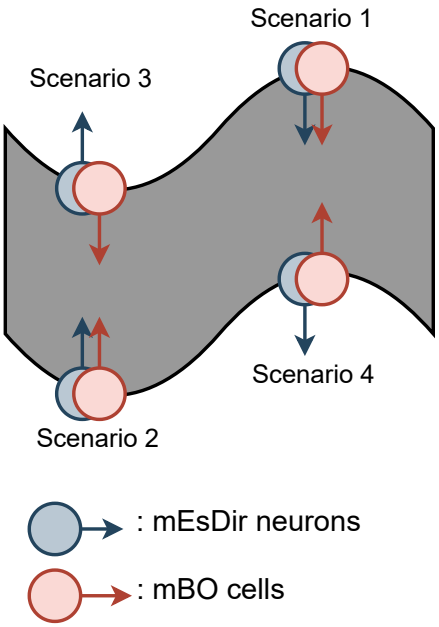


Figure 4.15: Curvature sign can be implemented with two pieces of information in hand: tangent vector derivative and the normal vector at each point along the bounding contour. These information are implemented with mEsDir and mBO neurons in SparseShape representing tangent vector derivative and normal vector directions respectively. At each visual field location, between pairs of winning mEsDir and mBO, four cases might happen. In two scenarios, where the direction of winning mBO and mEsDir, shown by an arrow attached to the RF, are the same, the contour segment corresponds to a convexity (See Equation 4.1). Whenever these neurons point toward opposite directions in the other two cases, the segment is concave.

4.3.1.4 Model Local Curvature

With mEsSign and mEsDeg neurons encoding curvature sign and magnitude, computing signed curvature, κ_s, is not difficult. The formulation in SparseShape is only slightly different from that of 2DSIL. In SparseShape, similar to mEsDir and mEsSign, signed curvature is encoded by a pair of neurons to represent positive and negative signs. At each scale in the hierarchy, a single map of mEsDeg and a pair of mEsSign (mEsSign_convex and mEsSign_concave) are implemented. Combining these maps yields pairs of local curvature cells with responses computed as:

$$\begin{aligned} R_{\text{mLocalCurv\_pos}}(x,y,\theta) &= R_{\text{mEsDeg}}(x,y,\theta) \cap \\ &\quad (R_{\text{mEsSign\_convex}}(x,y,\theta) > R_{\text{mEsSign\_concave}}(x,y,\theta)) \\ R_{\text{mLocalCurv\_neg}}(x,y,\theta) &= R_{\text{mEsDeg}}(x,y,\theta) \cap \\ &\quad (R_{\text{mEsSign\_convex}}(x,y,\theta) < R_{\text{mEsSign\_concave}}(x,y,\theta)). \end{aligned} \tag{4.9}$$

Similar to 2DSIL, the responses of mLocalCurv cells are combined by employing a max operation over all orientations. Consequently, the final layer feeding to mV4 neurons consists of eight curvature maps, four scales $\times$ two signs, representing four levels of acuteness for convexities and concavities.

4.3.2 Learning V4 Receptive Fields

In a way, Pasupathy and Connor had learned V4 receptive fields by fitting Gaussians to the biological data. However, as mentioned earlier, their model was simple and could not capture complex interactions between stimulus parts. Our goal is to learn V4 receptive fields in a way that complex and long-range interactions between shape parts, if they exist, can be captured. Recovering existence or equivalently lack of such interactions in V4 imparts significant insight into shape processing mechanisms in this visual area. For this purpose, we learn V4 receptive fields from the recordings provided to us by Dr. Anitha Pasupathy. Note that experimental data is limited with a total of 366 samples and data augmentation is not an option since any changes to the stimuli might result in changes in biological neuron responses.

In SparseShape, a naively-added fully-connected layer from mLocalCurv to mV4 cells has more than 14K weights to learn from less than 366 data points. To address the issue with the disproportionate number of parameters with respect to samples, we propose imposing sparsity priors that are compatible with discoveries of V4 [28] and other brain areas involved in shape representation [248]. Here, we leverage sparsity in a higher dimensional space and with a more relaxed model compared to the APC model, to learn V4 selectivities from their responses. The sparse coding procedure explained below is repeated for individual V4 neuron. Therefore, the equations that will appear below, will hold for each individual V4 cell and the procedure for learning the receptive fields will be repeated for each V4 neuron. With the learned RFs, mV4 responses to any stimuli set is a simple feedforward pass (with dorsal recurrence) in SparseShape.

Our proposed sparse coding method formulates V4 RF recovery as a supervised learning problem. Specifically, given the responses of mLocalCurv cells and biological V4 responses to a stimuli set, we seek a sparse combination of curvature components across the receptive field that can explain V4 responses. These contributions are weighted. Having found the desired curvature components and their weights, responses of the mV4 neurons corresponding to the biological V4 cell to an

arbitrary stimulus s can be obtained by:

$$R_{\mathrm{mV4}} = D_s^T \gamma, \tag{4.10}$$

where γ is the weight vector and D_s is a vector whose elements signal presence or absence of a shape part at a particular position within the receptive field. For example, the i-th element of D_s denoted as $d_s(i)$ might indicate presence of a broad convex part at the right top side of the receptive field for stimulus s. Consequently, D_s will be referred to as the part-based vector. To account for a variety of positions for shape parts, a 3×3 grid over the receptive field with Gaussian kernels in each cell of the grid filters each mLocalCurv map. The parameters of each Gaussian characterize the position and extent of a particular curvature component within a cell. Putting these together, $d_s(i)$ is computed as:

$$d_s(i) = \sum_{x,y} \mathbf{C}(x, y, s, k) \cdot \mathbf{G}(x, y, j; \mu_j, \Sigma_j), \tag{4.11}$$

with $\mathbf{C}(x, y, t, k)$ corresponding to the responses in the k-th mLocalCurv map to stimulus s at visual location (x, y) with $k \in \{1, 2, \ldots, 8\}$. In this equation, $\mathbf{G}(x, y, j; \mu_j, \Sigma_j)$ represents the Gaussian in the j-th cell with parameters μ_j, Σ_j and $j \in \{1, 2, \ldots, 9\}$. With 8 mLocalCurv maps and 9 cell positions, D_s and γ in our implementation are 72-dimensional vectors.

Minimizing the objective function:

$$\min_{\gamma, \mathbf{V}} \sum_{t=1}^{\tau} \frac{1}{2} \|R_t - D_t^T \gamma\|_2 + \lambda \|\gamma\|_1, \tag{4.12}$$

where R_t is the biological V4 response to stimulus t and τ is the number of stimuli in the training set results in a sparse combination of parts. In this equation, $\mathbf{V}$ is the set of parameters of the Gaussian kernels and λ balances between sparsity and the error term.

The diagram in Figure 4.16 shows the different components of our sparse coding formulation. Note that the green box in this figure corresponds to the green arrow in Figure 4.13. In order to perform the optimization, we alternate between optimizing for the sparse code vector γ and the parameters of the Gaussian kernels $\mathbf{V}$. Optimization over the parameters of the Gaussian kernels is similar to parameterized dictionary learning algorithm explained in [243]. The details of our optimization can be found in Appendix B.

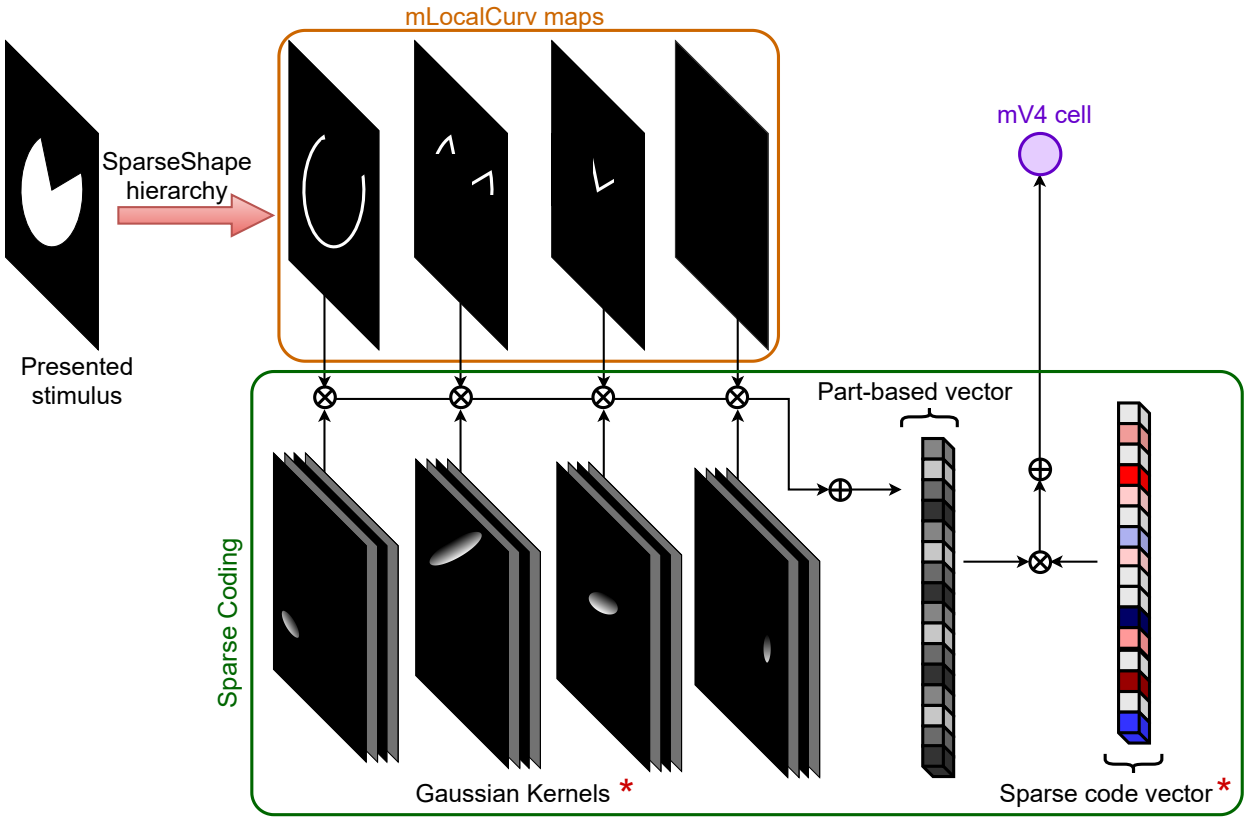


Figure 4.16: The proposed sparse coding step in SparseShape. For pairs of $(\{\text{mLocalCurv}\}, \text{mV4})$, to a set of presented shapes, the algorithm looks for a combination of shape conformations that could best explain mV4 responses. To account for the various positions of shape parts, a set of 9 Gaussian kernels spanning a 3×3 grid over the RF filter mLocalCurv responses at each cell. These filtered responses form the part-based vector. A weighted combination of part-based vector elements yields mV4 responses. These weights are determined by the sparse code vector. The green box in this figure, depicting the different components of our sparse coding, corresponds to the green arrow in the SparseShape architecture shown in Figure 4.13. The elements in this box that are identified with red asterisks are the parameters that are learned in this step.

Although our optimization formulation looks strikingly similar to that of Equation 4.2, there are three major differences:

1. Our proposed approach learns a mapping from mLocalCurv maps to V4 responses in a *supervised* manner.

2. In our formulation in Equation 4.12, one single sparse vector is learned for all the training samples for a given V4 cell ($\min_\gamma$ *outside* the sum in Equation 4.12).

3. Whereas the unsupervised algorithm iterates between feature learning and feature selection steps, in our formulation, the features are known and instead, both steps in each iteration

perform feature selection.

Despite these differences, we make use of advances in the sparse coding algorithms in ML as explained in Appendix B.

4.4 Evaluation

To evaluate the proposed model and learning approach, two categories of experiments were carried out. In one set of experiments, the goal was to evaluate model performance to the stimuli shown in Figure 4.6 and examine what the learned model could reveal about V4 cells. We refer to any experiment performed on shapes shown in Figure 4.3 as a standard experiment. The second category consists of an experiment with a new stimuli set to test scale invariance in the learned mV4 neurons, similar to the reported scale invariance in V4 responses [29].

4.4.1 Stimuli and Data

Two sets of stimuli were employed. In the standard experiments, the stimuli in Figure 4.6 from the study by Pasupathy and Connor [20] with 366 closed shapes were employed. These shapes, which we call standard shapes, were scaled such that stimulus edges were offset from the RF center by 0.75× RF size. The code to generate the shape as well as the V4 recordings from [20] were kindly provided by Dr. Anitha Pasupathy. In the scale invariance experiment, we generated stimuli similar to those in [29] by using the code from the standard experiments. The scales were matched to those in the V4 scale invariance study [29].

Similar to the RBO network, input to the network was set to 400×400 pixels. We measured 1° visual angle at 50cm to be 32×32 pixels for our stimuli size. Model V4 RFs were set to 4° following [92] equivalent to 128 × 128 pixels.

4.4.2 Training Procedure

SparseShape parameters in the last layer are learned through the sparse coding step. The parameter λ in Equation 4.12 is set using cross-validation over the training set for each neuron. Having determined the optimum lambda, the model neuron is trained on the entire training set. All

other non-learnable parameters, those known to us through previous experimental findings, are set according to RBO and 2DSIL.

To learn the model weights from mLocalCurv maps to mV4 cells, the shapes in the standard experiments were divided into a 60%-40% train and test splits. This procedure was employed to report all the experimental results below except for one: where we compare our model with the HMAX network, we follow their cross-validation steps. Specifically, in HMAX, they carried out a six-fold cross-validation over the whole stimuli set and reported the average correlation coefficients. This way, each shape will appear in the test set once and therefore, responses of the model to all the shapes in the set will be in hand. For comparison purposes, we adhered to their procedure.

The data from the biological V4 cells shows a variety of selectivities. For example, one V4 cell might be selective to sharp convexities at the left middle part of its RF while another is strongly activated with broad concavities at the bottom right side of the RF. Due to this diversity in V4 selectivities, stratified division of shapes into train-test sets was executed for each individual neuron to ensure an assortment of responses are present in each set. Our stratification procedure considered three folds of shapes in the stimuli according to weak, medium and strong activations of the neuron. Consequently, during learning, the model is fit with respect to shapes to which neuron responses span weak to strong activations.

4.4.3 Quantification

To quantify model performance, two measurements with respect to biological V4 responses are reported: mean error and correlation coefficients. The mean error comparison, following [16], computes the mean absolute difference between the normalized responses of each mV4 and V4 cell as:

$$\text{mean error} = \frac{\sum_{i=1}^{N} |R_{mV4} - R_{V4}|}{N}, \tag{4.13}$$

where N is set according to the number of stimuli in each of train and test sets. Similarly, correlation coefficients (Pearson's r) are computed separately for train and test sets between model and biological responses.

4.4.4 Comparison with Biological Responses, 2DSIL, and HMAX

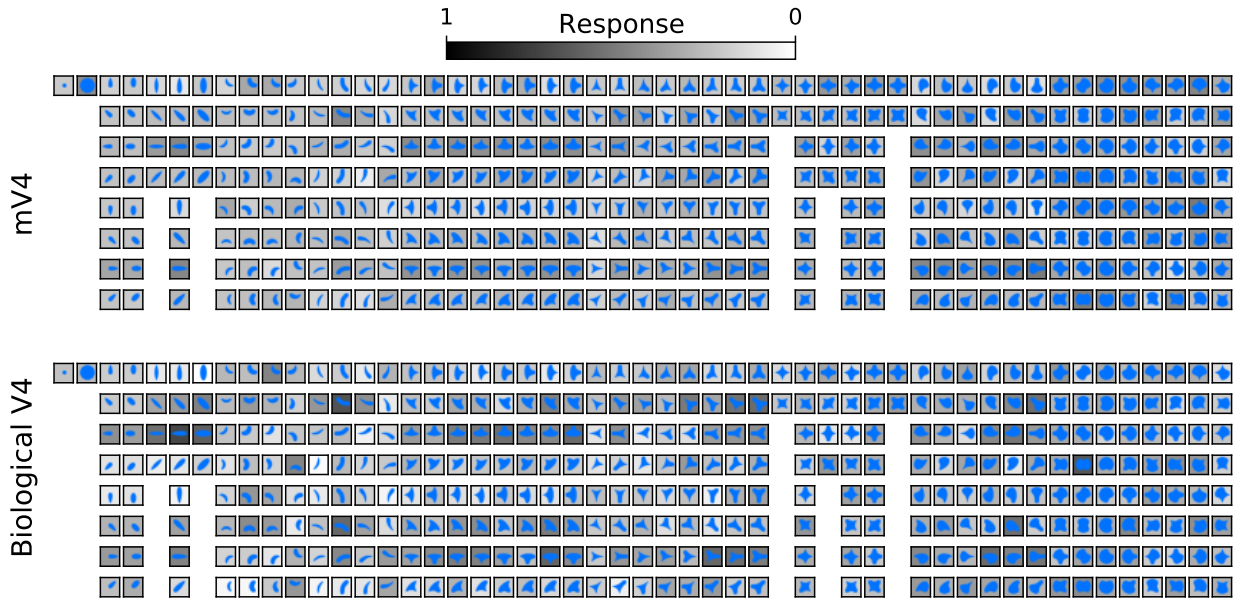
To evaluate the performance of our model, we presented standard shapes in the training set to the SparseShape network and recorded mLocalCurv maps. For each of the 109 biological V4 recording, we assigned responses to a corresponding mV4 cell. We call responses of biological and model V4 cells as biological V4 and mV4 responses respectively. Then, given the curvature maps and biological V4 responses, we learned a sparse mapping from mLocalCurv responses to mV4. Having learned the sparse mapping including the Gaussian maps from Equation 4.11 and the sparse code vector, we determined responses of each mV4 cell to unseen shapes as explained in Equation 4.10. Examples in Figures 4.17, 4.18, 4.19 depict responses of mV4 and biological cells to all the 366 stimuli in the standard shape set.

In these figures, each shape in the standard stimuli set is shown as a blue icon within a square. The ground intensity in each square signifies the strength of cell response to the shape in that square. Overall, these figures demonstrate similar responses between mV4 and V4 neurons.

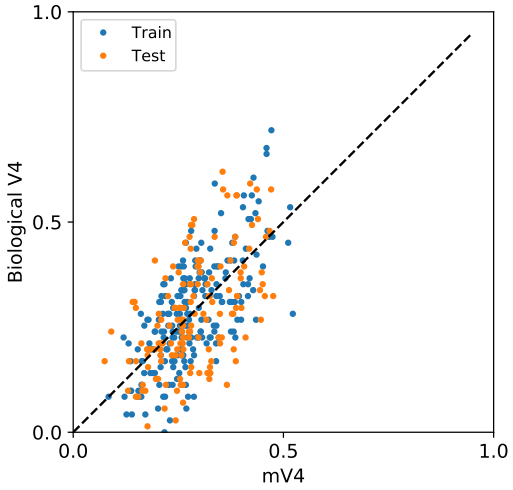
To measure this similarity, following 2DSIL, we computed the mean errors from Equation 4.13 for train and test sets separately. Additionally, to evaluate the performance improvement of the proposed steps compared to 2DSIL, we plotted mean errors of 2DSIL alongside ours. Figure 4.20 summarizes train and test mean errors for 109 mV4 cells in SparseShape and the mean error for 75 model neurons for which a shape template could be found in 2DSIL. This figure demonstrates that compared to 2DSIL, SparseShape mean errors are decreased compared to all the 75 2DSIL model neurons. Additionally, the remaining 34 cells follow a similar trend in train and test mean errors. Over the 109 mV4 SparseShape cells, the mean error is less than 0.18 whereas the 2DSIL error gets as high as 0.34 for some of the neurons. Mean errors from HMAX were not available and therefore, this model is missing in Figure 4.20.

Both APC and HMAX reported correlation coefficients as a goodness-of-fit measure. While the APC fitting was based on responses to all the 366 stimuli, HMAX employed a six-fold cross-validation and reported the average correlation coefficient (r). As explained in Section 4.4.2 and only for this part of our experiment, we report the average r over six-fold cross-validation over the whole stimuli set for fair comparison with HMAX.

One goal in this chapter was to relax the strong prior imposed by Pasupathy and Connor to



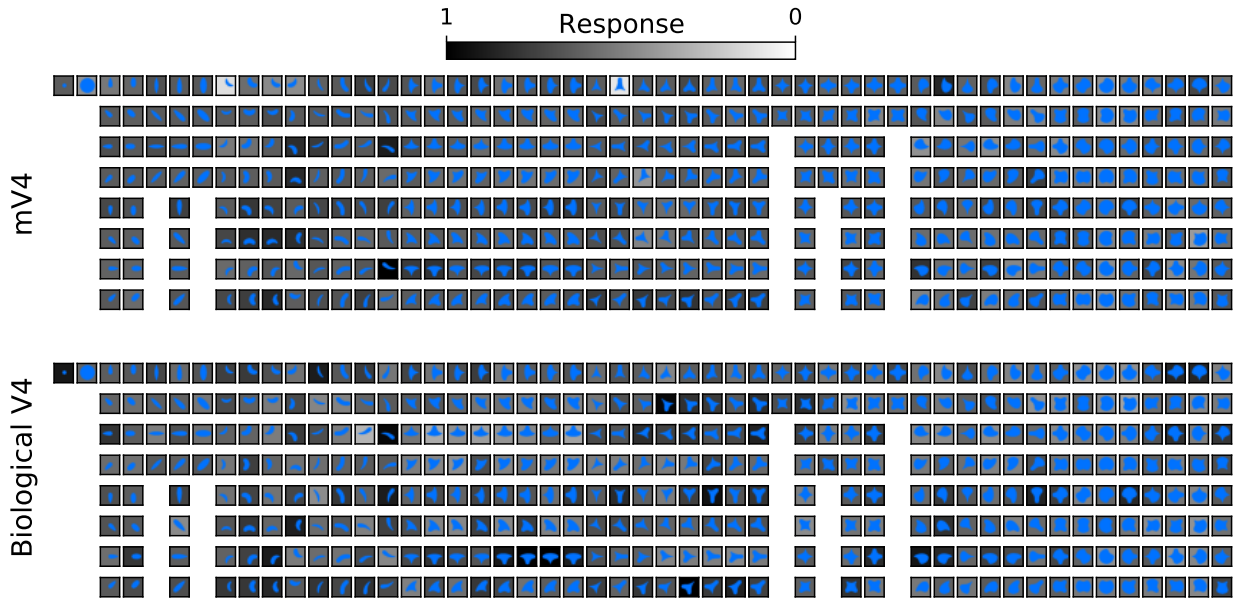
(a)



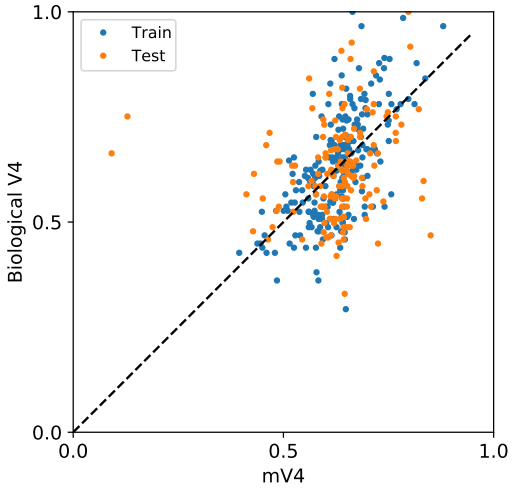
(b)

Figure 4.17: Responses of a learned mV4 cell. (a) shows two rows of mV4 and biological V4 responses. Each shape stimuli is depicted as a blue icon in a square. The intensity of the ground in each square indicates the response strength. The mV4 cell in this figure exhibits a similar pattern of responses as that of the biological V4 cell. (b) Biological V4 responses plotted against mV4 activations illustrates a strong correlation between these responses ($r = 0.69$).

recover neuron tunings. To compare the proposed approach, we also consider the APC model. Pasupathy and Conner also evaluated a four-dimensional APC model (3 curvatures and 1 angular position) by considering the two neighboring curvature components at either side of a shape part. We call this model as APC–4D and the 2D version as APC–2D. For completeness, we also considered



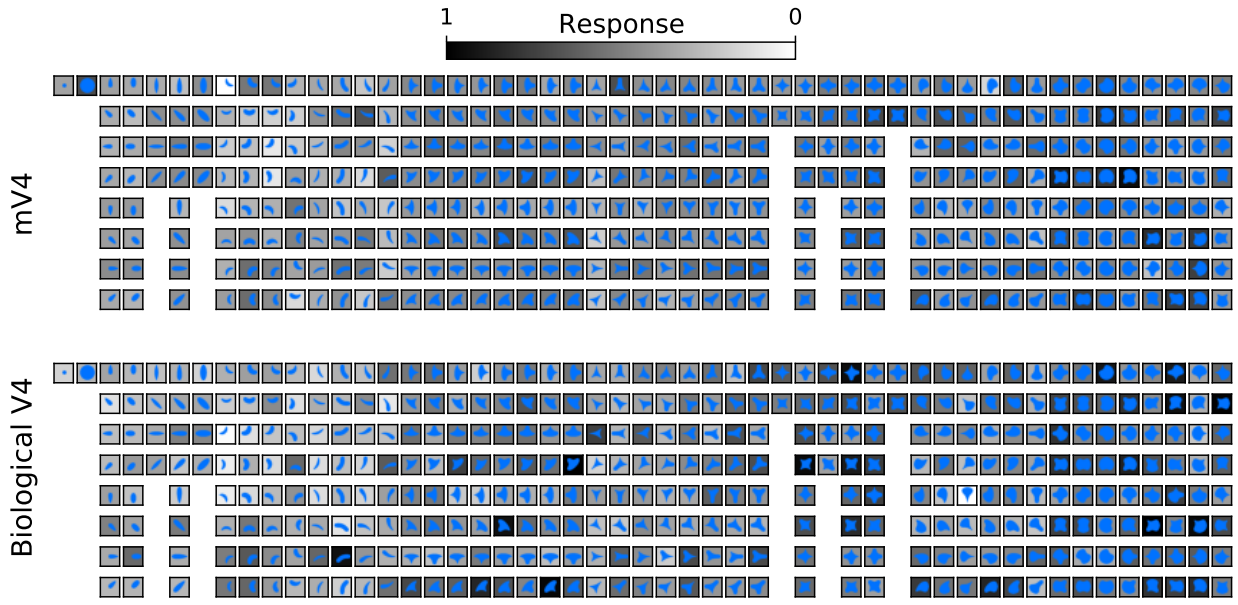
(a)



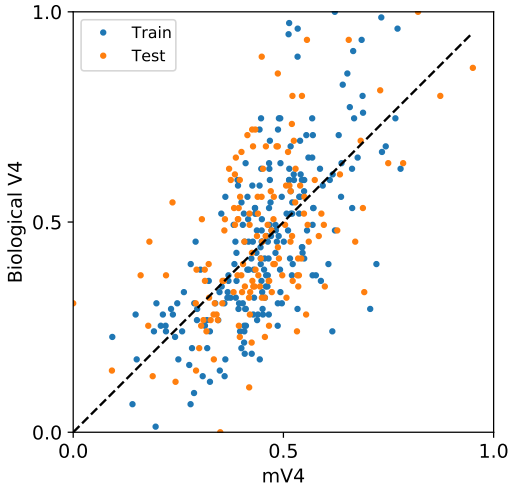
(b)

Figure 4.18: Responses of a learned mV4 cell. (a) shows two rows of mV4 and biological V4 responses. Each shape stimuli is depicted as a blue icon in a square. The intensity of the ground in each square indicates the response strength. The mV4 cell in this figure exhibits a similar pattern of responses as that of the biological V4 cell. (b) Biological V4 responses plotted against mV4 activations illustrates a strong correlation between these responses ($r = 0.63$).

the correlations for 100 random neurons selected from convolutional layer 2 (conv2) and fully-connected layer 7 (FC7) of Alexnet reported by Pospisil *et al.* [56]. Figure 4.21 illustrates the correlation coefficients for all the APC–2D, APC–4D, Alexnet–conv2, Alexnet–FC7, HMAX, 2DSIL and SparseShape models. A ridgeline plot separates theses distributions according to their overlap



(a)



(b)

Figure 4.19: Responses of a learned mV4 cell. (a) shows two rows of mV4 and biological V4 responses. Each shape stimuli is depicted as a blue icon in a square. The intensity of the ground in each square indicates the response strength. The mV4 cell in this figure exhibits a similar pattern of responses as that of the biological V4 cell. (b) Biological V4 responses plotted against mV4 activations illustrates a strong correlation between these responses ($r = 0.70$).

for easier comparison. This figure shows that SparseShape and HMAX have similar correlation coefficient distributions. In contrast, 2DSIL correlations are shifted to smaller values compared to both train and test plots for SparseShape.

The APC distributions are most interesting. Note that the APC model is not a hierarchical

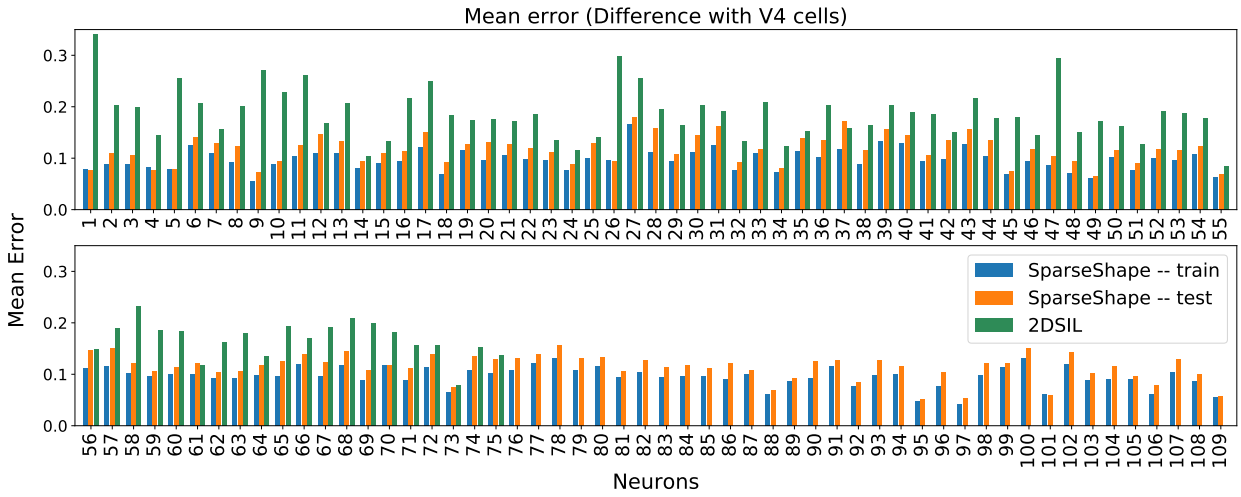


Figure 4.20: Mean error in responses of model V4 cells for 2DSIL and SparseShape. Neuron errors are plotted in two rows for better visualization. SparseShape improves the performance measured as the mean difference of responses with biological V4 cells compared to 2DSIL.

model and represents each stimulus in the suggested angular position and curvature space. By considering signed curvature in their model, they encode an object-centered representation and in that sense the APC model is most similar to SparseShape representations. Although the APC model was fitted to the complete data, the correlation distributions in Figure 4.21 show that the APC–2D correlations are smaller than that of SparseShape–test. Combining three shape components in APC–4D shifts the correlations to larger values. This shift in distribution with incorporating more than a single contributing part, as Pasupathy and Connor pointed out, suggests V4 cells encode complex configurations within their RF. Still, APC-4D distribution matches that of SparseShape–test and the shift to larger correlations in SparseShape–train suggests a more complicated pattern of interactions between shape parts than three neighboring ones within V4 RFs.

Figure 4.21 shows Alexnet–conv2 and Alexnet–FC7 correlations are comparable to those of APC–4D with slightly larger correlations in FC7 neurons. Alexnet is a blackbox model as discussed in Chapter 1 that does not lend insight into processing in the ventral stream. Moreover, Alexnet is a feedforward network and the same argument on object-centered representations due to global context applies to this network.

Table 4.2 summarizes the median correlation coefficients for all the distributions plotted in Figure 4.21, confirming comparable distributions between HMAX and SparseShape. Despite comparable distributions, HMAX representations are not object-centered. When the goal is to replicate

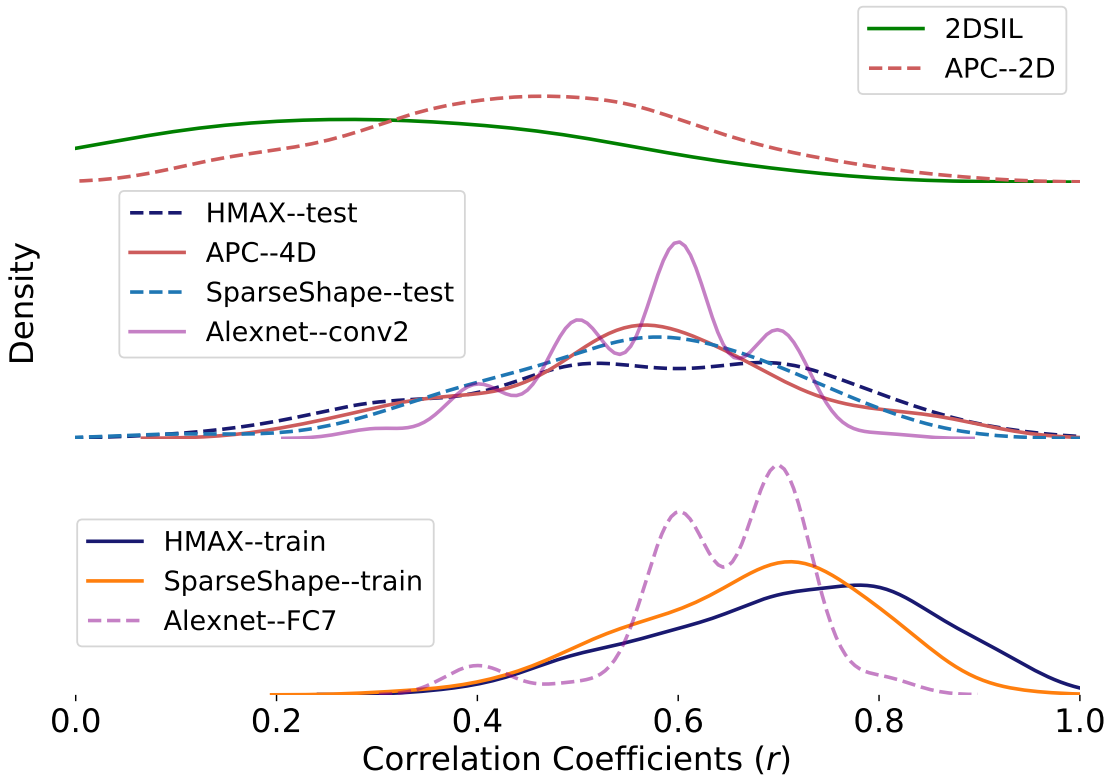


Figure 4.21: Population distribution of correlation coefficients (r) for Alexnet, HMAX, APC, 2DSIL and SparseShape. The proposed approach improves the correlation coefficients compared to 2DSIL and the APC model. The standard two-dimensional APC is identified as APC–2D. Pasupathy and Connor [20] also considered the two neighboring conformations for each shape parts presenting shapes in a four-dimensional space. This presentations is called APC–4D. The plot illustrates that by incorporating the influence of more than one conformation to V4 responses, the correlations are improved. Note that in both APC–2D and APC–4D, the model fitting is performed with all the 366 stimuli. Both conv2 and FC7 layers in Alexnet perform comparable to APC–4D with a slightly higher correlations in FC7 neurons. Note that Alexnet is a feedforward network and the same argument on global context for modeling convexities and concavities applies to this network. HMAX is the model with best correlations. As explained in the text, this superior performance could be due to overfitting, although this conjecture needs further investigation. Additionally, since HMAX compresses V2 and V4 processings into a single layer that is learned using a heuristic greedy approach, this network does not match the rest of the models in terms of explanatory capacity.

the observed behavior of V4 neurons, this model lacks a proper explanation of V4 shape processing mechanisms. Additionally, in a study, Bair *et al.* [249] reported that the learned HMAX neurons could not achieve the levels of invariance to scale and position observed in V4 neurons. Together, these results suggest that HMAX, despite impressive correlation coefficients, does not replicate V4

Model – Set	Correlation Coefficient
HMAX–train	0.72
HMAX–test	0.57
SparseShape – train	0.70
SparseShape – test	0.56
2DSIL	0.28
APC –2D	0.46
APC–4D	0.57

Table 4.2: Medians of correlation coefficients for the models which population distribution was shown in Figure 4.21. HMAX–train is the best performing model followed by SparseShape–train.

responses.

As a summary, our results show that mV4 responses of SparseShape are highly correlated with biological V4 responses with small mean errors. In contrast to HMAX and 2DSIL, mV4 representations are object-centered. Furthermore, comparison with APC–4D confirms complex interactions with multiple shape part contributions to V4 responses.

4.4.5 On Sparsity

The APC–2D and APC–4D models take one and three contributing shape parts into account respectively. Our sparse coding scheme, compared to APC, is more relaxed in incorporating more shape parts and not only those in spatial proximity, but also components across the RF. Asking the question of how many shape components were determined as important to shape encoding in V4, we examined the sparsity of the learned sparse code vector by considering the percentage of its nonzero elements for each of the 109 neurons. To better evaluate the impact of convexities versus concavities, we calculated the nonzero percentage separately for each of convex and concave parts. The plot in Figure 4.22 demonstrates the distribution of convex/concave sparsity for the population. Each neuron is displayed with a dot and when neurons coincide in this space, a single dot is shown. The distribution of dots in this space suggests contributions from more than a single or even three shape parts to V4 responses. Moreover, the marginal distributions on the top and right sides of this plot suggest that most V4 cells integrate a higher percentage of convex part compared to concave components. These results are in agreement with convexity-bias observations reported in [19, 28].

With more convex parts feeding to V4 cells, one possibility is that the weights from these con-

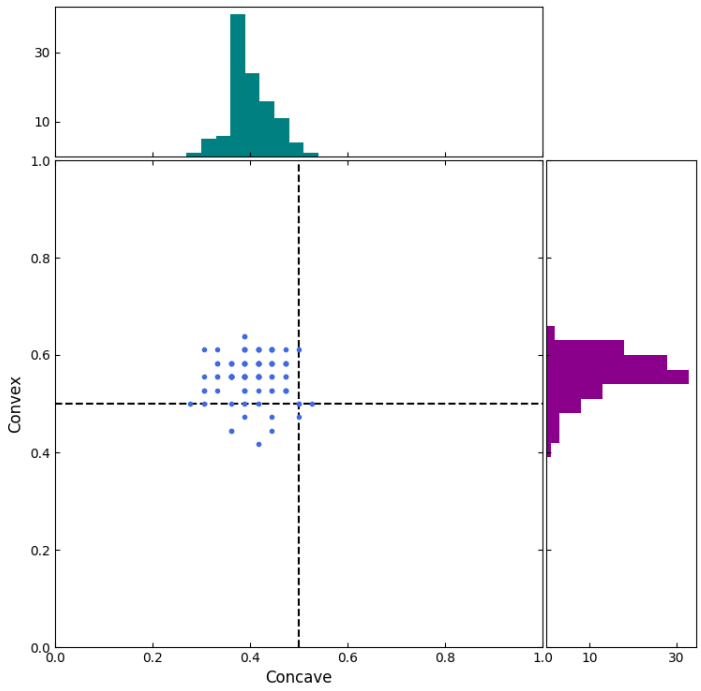


Figure 4.22: Sparsity analysis. For each neuron in the mV4 population, we calculated the percentage of nonzero convex and concave elements of the sparse code vector. Each neuron is shown with a dot and when neurons have matching sparsity, a single dot is shown. This figure demonstrates that more than a single or three curvature components contribute to V4 activations. Moreover, the skewness in the distribution of convex components (right side histogram) confirms previous observations of bias toward convexities in V4 neurons [19, 28].

formations is smaller compared to the fewer concave parts impacting V4 responses, compensating for this imbalance. To evaluate this possibility, we did the following: for each neuron: we normalized the sparse code vector with the maximum set at 1. Then, for each mLocalCurv map, we took the maximum weight from the map to mV4. This process will result in a vector of length 8 corresponding to the eight local curvature maps for each neuron. Then, for the population of mV4 neurons, we plotted the distributions of weights for each local curvature map. If weights from concave maps were larger compared to those of convex ones, then, one would expect the concave distributions to be shifted to larger weights. The eight distributions for the mV4 population is illustrated in Figure 4.23. Interestingly, all distributions are relatively overlapping, except for the one for acute convexities that peaks at one. This plot, along with that of Figure 4.22, confirms bias toward convexities in the learned model.

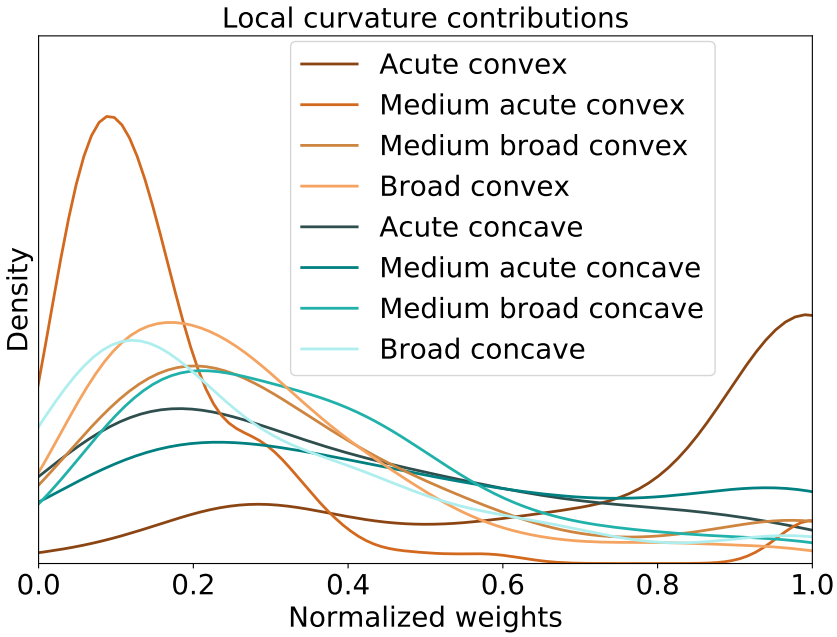


Figure 4.23: Distributions of normalized weights from mLocalCurv maps to mV4 neurons separated according to signed curvature. overall, all distributions overlap except for the acute convex weights that peaks at one. This plot in conjunction with results from Figure 4.22 indicate a bias toward convexities in mV4 neurons.

4.4.6 Facilitatory versus Inhibitory Contributions

In all other models, including APC, contributions to V4 responses are limited to facilitatory ones. If V4 neurons indeed integrate V2 responses in both excitatory and inhibitory manners, these models neglect accounting for important contributions to V4 responses. In SparseShape, both types of contributions are accounted for in the model and a combination of both types of effects in V4 activations are explored during learning.

In order to assess the effect of each type of contribution to mV4 responses, we evaluated those effects in the learned weights in the sparse code vector. Specifically, similar to above, we normalized the sparse code vector for each neuron. Then, we plotted the distribution of the sum of facilitatory and inhibitory weights, separately for convex and concave components, for the population of mV4 neurons. These four distributions are depicted in Figure 4.24 demonstrating overlapping distributions with a slight shift for facilitatory convex weights. Overlapping population distributions indicate similar effects from both facilitatory and inhibitory contributions, suggesting that without

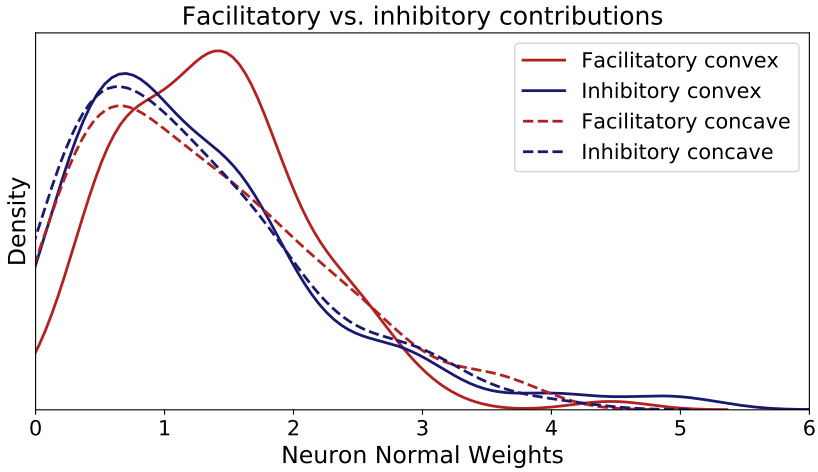


Figure 4.24: Facilitatory versus inhibitory weight distributions. The distribution of facilitatory and inhibitory contributions from shape parts greatly overlap, with a slight shift for facilitatory convex weights, suggesting both types of contributions determine V4 responses.

taking inhibitory parts into account, our understanding of V4 shape processing is incomplete.

4.4.7 Visualizing Receptive Fields

An advantage of combining an analytically-defined model with a learning step is effortless understanding of the learned features. Visualizing the recovered receptive fields is a means to that purpose, which helps the goal of furthering our understanding of shape selectivities in V4.

In APC and 2DSIL, visualizing the recovered RFs is not difficult. In HMAX, however, absence of intermediate processing stages encapsulates all those layers into a single learned layer. This procedure yields recovered receptive fields that are difficult to interpret. For example, consider the recovered V4 receptive fields by the HMAX model in Figure 4.8. Aside from the fact that no curvature-like combinations of C1 units were learned in some of the cells (see those denoted with red squares), examples could be found which combination of C1 units appear difficult to interpret in terms of curvature components (green-squared examples in Figure 4.8).

Unlike HMAX, the more abstract representation of signed curvature in the mLocalCurv layer of SparseShape makes the task of explaining the recovered receptive fields effortless. For example, Figure 4.25 shows an example neuron responses and its biological V4 versus mV4 response plot along with a visualization of its recovered receptive field. One look at this RF indicate selectivity to acute to mild convexities in the left and bottom part of the RF with neuron responses inhibited

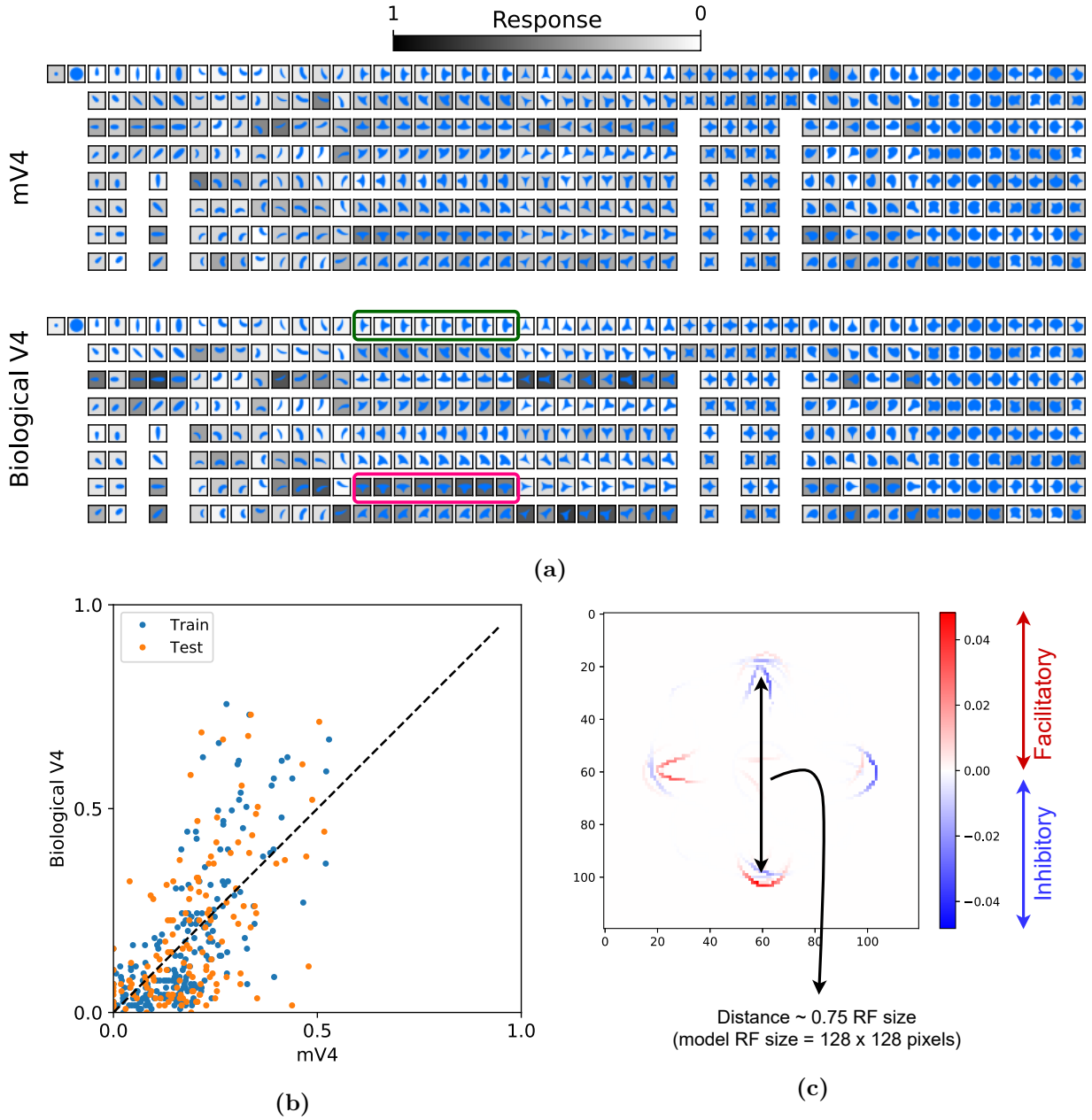


Figure 4.25: Responses of an example neuron and its recovered RF. (a) Responses of a mV4 cell are shown. The shapes inside the magenta box evoke strong activations in the biological neuron. Same shapes but rotated at 45deg inside the green box do not excite this cell. (b) Biological V4 responses plotted against mV4 responses. (c) The recovered RF of the neuron using the proposed approach with red and blue indicating facilitatory and inhibitory contributions. Shapes with acute and medium broad convexities at the left and bottom part of the RF respectively, evoke strong activations in the cell. In contrast acute and medium broad convexities in the top and right parts of the RF inhibit neuronal responses. This RF shows a relative clustering of excitatory and inhibitory parts raising the question of if neurons V4 neurons exhibit contiguous excitatory/inhibitory regions. This recovered receptive field also displays that shape conformations from the spatial extent of the neuron's RF are combined to account for its selectivity. For example, in this RF, the distance between two contributing components is as large as 0.75 $\times$ RF. Such long-range interactions between shape conformations were not captured in the APC model due to fitting a single Gaussian to neuron responses.

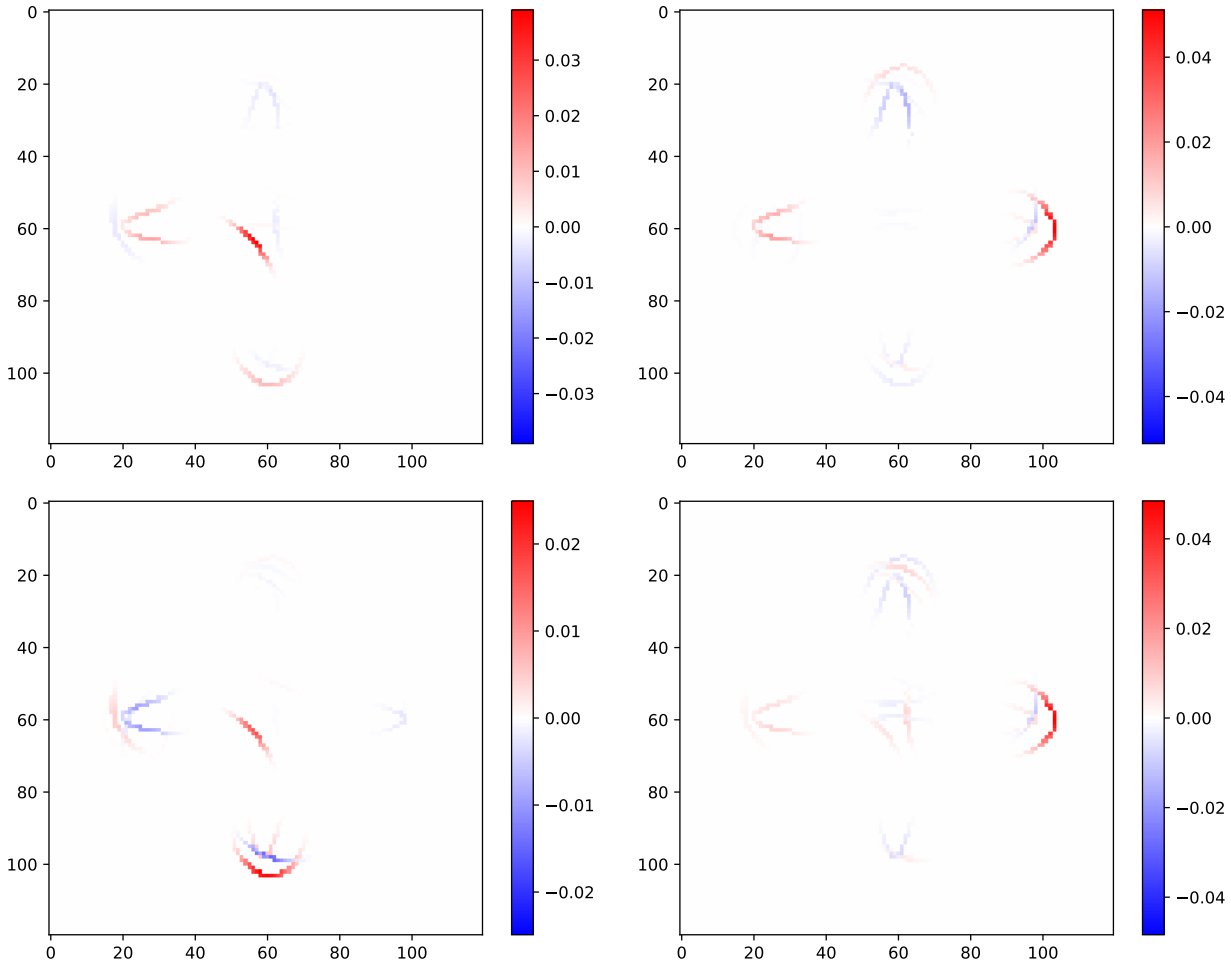


Figure 4.26: Examples of learned receptive fields in SparseShape. In each plot, red and blue shape parts in the RF are for facilitatory and inhibitory contributions respectively.

with appearance of acute and mild convexities on top and right parts of the RF respectively. The recovered RF can be qualitatively verified with the biological V4 responses in Figure 4.25a. For example, the shapes within the magenta rectangle in Figure 4.25a have two convex parts on facilitatory positions and only one part matching an inhibitory component causing relatively strong activations of the cell. The same shapes rotated with two convex part within the inhibitory parts of the RF (encompassed with a green rectangle) have less success in activating the neuron. A few more examples of recovered RFs are presented in Figure 4.26.

Looking at the recovered RF in Figure 4.25c presents an interesting observation: the contributions to this neuron's responses are from the full spatial extent of the cell. As the arrow in Figure 4.25c shows, the distance between interacting shape parts is almost 0.75 $\times$ RF size. In order

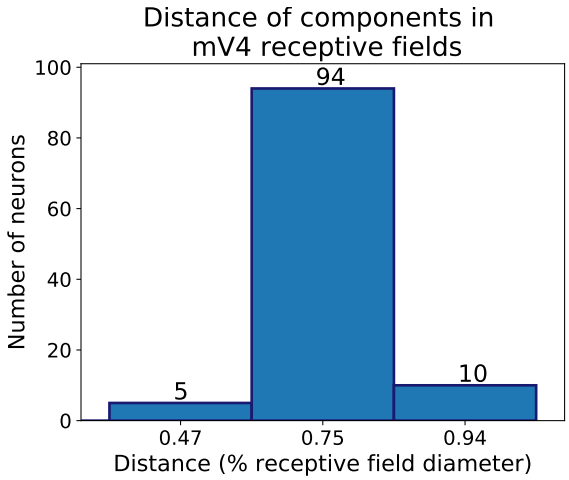


Figure 4.27: Histogram of distance of shape parts with RF of mV4 cells. The majority of mV4 cells (104 out of 109) integrate shape parts that are apart as far as three forth of the RF diameter, suggesting V4 neurons incorporate V2 responses from the full span of their RF.

to quantify the extent of shape part interactions within the RF, we did the following: for each neuron, we computed the distance of all pairs of contributing parts in the RF. For each cell, we took the maximum distance among all parts and plotted the distance histogram for the population of mV4 neurons as shown in Figure 4.27. As the figure illustrates, the majority of neurons integrate shape parts from as far as two-third of their RF diameter. Such interactions across the RF could not be accounted for in the APC model unless extending its representation to a nine-dimensional space (for eight curvature parts and one angular position) and only for excitatory ones.

4.4.8 Scale Invariance

Studying V4 responses to stimuli of various sizes, El-Shamayleh *et al.* [29] found these cells encode shapes in a size-invariant manner. More specifically, they reported that V4 neurons encode "normalized" rather than "absolute" curvature of shape parts. Figure 4.28 depicts how absolute curvature changes with variations in shape scale with normalized curvature remaining unchanged. They prepared a stimuli set of parametric shapes at various scales and hypothesized that if V4 cells were sensitive to shape size, their tuning peaks would shift with scale. However, if these neurons exhibit scale invariance, their tuning peak would not shift with changes in scale. Examples of schematic responses for these two hypotheses are presented in Figure 4.29. To test for the two hypotheses, they measured the shift in tuning centroids as a function of scale. The size-dependent

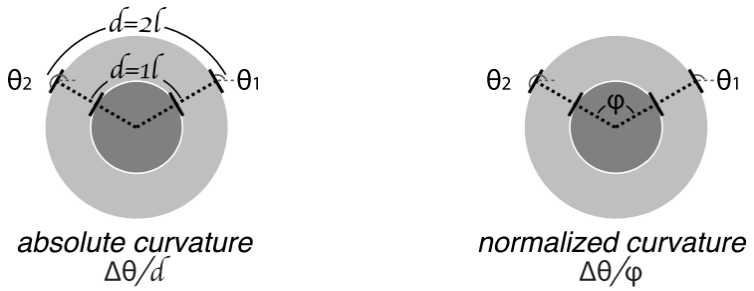


Figure 4.28: When a contour is scaled, its shape is preserved but its curvature changes. To test whether V4 neurons are sensitive to scale or encode curvature as a size-invariant property of shape, El-Shamayleh *et al*. [29] defined absolute curvature as the rate of change of the tangent angle per unit of contour length (left) and normalized curvature as the rate of change of the tangent angle per unit of angular length (right). These two examples show that absolute curvature is size-dependent while normalized curvature is preserved with changes in shape scale. Figure adapted from [29].

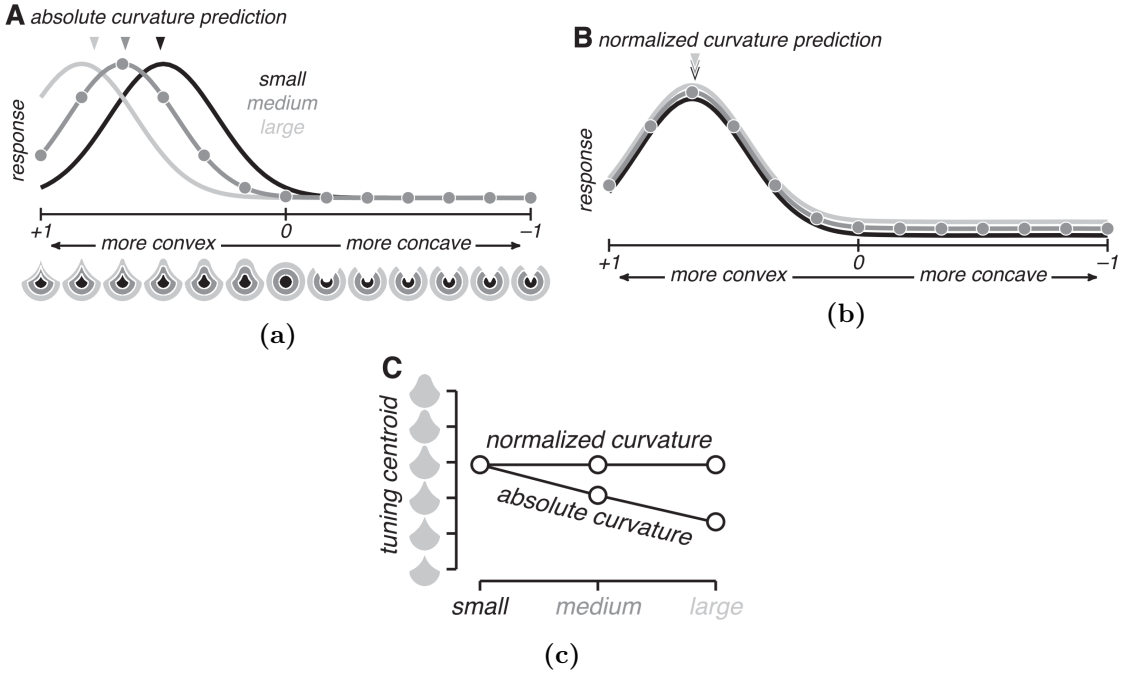


Figure 4.29: Predictions for selectivity to absolute and normalized curvature. (a) If representations in V4 neurons are size-dependent, then, with changes in shape scale the tuning centroids shift. (b) A scale-invariant encoding in V4 neurons keep the tuning centroids to changes to shape scale. (b) With the tuning centroids plotted as a function of scale, the scale-invariant hypothesis predicts yielding a line with zero slope. The absolute curvature hypothesis predicts a non-zero slope for the same function. Figure adapted from [29].

hypothesis would yield a non-zero slope while the slope would remain close to zero in the scale-invariance hypothesis (See the example plot in Figure 4.29c). For their population of neurons, they found a notable peak around zero in the tuning centroid slope histogram. To examine if the

observed invariance in responses could be attributed to the shift in the position of boundary conformation, they performed a similar experiment with only the convex shapes at various positions within the RF. Having found the slope of the tuning centroids peak around zero for the population, they concluded that the observed scale invariance could not be caused by the changes in position.

We performed a similar experiment on the learned mV4 cells. For this purpose, using the code provided to us for the standard stimuli set, we constructed shapes similar to those employed by El-Shamayleh *et al*. [29] and evaluated systematic shifts in tuning centroids in mV4 cells. Figure 4.30 demonstrates a few examples of mV4 neuron responses to changes in shape size. These neuron, as is evident from the plots in Figure 4.30, exhibit scale-invariance representations. For the population of 109 mV4 cells learned in SparseShape, we evaluated the tuning centroid slope distribution to changes in size and position and we found that similar to biological V4 cells, our learned mV4 neurons encode "normalized" rather than "absolute" curvature and are position-invariant. Figure 4.31 illustrates two histograms with peaks at zero for scale and position invariance test.

Results from this experiment are interesting in that our mV4 neurons were trained with a single scale stimuli set. However, unlike HMAX whose responses undergo substantial changes with variations in scale and position [249], the learned weights in SparseShape accounted for that variation in responses of biological V4 cells.

4.5 Discussion

Our goal was to demonstrate how combining an analytical model with learning can facilitate use of biologically-inspired models when there is an evident lack of knowledge about certain processing aspects. To this end, we studied shape processing mechanisms in V4 and showed that by combining an analytical model for shape processing with a simple learning method, unknown shape processing mechanisms in this visual area could be recovered.

Despite years of research, shape processing mechanisms in particular brain areas are still an enigma. While we know neurons in V1 are selective to oriented edges and bars and V2 cells encode endstopping, how V4 neurons achieve their object-centered shape representations is still unknown. To suggest a plausible processing flow of the visual signal to such encodings, we proposed the SparseShape network with all the early and intermediate processing layers implemented.

Our model and the geometrical interpretations of our model components demonstrate how combining previously discovered representations, namely endstopping and side-of-figure, in the brain are essential components for the object-centered encoding reported in V4 neurons. Removing any of the components from this unified framework disturbs the object-centered representations. Consequently, with missing components in previously introduced models, those approaches were not successful in encoding such a representation. Additionally, combining SparseShape with a simple learning method revealed more than what was previously discovered about V4 cells: our results indicated contributions from multiple parts from the spatial extent of the receptive field in both facilitatory and inhibitory manners in V4 responses. In short, an important contribution of this work was to relax tuning priors and reveal more by revisiting data from nearly two decades ago.

Although HMAX matches our performance in correlation coefficients with biological V4 responses, due to lack of modeling intermediate processing stages corresponding to V2, their network can be viewed as a gray-box (if not a black-box) for shape processing. Additionally, HMAX failed to exhibit similar levels of invariance to changes in position and scale as V4 neurons [249].

Even modeling intermediate stages is not adequate when important mechanisms are absent from a model. We demonstrated that the 2DSIL hierarchy required additional mechanisms to incorporate global context that enable curvature sign encoding. Without such mechanisms, the model is not complete. Moreover, we showed that the extended model, with assistance from the learning step, could substantially improve the prediction performance compared to that of 2DSIL.

When only the processing mechanisms in V4 are the focus of research, some might behold no added value in modeling all the stages from V1 to curvature encoding in V4. In that case, a simple transformation of shapes to a more abstract space, such as the APC model, might suffice. In that case, one must employ such models with care. For example, our various experiments demonstrated that a strong prior such as fitting a single Gaussian in APC could neglect recovering certain contributing factors such as inhibitory ones from across the RF. We investigated the effect of loosening such priors and showed that with help from learning methods, the performance was improved compared to that of the APC approach. Furthermore, very interesting observations about how V4 neurons combine curvature components within their RF could be made.

We explored representational sparseness in the learned mV4 neurons. Our results all well-aligned with the reported observations in biological V4 cells for bias toward acute convexities both in terms

of percentage of nonzero weights and the strength of weights from each curvature component to mV4 responses.

The SparseShape hierarchy is a step forward to further our understanding of V4 shape processing. However, many more questions will arise in the wake of our discoveries. For example, the facilitatory and inhibitory influence of shape parts could be further examined in biological V4 cells by adding/removing figure conformations. With object-centered encodings in V4 studying facilitatory/inhibitory contributions by removing shape parts, which alter the representation from a closed shape to that of an open curve, might prove difficult. However, we discussed in Chapter 3 that the RBO network could explain illusory contours. Moreover, Zhang *et al.* [14] explored the effect of removing boundary parts in RBO neurons and found even when boundary fragments were removed, the side-of-figure preference was present in border ownership responses. Perhaps in a similar manner, the standard stimuli with Cornsweet shapes could be employed to pinpoint the influence of parts in V4 responses. Note that such a study will be different from the one in which V4 responses to occlusion were examined [221, 220]. In the suggested experimental design, there is no occlusion, but absence of a parts whereas in the occlusion studies, the shape is occluded with another form affecting border ownership and consequently shape encoding.

Our proposed model can be further improved. For example, in SparseShape, unlike HMAX, we did not perform cross-validation for individual neurons to determine the optimal parameter (λ in SparseShape) for each cell. Other improvements with more massive modelings are also possible. For example, adding neurons selective to texture could lend insight into joint shape and texture processing in this area. Similarly, incorporating horizontal interactions could help in explaining the dynamics of V4 responses. These steps are left for future work.

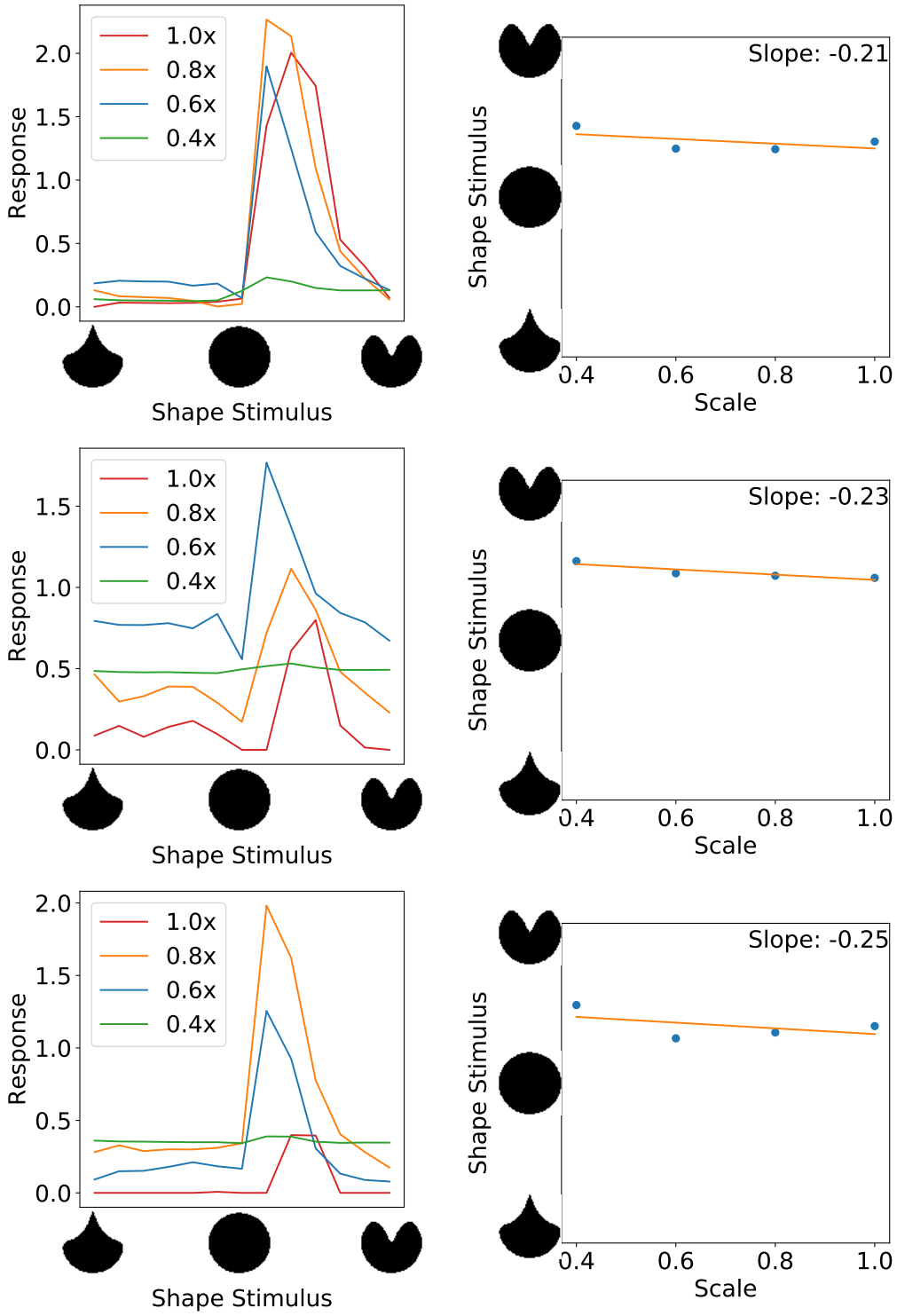
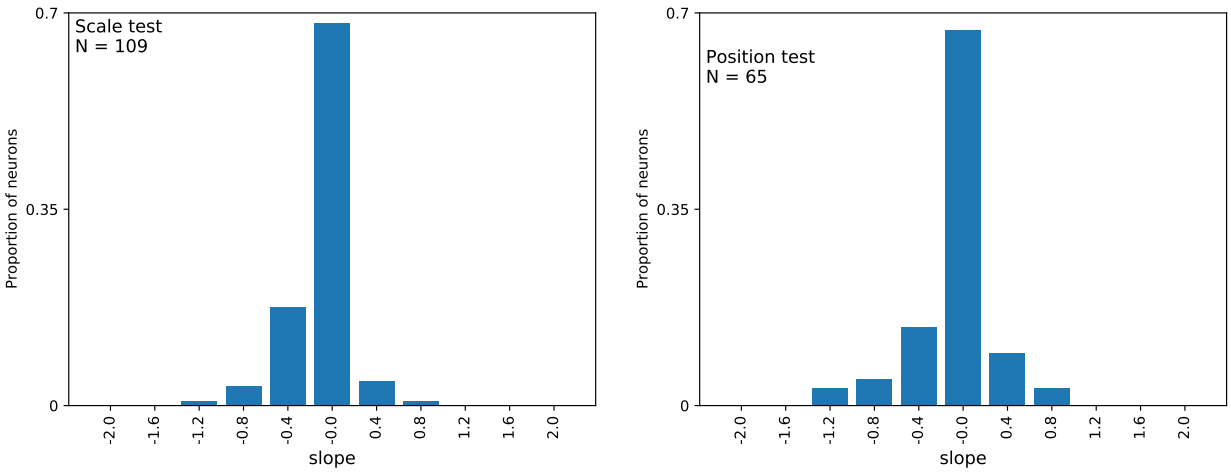


Figure 4.30: Scale-invariance in responses of mV4 neurons. Each row in this figure consists of responses of a mV4 cell to stimuli of various scales (left) and its corresponding tuning centroid slope as a function of scale (right). In light of the two hypotheses for scale-invariance in V4 responses, and compared to their schematic predictions in Figure 4.29, these mV4 cells exhibit "normalized" curvature encoding.



(a) **(b)**

Figure 4.31: Tuning centroid slope distribution for the population of mV4 neurons in Sparse-Shape. (a) Slope distribution for the scale test, (b) Slope distribution for the position test. These histograms match those of biological V4 cells illustrated in Figure 4.31 implying scale-invariance in the learned mV4 responses despite the training being performed on a single scale of shapes.

Chapter 5

Conclusions and Future Directions

5.1 Summary of Contributions

All the models proposed here were constrained to meet known biological aspects. With amazing new discoveries about the brain, and the visual system in particular, our knowledge grows, as well as our set of known unknowns. That being said, the existing models can be further extended or modified to explain new uncovered processing aspects or similarly to make new predictions. The loop between neurophysiological experiments and computational modeling, such as those proposed in this dissertation, is both exciting and necessary. Where testing for certain hypotheses might be costly with physiological experiments, those can be scrutinized with computational modeling. With new physiological data available, previously introduced computational models can be tested and adjustments can be made if necessary. In short, both methods need to go hand in hand for an efficient maneuver to further our understanding of the brain. This dissertation was an attempt in computational modeling with the current available knowledge and unknowns. My hope is that the present manuscript could illustrate the importance of analytically-defined neural networks in the loop.

In this thesis, I pursued two goals:

1. To incorporate modern neuroscience into biologically-inspired models and demonstrate how certain representations can be implemented in a vision system in accordance to recent discoveries about the brain.

2. To demonstrate that marrying biologically-inspired models with learning algorithms can effectively rectify the challenge of modeling still-unknown processing mechanisms.

The hue network and the RBO model covered in Chapters 2 and 3 were attempts toward the first goal. The SparseShape network explained in Chapter 4 suggests that pairing analytical neural networks with learning can reach beyond boundaries of our current findings. With these models explained, the following contributions, repeated from the introduction, are apparent:

- **Direct understanding of model features:** All three models provide clear explanations for the various components in the model and allow for an effortless interpretation of features present in the model. For example, consider answering the question "what features are extracted at the endstop-curvature sign layer (layer 4 in SparseShape)?" The answer to this

question is trivial: neuron responses in this layer signal presence/absence of convex or concave contour segments in an object-centered encoding.

- **Effortless verification of contributing factors:** With analytical models, the factors contributing to a specific representation can be easily tested. For example, in the hue network, we evaluated the contributions from single-opponent and multiplicative mV2 neurons to mV4 cells with a simple approach. As another example, in the RBO network, evaluating the effects of global versus local context did not require any complicated attempts. In SparseShape, pinpointing contributing curvature components to V4 responses was effortless by simply analyzing the learned sparse code vector.

- **Explicit explanation of learned features:** When combined with learning approaches, due to the easy interpretation of features, the learned mechanisms are easy to understand and explain; the visualized receptive fields in SparseShape are a testament to ease in explaining the learned features.

In addition to the overall importance of models studied in my thesis, each hierarchy delivers its own unique advantages in addressing specific visual processing problems. For example, in the case of the hue network, by proposing a mechanistic computational model inspired by neural mechanisms in the visual system:

- The contributions of each visual area LGN, V1, V2, and V4 in hue representation were revealed.

- A plausible process for the progression of hue representation in the brain was provided.

- We suggested the answer to the question "which region in the brain represents unique hues?" is not limited to a single brain area, which in turn could be the source of disagreement among color vision researchers.

The RBO network:

- Is the first to combine global and local context for border ownership assignment.

- Lends detailed insight to the role each of the global and local context play in a quantified way.

- Eliminates the need for any built-in priors for BO assignment.
- Explains the perception of illusory contours such as the Kanizsa's square.

Finally, the SparseShape network:

- Is the first computational model to suggest mechanisms for an object-centered representation in model layer V4.
- Demonstrates how simple learning methodologies can be employed in an analytical neural network allowing recovering of unknowns of visual processing in the brain.
- Suggests an approach that eliminates the need for any strong priors to recover neuron tunings in V4.

5.2 Future Directions

This dissertations aimed at answering some questions. But along the way, many more were raised to be investigated in the future. Possible research directions for each model are reviewed separately below:

5.2.1 The Hue Hierarchy

The hue network can be further extended to encode lightness and saturation. Modeling double-opponent neurons encoding color border makes it possible to combine this hierarchy with the RBO network to represent border ownership along color boundaries. In our proposed model, mV4 responses were obtained by linearly combining single-opponent and multiplicative mV2 responses. An important question, for which I am not aware of any physiological study or evidence, is if hue-selective neurons in V4 combine V2 responses in a nonlinear fashion, perhaps with mutliplicative modulations. A computational model with predictions could help an informed experimental design for further physiological study.

5.2.2 The RBO Model

In the RBO network, we provided a detail discussion on plausibility of dorsal modulations for border ownership assignment. This hypothesis need to be tested. For now, we assume dorsal modulations

is one mechanism employed in the brain to encode border ownership. However, the brain is a complex neural network with complicated and entangle neuronal connections within and between visual areas. As such, this complex organ has a variety of mechanisms at its disposal and a subset of those could be employed for encoding certain features, not just a single mechanism. In that case, one need to investigate incorporating more than dorsal and lateral modulations into border ownership assignment. For example, feedback from higher ventral areas could enhance the difference of responses later in the time course. Additionally, in absence of dorsal modulations, for example for color borders where colors at abutting regions of the border give rise to the same intensity, feedback from higher ventral areas could be key to achieving border ownership assignment. One can investigate these possibilities with an extended version of the RBO network, perhaps combined with an extended version of the hue hierarchy, prior to any physiological study.

5.2.3 The SparseShape Network

The SparseShape network provided the first object-centered representation for modeling V4 responses. More experiments can be performed on SparseShape to evaluate its predictive power in modeling V4 responses. For example, testing the learned mV4 responses to occlusion. Although we studied position-invariance in mV4 responses as part of our scale-invariance experiment, another study with focus on mV4 responses to translated shapes within its RF could provide better insight. SparseShape can be further extended to model both filled/outlined shapes in mV2 and mV4 to model V4 responses to similar stimuli. Adding texture-selective cells allow for modeling V4 responses to joint texture and shape stimuli.

We demonstrated how learning can be incorporated into an analytically-defined neural network to compensate for lack of knowledge. In a similar fashion, a learning step could help to account for variability in the data, for example, those present in natural images. More specifically, the SparseShape network can be extended with additional neuronal layers to learn natural object shapes.

Finally, all the three models, and their extended versions, can be combined to make a bigger

analytically-defined neural network that can better account for variations in the visual signal just the same way our brain has combined them all in a single intriguing machine.

Bibliography

[1] J. H. Maunsell, "Physiological evidence for two visual subsystems," in *Matters of intelligence*, pp. 59–87, Springer, 1987.

[2] M. T. Schmolesky, Y. Wang, D. P. Hanes, K. G. Thompson, S. Leutgeb, J. D. Schall, and A. G. Leventhal, "Signal timing across the macaque visual system," *Journal of neurophysiology*, vol. 79, no. 6, pp. 3272–3278, 1998.

[3] Q. Zhang, Y. N. Wu, and S.-C. Zhu, "Interpretable convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8827–8836, 2018.

[4] L. G. Ungerleider and J. V. Haxby, "'what' and 'where' in the human brain," *Current opinion in neurobiology*, vol. 4, no. 2, pp. 157–165, 1994.

[5] B. A. Wandell and J. Winawer, "Computational neuroimaging and population receptive fields," *Trends in cognitive sciences*, vol. 19, no. 6, pp. 349–357, 2015.

[6] M. Li, F. Liu, M. Juusola, and S. Tang, "Perceptual color map in macaque visual area v4," *Journal of Neuroscience*, vol. 34, no. 1, pp. 202–217, 2014.

[7] D. C. Kiper, S. B. Fenstemaker, and K. R. Gegenfurtner, "Chromatic properties of neurons in macaque area V2," *Visual neuroscience*, vol. 14, no. 6, pp. 1061–1072, 1997.

[8] E. Miyahara, "Focal colors and unique hues," *Perceptual and motor skills*, vol. 97, no. 3_suppl, pp. 1038–1042, 2003.

[9] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729, 2008.

[10] J. Zhang, Y. Barhomi, and T. Serre, "A new biologically inspired color image descriptor," in *Proceedings of the European Conference on Computer Vision*, pp. 312–324, Springer, 2012.

[11] H. Zhou, H. S. Friedman, and R. von der Heydt, "Coding of Border Ownership in Monkey Visual Cortex," *Journal of Neuroscience*, vol. 20, no. 17, pp. 6594–6611, 2000.

[12] C. C. Fowlkes, D. R. Martin, and J. Malik, "Local figure–ground cues are valid for natural images," *Journal of Vision*, vol. 7, no. 8, pp. 2–2, 2007.

[13] X. Ren, C. C. Fowlkes, and J. Malik, "Figure/ground assignment in natural images," in *Proceedings of the European Conference on Computer Vision*, pp. 614–627, Springer, 2006.

[14] N. R. Zhang and R. von der Heydt, "Analysis of the Context Integration Mechanisms Underlying Figure–Ground Organization in the Visual Cortex," *Journal of Neuroscience*, vol. 30, no. 19, pp. 6482–6496, 2010.

[15] J. Bullier, "Integrated model of visual processing," *Brain Research Reviews*, vol. 36, no. 2, pp. 96–107, 2001.

[16] A. J. Rodríguez-Sánchez and J. K. Tsotsos, "The Roles of Endstopped and Curvature Tuned Computations in a Hierarchical Representation of 2D Shape," *PLoS ONE*, vol. 7, pp. 1–13, 08 2012.

[17] D. J. Felleman and J. H. Kaas, "Receptive-field properties of neurons in middle temporal visual area (MT) of owl monkeys," *Journal of neurophysiology*, vol. 52, no. 3, pp. 488–513, 1984.

[18] E. Craft, H. Schütze, E. Niebur, and R. von der Heydt, "A Neural Model of Figure–Ground Organization," *Journal of neurophysiology*, vol. 97, no. 6, pp. 4310–4326, 2007.

[19] A. Pasupathy and C. E. Connor, "Responses to contour features in macaque area V4," *Journal of neurophysiology*, vol. 82, no. 5, pp. 2490–2502, 1999.

[20] A. Pasupathy and C. E. Connor, "Shape representation in area V4: position-specific tuning for boundary conformation," *Journal of neurophysiology*, vol. 86, no. 5, pp. 2505–2519, 2001.

[21] A. Pasupathy and C. E. Connor, "Population coding of shape in area V4," *Nature neuroscience*, vol. 5, no. 12, pp. 1332–1338, 2002.

[22] A. N. Pressley, *Elementary differential geometry*. Springer Science & Business Media, 2010.

[23] A. Pasupathy, D. V. Popovkina, and T. Kim, "Visual functions of primate area V4," *Annual Review of Vision Science*, vol. 6, pp. 363–385, 2020.

[24] 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*, vol. 98, no. 3, pp. 1733–1750, 2007.

[25] T. Serre, M. Kouh, C. Cadieu, U. Knoblich, G. Kreiman, and T. Poggio, "A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex," tech. rep., MASSACHUSETTS INST OF TECH CAMBRIDGE MA CENTER FOR BIOLOGICAL AND COMPUTATIONAL LEARNING, 2005.

[26] C. Cadieu, M. Kouh, M. Riesenhuber, and T. Poggio, "Shape Representation in : Investigating Position-Specific Tuning for Boundary Confirmation with the Standard Model of Object Recognition," tech. rep., MASSACHUSETTS INST OF TECH CAMBRIDGE MA CENTER FOR BIOLOGICAL AND COMPUTATIONAL LEARNING, 2004.

[27] A. J. Rodríguez-Sánchez and J. K. Tsotsos, "The importance of intermediate representations for the modeling of 2D shape detection: Endstopping and curvature tuned computations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4321–4326, 2011.

[28] E. T. Carlson, R. J. Rasquinha, K. Zhang, and C. E. Connor, "A sparse object coding scheme in area V4," *Current Biology*, vol. 21, no. 4, pp. 288–293, 2011.

[29] Y. El-Shamayleh and A. Pasupathy, "Contour curvature as an invariant code for objects in visual area V4," *Journal of Neuroscience*, vol. 36, no. 20, pp. 5532–5543, 2016.

[30] G. Shmueli *et al.*, "To explain or to predict?," *Statistical science*, vol. 25, no. 3, pp. 289–310, 2010.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the International Conference on Computer Vision*, pp. 1026–1034, 2015.

[32] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[33] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 427–436, 2015.

[34] B. R. Webster, S. E. Anthony, and W. J. Scheirer, "Psyphy: A psychophysics driven evaluation framework for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2280–2286, 2018.

[35] S. Dodge and L. Karam, "A study and comparison of human and deep learning recognition performance under visual distortions," in *2017 26th International Conference on Computer Communication and Networks (ICCCN)*, pp. 1–7, 2017.

[36] H. Hosseini, B. Xiao, M. Jaiswal, and R. Poovendran, "On the limitation of convolutional neural networks in recognizing negative images," in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 352–358, 2017.

[37] M. Ghodrati, A. Farzmahdi, K. Rajaei, R. Ebrahimpour, and S.-M. Khaligh-Razavi, "Feedforward object-vision models only tolerate small image variations compared to human," *Frontiers in computational neuroscience*, vol. 8, p. 74, 2014.

[38] R. Rajalingham, E. B. Issa, P. Bashivan, K. Kar, K. Schmidt, and J. J. DiCarlo, "Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks," *Journal of Neuroscience*, vol. 38, no. 33, pp. 7255–7269, 2018.

[39] Y. Xu and M. Vaziri-Pashkam, "Examining the coding strength of object identity and non-identity features in human occipito-temporal cortex and convolutional neural networks," *Journal of Neuroscience*, 2021.

[40] Y. Xu and M. Vaziri-Pashkam, "Limits to visual representational correspondence between convolutional neural networks and the human brain," *Nature Communications*, vol. 12, no. 1, pp. 1–16, 2021.

[41] R. Geirhos, D. H. Janssen, H. H. Schütt, J. Rauber, M. Bethge, and F. A. Wichmann, "Comparing deep neural networks against humans: object recognition when the signal gets weaker," *arXiv preprint arXiv:1706.06969*, 2017.

[42] B. Hu, S. Khan, E. Niebur, and B. Tripp, "Figure-ground representation in deep neural networks," in *Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–6, 2019.

[43] J. Kim, M. Ricci, and T. Serre, "Not-So-CLEVR: learning same–different relations strains feedforward neural networks," *Interface focus*, vol. 8, no. 4, pp. 00–11, 2018.

[44] T. Horikawa, S. C. Aoki, M. Tsukamoto, and Y. Kamitani, "Characterization of deep neural network features by decodability from human brain activity," *Scientific data*, vol. 6, no. 1, pp. 1–12, 2019.

[45] C. Wloka and J. K. Tsotsos, "Flipped on its Head: Deep Learning-Based Saliency Finds Asymmetry in the Opposite Direction Expected for Singleton Search of Flipped and Canonical Targets," *Journal of Vision*, vol. 19, no. 10, pp. 318–318, 2019.

[46] C. Wloka, I. Kotseruba, and J. K. Tsotsos, "Active fixation control to predict saccade sequences," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3184–3193, 2018.

[47] A. Rosenfeld, R. Zemel, and J. K. Tsotsos, "The elephant in the room," *arXiv preprint arXiv:1808.03305*, 2018.

[48] C. M. Altimus, B. J. Marlin, N. E. Charalambakis, A. Colón-Rodríguez, E. J. Glover, P. Izbicki, A. Johnson, M. V. Lourenco, R. A. Makinson, J. McQuail, I. Obeso, N. Padilla-Coreano, and M. F. Wells, "The Next 50 Years of Neuroscience," *Journal of Neuroscience*, vol. 40, no. 1, pp. 101–106, 2020.

[49] J. W. Brown, "The tale of the neuroscientists and the computer: why mechanistic theory matters," *Frontiers in neuroscience*, vol. 8, p. 349, 2014.

[50] D. H. Hubel and T. N. Wiesel, "Receptive fields of single neurones in the cat's striate cortex," *The Journal of physiology*, vol. 148, no. 3, pp. 574–591, 1959.

[51] C. F. Cadieu, H. Hong, D. L. Yamins, N. Pinto, D. Ardila, E. A. Solomon, N. J. Majaj, and J. J. DiCarlo, "Deep neural networks rival the representation of primate IT cortex for core visual object recognition," *PLoS computational biology*, vol. 10, no. 12, pp. 1–18, 2014.

[52] S. R. Kheradpisheh, M. Ghodrati, M. Ganjtabesh, and T. Masquelier, "Deep networks can resemble human feed-forward vision in invariant object recognition," *Scientific reports*, vol. 6, no. 1, pp. 1–24, 2016.

[53] S.-M. Khaligh-Razavi and N. Kriegeskorte, "Deep supervised, but not unsupervised, models may explain IT cortical representation," *PLoS computational biology*, vol. 10, no. 11, pp. 1–29, 2014.

[54] R. M. Cichy, A. Khosla, D. Pantazis, A. Torralba, and A. Oliva, "Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence," *Scientific reports*, vol. 6, no. 1, pp. 1–13, 2016.

[55] J. Kubilius, S. Bracci, and H. P. Op de Beeck, "Deep neural networks as a computational model for human shape sensitivity," *PLoS computational biology*, vol. 12, no. 4, pp. 1–26, 2016.

[56] D. A. Pospisil, A. Pasupathy, and W. Bair, "'Artiphysiology' reveals V4-like shape tuning in a deep network trained for image classification," *eLife*, vol. 7, p. e38242, 2018.

[57] I. Kuzovkin, R. Vicente, M. Petton, J.-P. Lachaux, M. Baciu, P. Kahane, S. Rheims, J. R. Vidal, and J. Aru, "Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex," *Communications biology*, vol. 1, no. 1, pp. 1–12, 2018.

[58] C. Zhuang, S. Yan, A. Nayebi, M. Schrimpf, M. C. Frank, J. J. DiCarlo, and D. L. K. Yamins, "Unsupervised neural network models of the ventral visual stream," *Proceedings of the National Academy of Sciences, U.S.A.*, vol. 118, no. 3, 2021.

[59] M. Eickenberg, A. Gramfort, G. Varoquaux, and B. Thirion, "Seeing it all: Convolutional network layers map the function of the human visual system," *NeuroImage*, vol. 152, pp. 184–194, 2017.

[60] M. Schrimpf, J. Kubilius, M. J. Lee, N. A. R. Murty, R. Ajemian, and J. J. DiCarlo, "Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence," *Neuron*, vol. 108, no. 3, pp. 413–423, 2020.

[61] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

[62] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[63] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proceedings of the International Conference on Machine Learning*, pp. 10347–10357, 2021.

[64] N. Baker, H. Lu, G. Erlikhman, and P. J. Kellman, "Deep convolutional networks do not classify based on global object shape," *PLoS computational biology*, vol. 14, no. 12, pp. 1–43, 2018.

[65] N. Baker, H. Lu, G. Erlikhman, and P. J. Kellman, "Local features and global shape information in object classification by deep convolutional neural networks," *Vision research*, vol. 172, pp. 46–61, 2020.

[66] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness.," in *Proceedings of International Conference on Learning and Representation*, 2019.

[67] R. Geirhos, K. Narayanappa, B. Mitzkus, T. Thieringer, M. Bethge, F. A. Wichmann, and W. Brendel, "Partial success in closing the gap between human and machine vision," *arXiv:2106.07411*, 2021.

[68] K. Kar and J. J. DiCarlo, "Fast recurrent processing via ventrolateral prefrontal cortex is needed by the primate ventral stream for robust core visual object recognition," *Neuron*, vol. 109, no. 1, pp. 164–176, 2021.

[69] K. N. Kay, "Principles for models of neural information processing," *NeuroImage*, vol. 180, pp. 101–109, 2018.

[70] A. Stevenson, *Oxford Dictionary of English*. Oxford University Press, 2010.

[71] T. J. Sejnowski, "The unreasonable effectiveness of deep learning in artificial intelligence," *Proceedings of the National Academy of Sciences, U.S.A.*, vol. 117, no. 48, pp. 30033–30038, 2020.

[72] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo, "Modeling visual attention via selective tuning," *Artificial Intelligence*, vol. 78, no. 1-2, pp. 507–545, 1995.

[73] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proceedings of the European Conference on Computer Vision*, pp. 818–833, 2014.

[74] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *In Workshop at International Conference on Learning Representations*, 2014.

[75] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5188–5196, 2015.

[76] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10, no. 7, pp. 1–46, 2015.

[77] T. Zhou, P. Krahenbuhl, M. Aubry, Q. Huang, and A. A. Efros, "Learning dense correspondence via 3D-guided cycle consistency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 117–126, 2016.

[78] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *Proceedings of the International Conference on Computer Vision*, pp. 3429–3437, 2017.

[79] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the international conference on neural information processing systems*, pp. 4768–4777, 2017.

[80] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the International Conference on Computer Vision*, pp. 618–626, 2017.

[81] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the International Conference on Machine Learning*, pp. 3319–3328, 2017.

[82] M. Sundararajan, J. Xu, A. Taly, R. Sayres, and A. Najmi, "Exploring principled visualizations for deep network attributions.," in *IUI Workshops*, vol. 4, 2019.

[83] A. Kapishnikov, T. Bolukbasi, F. Viégas, and M. Terry, "XRAI: Better Attributions Through Regions," in *Proceedings of the International Conference on Computer Vision*, pp. 4948–4957, 2019.

[84] V. Petsiuk, A. Das, and K. Saenko, "RISE: Randomized Input Sampling for Explanation of Black-box Models," in *Proceedings of the British Machine Vision Conference*, 2018.

[85] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.

[86] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, *et al.*, "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)," in *Proceedings of the International Conference on Machine Learning*, pp. 2668–2677, 2018.

[87] C. Feichtenhofer, A. Pinz, R. P. Wildes, and A. Zisserman, "Deep Insights into Convolutional Networks for Video Recognition," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 420–437, 2020.

[88] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene cnns," *arXiv preprint arXiv:1412.6856*, 2014.

[89] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network Dissection: Quantifying Interpretability of Deep Visual Representations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6541–6549, 2017.

[90] B. Zhou, D. Bau, A. Oliva, and A. Torralba, "Interpreting Deep Visual Representations via Network Dissection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2131–2145, 2019.

[91] D. Bau, J.-Y. Zhu, H. Strobelt, A. Lapedriza, B. Zhou, and A. Torralba, "Understanding the role of individual units in a deep neural network," *Proceedings of the National Academy of Sciences, U.S.A.*, vol. 117, no. 48, pp. 30071–30078, 2020.

[92] D. J. Felleman and D. C. Van Essen, "Distributed hierarchical processing in the primate cerebral cortex.," *Cerebral Cortex*, vol. 1, no. 1, pp. 1–47, 1991.

[93] T. D. Albright, "Direction and orientation selectivity of neurons in visual area MT of the macaque," *Journal of neurophysiology*, vol. 52, no. 6, pp. 1106–1130, 1984.

[94] S. Raiguel, D.-K. Xiao, V. Marcar, and G. Orban, "Response latency of macaque area MT/V5 neurons and its relationship to stimulus parameters," *Journal of neurophysiology*, vol. 82, no. 4, pp. 1944–1956, 1999.

[95] H. Kolster, R. Peeters, and G. A. Orban, "The retinotopic organization of the human middle temporal area MT/V5 and its cortical neighbors," *Journal of Neuroscience*, vol. 30, no. 29, pp. 9801–9820, 2010.

[96] G. M. Boynton and J. Hegdé, "Visual cortex: The continuing puzzle of area V2," *Current Biology*, vol. 14, no. 13, pp. R523–R524, 2004.

[97] K. Tanaka, "Inferotemporal cortex and object vision," *Annual review of neuroscience*, vol. 19, no. 1, pp. 109–139, 1996.

[98] E. Kobatake, G. Wang, and K. Tanaka, "Effects of shape-discrimination training on the selectivity of inferotemporal cells in adult monkeys," *Journal of neurophysiology*, vol. 80, no. 1, pp. 324–330, 1998.

[99] J. Freeman and E. Simoncelli, "Metamers of the ventral stream," *Nature neuroscience*, vol. 14, pp. 1195–1204, 9 2011.

[100] J. Hupé, A. James, B. Payne, S. Lomber, P. Girard, and J. Bullier, "Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons," *Nature*, vol. 394, no. 6695, pp. 784–787, 1998.

[101] J.-M. Hupe, A. C. James, P. Girard, S. G. Lomber, B. R. Payne, and J. Bullier, "Feedback connections act on the early part of the responses in monkey visual cortex," *Journal of neurophysiology*, vol. 85, no. 1, pp. 134–145, 2001.

[102] A. Angelucci and J. Bullier, "Reaching beyond the classical receptive field of V1 neurons: horizontal or feedback axons?," *Journal of Physiology-Paris*, vol. 97, no. 2, pp. 141–154, 2003.

[103] R. C. Reid and R. M. Shapley, "Space and time maps of cone photoreceptor signals in macaque lateral geniculate nucleus," *Journal of Neuroscience*, vol. 22, no. 14, pp. 6158–6175, 2002.

[104] L. E. Wool, J. D. Crook, J. B. Troy, O. S. Packer, Q. Zaidi, and D. M. Dacey, "Nonselective wiring accounts for red-green opponency in midget ganglion cells of the primate retina," *Journal of Neuroscience*, vol. 38, no. 6, pp. 1520–1540, 2018.

[105] L. E. Wool, O. S. Packer, Q. Zaidi, and D. M. Dacey, "Connectomic identification and three-dimensional color tuning of S-OFF midget ganglion cells in the primate retina," *Journal of Neuroscience*, vol. 39, no. 40, pp. 7893–7909, 2019.

[106] E. Hering, *Outlines of a theory of the light sense.* Harvard University Press, 1964.

[107] A. M. Derrington, J. Krauskopf, and P. Lennie, "Chromatic mechanisms in lateral geniculate nucleus of macaque.," *The Journal of physiology*, vol. 357, no. 1, pp. 241–265, 1984.

[108] R. L. De Valois, N. P. Cottaris, S. D. Elfar, L. E. Mahon, and J. A. Wilson, "Some transformations of color information from lateral geniculate nucleus to striate cortex," *Proceedings of the National Academy of Sciences, U.S.A.*, vol. 97, no. 9, pp. 4997–5002, 2000.

[109] H. Sun, H. E. Smithson, Q. Zaidi, and B. B. Lee, "Specificity of cone inputs to macaque retinal ganglion cells," *Journal of neurophysiology*, vol. 95, no. 2, pp. 837–849, 2006.

[110] D. I. MacLeod and R. M. Boynton, "Chromaticity diagram showing cone excitation by stimuli of equal luminance," *Journal of the Optical Society of America*, vol. 69, no. 8, pp. 1183–1186, 1979.

[111] A. Hanazawa, H. Komatsu, and I. Murakami, "Neural selectivity for hue and saturation of colour in the primary visual cortex of the monkey," *European Journal of Neuroscience*, vol. 12, no. 5, pp. 1753–1763, 2000.

[112] I. Kuriki, P. Sun, K. Ueno, K. Tanaka, and K. Cheng, "Hue selectivity in human visual cortex revealed by functional magnetic resonance imaging," *Cerebral Cortex*, vol. 25, no. 12, pp. 4869–4884, 2015.

[113] E. N. Johnson, M. J. Hawken, and R. Shapley, "Cone inputs in macaque primary visual cortex," *Journal of neurophysiology*, vol. 91, no. 6, pp. 2501–2514, 2004.

[114] R. Shapley and M. Hawken, "Neural mechanisms for color perception in the primary visual cortex," *Current opinion in neurobiology*, vol. 12, no. 4, pp. 426–432, 2002.

[115] P. Lennie, J. Krauskopf, and G. Sclar, "Chromatic mechanisms in striate cortex of macaque," *Journal of Neuroscience*, vol. 10, no. 2, pp. 649–669, 1990.

[116] T. Wachtler, T. J. Sejnowski, and T. D. Albright, "Representation of color stimuli in awake macaque primary visual cortex," *Neuron*, vol. 37, no. 4, pp. 681–691, 2003.

[117] T. Namima, M. Yasuda, T. Banno, G. Okazawa, and H. Komatsu, "Effects of luminance contrast on the color selectivity of neurons in the macaque area V4 and inferior temporal cortex," *Journal of Neuroscience*, vol. 34, no. 45, pp. 14934–14947, 2014.

[118] B. R. Conway, S. Moeller, and D. Y. Tsao, "Specialized color modules in macaque extrastriate cortex," *Neuron*, vol. 56, no. 3, pp. 560–573, 2007.

[119] B. R. Conway and D. Y. Tsao, "Color-tuned neurons are spatially clustered according to color preference within alert macaque posterior inferior temporal cortex," *Proceedings of the National Academy of Sciences, U.S.A.*, vol. 106, no. 42, pp. 18034–18039, 2009.

[120] H. B. Barlow, "Why have multiple cortical areas?," *Vision research*, vol. 26, no. 1, pp. 81–90, 1986.

[121] Y. Xiao, Y. Wang, and D. J. Felleman, "A spatially organized representation of colour in macaque cortical area V2," *Nature*, vol. 421, no. 6922, pp. 535–539, 2003.

[122] K. S. Bohon, K. L. Hermann, T. Hansen, and B. R. Conway, "Representation of perceptual color space in macaque posterior inferior temporal cortex (the V4 complex)," *eNeuro*, vol. 3, no. 4, pp. ENEURO–0039, 2016.

[123] S. J. Schein and R. Desimone, "Spectral properties of V4 neurons in the macaque," *Journal of Neuroscience*, vol. 10, no. 10, pp. 3369–3389, 1990.

[124] S. Zeki, "The representation of colours in the cerebral cortex," *Nature*, vol. 284, no. 5755, pp. 412–418, 1980.

[125] Q. Zaidi, J. Marshall, H. Thoen, and B. R. Conway, "Evolution of neural computations: Mantis shrimp and human color decoding," *i-Perception*, vol. 5, no. 6, pp. 492–496, 2014.

[126] M. A. Webster, E. Miyahara, G. Malkoc, and V. E. Raker, "Variations in normal color vision. II. unique hues," *Journal of the Optical Society of America A*, vol. 17, no. 9, pp. 1545–1555, 2000.

[127] S. M. Wuerger, P. Atkinson, and S. Cropper, "The cone inputs to the unique-hue mechanisms," *Vision research*, vol. 45, no. 25, pp. 3210–3223, 2005.

[128] C. M. Stoughton and B. R. Conway, "Neural basis for unique hues," *Current Biology*, vol. 18, no. 16, pp. R698–R699, 2008.

[129] H. Komatsu, Y. Ideura, S. Kaji, and S. Yamane, "Color selectivity of neurons in the inferior temporal cortex of the awake macaque monkey," *Journal of Neuroscience*, vol. 12, no. 2, pp. 408–424, 1992.

[130] J. Mollon, "A neural basis for unique hues?," *Current Biology*, vol. 19, no. 11, pp. R441–R442, 2009.

[131] Q. Zaidi, R. Ennis, L. Wool, S. Komban, J.-M. Alonso, B. Conway, G. Gagin, and K. Bohon, "The enigma of unique hues," in *i-Perception*, vol. 5, pp. 421–421, 2014.

[132] R. J. Ennis and Q. Zaidi, "Geometrical structure of perceptual color space: mental representations and adaptation invariance," *Journal of Vision*, vol. 19, no. 12, pp. 1–1, 2019.

[133] L. E. Wool, S. J. Komban, J. Kremkow, M. Jansen, X. Li, J.-M. Alonso, and Q. Zaidi, "Salience of unique hues and implications for color theory," *Journal of Vision*, vol. 15, no. 2, pp. 10–10, 2015.

[134] A. Rasouli and J. K. Tsotsos, "The effect of color space selection on detectability and discriminability of colored objects," *arXiv preprint arXiv:1702.05421*, 2017.

[135] M. Engilberge, E. Collins, and S. Süsstrunk, "Color representation in deep neural networks," in *Proceedings of the IEEE International Conference on Image Processing*, pp. 2786–2790, 2017.

[136] I. Rafegas and M. Vanrell, "Color encoding in biologically-inspired convolutional neural networks," *Vision research*, vol. 151, pp. 7–17, 2018.

[137] P. A. Dufort and C. J. Lumsden, "Color categorization and color constancy in a neural network model of V4," *Biological cybernetics*, vol. 65, no. 4, pp. 293–303, 1991.

[138] S. M. Courtney, L. H. Finkel, and G. Buchsbaum, "A multistage neural network for color constancy and color induction," *IEEE Trans. Neural Networks*, vol. 6, no. 4, pp. 972–985, 1995.

[139] J. Wray and G. M. Edelman, "A model of color vision based on cortical reentry," *Cerebral Cortex*, vol. 6, no. 5, pp. 701–716, 1996.

[140] R. L. De Valois and K. K. De Valois, "A multi-stage color model," *Vision research*, vol. 33, no. 8, pp. 1053–1065, 1993.

[141] S. R. Lehky and T. J. Sejnowski, "Seeing white: Qualia in the context of decoding population codes," *Neural computation*, vol. 11, no. 6, pp. 1261–1280, 1999.

[142] B. R. Conway, "Spatial structure of cone inputs to color cells in alert macaque primary visual cortex (V-1)," *Journal of Neuroscience*, vol. 21, no. 8, pp. 2768–2783, 2001.

[143] B. R. Conway and M. S. Livingstone, "Spatial and temporal properties of cone signals in alert macaque primary visual cortex," *Journal of Neuroscience*, vol. 26, no. 42, pp. 10826–10846, 2006.

[144] R. A. Silver, "Neuronal arithmetic," *Nature reviews neuroscience*, vol. 11, no. 7, p. 474, 2010.

[145] F. Gabbiani, H. G. Krapp, C. Koch, and G. Laurent, "Multiplicative computation in a visual neuron sensitive to looming," *Nature*, vol. 420, no. 6913, p. 320, 2002.

[146] J. L. Peña and M. Konishi, "Auditory spatial receptive fields created by multiplication," *Science*, vol. 292, no. 5515, pp. 249–252, 2001.

[147] J. B. D. Paula, "Converting RGB images to LMS cone activations," tech. rep., Department of Computer Sciences, The University of Texas at Austin, 2006. Technical Report 06-49.

[148] A. L. Rothenstein, A. Zaharescu, and J. K. Tsotsos, "Tarzann : A general purpose neural network simulator for visual attention modeling," in *Attention and Performance in Computational Vision*, pp. 159–167, 2005.

[149] R. Shapley and M. J. Hawken, "Color in the cortex: single-and double-opponent cells," *Vision research*, vol. 51, no. 7, pp. 701–717, 2011.

[150] Q. Zaidi and B. Conway, "Steps towards neural decoding of colors," *Current Opinion in Behavioral Sciences*, vol. 30, pp. 169–177, 2019.

[151] E. Koch, J. Jin, J. M. Alonso, and Q. Zaidi, "Functional implications of orientation maps in primary visual cortex," *Nature communications*, vol. 7, no. 1, pp. 1–13, 2016.

[152] K. Koh, S.-J. Kim, and S. Boyd, "An interior-point method for large-scale L1-regularized logistic regression," *Journal of Machine Learning Research*, vol. 8, no. Jul, pp. 1519–1555, 2007.

[153] J. R. Williford and R. von der Heydt, "Figure-ground organization in visual cortex for natural scenes," *eNeuro*, vol. 3, no. 6, pp. ENEURO–0127, 2016.

[154] J. K. Hesse and D. Y. Tsao, "Consistency of border-ownership cells across artificial stimuli, natural stimuli, and stimuli with ambiguous contours," *Journal of Neuroscience*, vol. 36, no. 44, pp. 11338–11349, 2016.

[155] F. T. Qiu and R. Von Der Heydt, "Figure and ground in the visual cortex: V2 combines stereoscopic cues with gestalt rules," *Neuron*, vol. 47, no. 1, pp. 155–166, 2005.

[156] F. T. Qiu, T. Sugihara, and R. Von Der Heydt, "Figure-ground mechanisms provide structure for selective attention," *Nature neuroscience*, vol. 10, no. 11, p. 1492, 2007.

[157] T. N. Cornsweet, *Visual perception*. Academic Press, 1970.

[158] K. Sakai and H. Nishimura, "Surrounding suppression and facilitation in the determination of border ownership," *Journal of Cognitive Neuroscience*, vol. 18, no. 4, pp. 562–579, 2006.

[159] L. Zhaoping, "Border ownership from intracortical interactions in visual area V2," *Neuron*, vol. 47, no. 1, pp. 143 – 153, 2005.

[160] T. Sugihara, F. T. Qiu, and R. von der Heydt, "The speed of context integration in the visual cortex," *Journal of neurophysiology*, vol. 106, no. 1, pp. 374–385, 2011.

[161] S. L. Brincat and C. E. Connor, "Dynamic shape synthesis in posterior inferotemporal cortex," *Neuron*, vol. 49, no. 1, pp. 17 – 24, 2006.

[162] H. Fu, C. Wang, D. Tao, and M. J. Black, "Occlusion boundary detection via deep exploration of context," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 241–250, 2016.

[163] P. Sundberg, T. Brox, M. Maire, P. Arbeláez, and J. Malik, "Occlusion boundary detection and figure/ground assignment from optical flow," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2233–2240, 2011.

[164] M. Nitzberg and D. Mumford, "The 2.1-D sketch," in *Proceedings of the International Conference on Computer Vision*, pp. 138–144, Dec 1990.

[165] L. G. Roberts, *Machine perception of three-dimensional solids.* PhD thesis, Massachusetts Institute of Technology, 1963.

[166] D. Hoiem, A. A. Efros, and M. Hebert, "Recovering Occlusion Boundaries from an Image," *International Journal of Computer Vision*, vol. 91, no. 3, pp. 328–346, 2011.

[167] P. Wang and A. Yuille, "DOC: Deep OCclusion Estimation from a Single Image," in *Proceedings of the European Conference on Computer Vision*, pp. 545–561, 2016.

[168] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics," in *Proceedings of the International Conference on Computer Vision*, vol. 2, pp. 416–423, 2001.

[169] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proceedings of the International Conference on Computer Vision*, pp. 2961–2969, 2017.

[170] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[171] J. M. Yau, A. Pasupathy, S. L. Brincat, and C. E. Connor, "Curvature Processing Dynamics in Macaque Area V4," *Cerebral Cortex*, vol. 23, no. 1, pp. 198 – 209, 2012.

[172] K. Sakai, H. Nishimura, R. Shimizu, and K. Kondo, "Consistent and robust determination of border ownership based on asymmetric surrounding contrast," *Neural Networks*, vol. 33, pp. 257–274, 2012.

[173] H. Supèr, A. Romeo, and M. Keil, "Feed-forward segmentation of figure-ground and assignment of border-ownership," *PLoS ONE*, vol. 5, pp. 1–14, 05 2010.

[174] J. F. Jehee, V. A. Lamme, and P. R. Roelfsema, "Boundary assignment in a recurrent network architecture," *Vision research*, vol. 47, no. 9, pp. 1153 – 1165, 2007.

[175] O. W. Layton, E. Mingolla, and A. Yazdanbakhsh, "Neural dynamics of feedforward and feedback processing in figure-ground segregation," *Frontiers in Psychology*, vol. 5, p. 972, 2014.

[176] S. Tschechne and H. Neumann, "Hierarchical representation of shapes in visual cortex—from localized features to figural shape segregation," *Frontiers in computational neuroscience*, vol. 8, p. 93, 2014.

[177] A. F. Russell, S. Mihalaş, R. von der Heydt, E. Niebur, and R. Etienne-Cummings, "A model of proto-object based saliency," *Vision research*, vol. 94, pp. 1 – 15, 2014.

[178] O. W. Layton, E. Mingolla, and A. Yazdanbakhsh, "Dynamic coding of border-ownership in visual cortex," *Journal of Vision*, vol. 12, no. 13, p. 8, 2012.

[179] W. Bair, J. R. Cavanaugh, M. A. Smith, and J. A. Movshon, "The timing of response onset and offset in macaque visual neurons," *Journal of Neuroscience*, vol. 22, no. 8, pp. 3189–3205, 2002.

[180] M. J.-M. Macé, A. Delorme, G. Richard, and M. Fabre-Thorpe, "Spotting animals in natural scenes: efficiency of humans and monkeys at very low contrasts," *Animal cognition*, vol. 13, no. 3, pp. 405–418, 2010.

[181] M. J.-M. Macé, S. J. Thorpe, and M. Fabre-Thorpe, "Rapid categorization of achromatic natural scenes: how robust at very low contrasts?," *European Journal of Neuroscience*, vol. 21, no. 7, pp. 2007–2018, 2005.

[182] S. L. Brincat and C. E. Connor, "Underlying principles of visual shape selectivity in posterior inferotemporal cortex," *Nature neuroscience*, vol. 7, no. 8, pp. 880–886, 2004.

[183] A. Angelucci, M. Bijanzadeh, L. Nurminen, F. Federer, S. Merlin, and P. C. Bressloff, "Circuits and mechanisms for surround modulation in visual cortex," *Annual review of neuroscience*, vol. 40, pp. 425–451, 2017.

[184] J. D. Schall, "Visuomotor functions in the frontal lobe," *Annual Review of Vision Science*, vol. 1, pp. 469–498, 2015.

[185] T. Kim, W. Bair, and A. Pasupathy, "Neural coding for shape and texture in macaque area v4," *Journal of Neuroscience*, vol. 39, no. 24, pp. 4760–4774, 2019.

[186] Y. LeCun, P. Haffner, L. Bottou, and Y. Bengio, "Object Recognition with Gradient-Based Learning," in *Shape, contour and grouping in computer vision*, pp. 319–345, 1999.

[187] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of physiology*, vol. 160, no. 1, pp. 106–154, 1962.

[188] D. H. Hubel and T. N. Wiesel, "Receptive fields and functional architecture of monkey striate cortex," *The Journal of physiology*, vol. 195, no. 1, pp. 215–243, 1968.

[189] H. Spitzer and S. Hochstein, "A complex-cell receptive-field model," *Journal of neurophysiology*, vol. 53, no. 5, pp. 1266–1286, 1985.

[190] A. Angelucci, J. B. Levitt, and J. S. Lund, "Anatomical origins of the classical receptive field and modulatory surround field of single neurons in macaque visual cortical area V1," in *Progress in brain research*, vol. 136, pp. 373–388, 2002.

[191] A. P. Witkin, "Scale-space Filtering," in *Proceedings of the International Joint Conference on Artificial Intelligence*, vol. 2, pp. 1019–1022, 1983.

[192] R. Gattass, A. P. Sousa, and M. G. Rosa, "Visual topography of V1 in the cebus monkey," *Journal of Comparative Neurology*, vol. 259, no. 4, pp. 529–548, 1987.

[193] M. Fiorani, R. Gattass, M. G. Rosa, and A. P. Sousa, "Visual area MT in the cebus monkey: location, visuotopic organization, and variability," *Journal of Comparative Neurology*, vol. 287, no. 1, pp. 98–118, 1989.

[194] D. H. Hubel, "The visual cortex of the brain," *Scientific American*, vol. 209, no. 5, pp. 54–63, 1963.

[195] R. Shapley, E. Kaplan, and R. Soodak, "Spatial summation and contrast sensitivity of X and Y cells in the lateral geniculate nucleus of the macaque," *Nature*, vol. 292, no. 5823, p. 543, 1981.

[196] E. Kaplan and R. M. Shapley, "The primate retina contains two types of ganglion cells, with high and low contrast sensitivity," *Proceedings of the National Academy of Sciences, U.S.A.*, vol. 83, no. 8, pp. 2755–2757, 1986.

[197] G. Sclar, J. H. Maunsell, and P. Lennie, "Coding of image contrast in central visual pathways of the macaque monkey," *Vision research*, vol. 30, no. 1, pp. 1–10, 1990.

[198] R. B. Tootell, J. B. Reppas, K. K. Kwong, R. Malach, R. T. Born, T. J. Brady, B. R. Rosen, and J. W. Belliveau, "Functional analysis of human MT and related visual cortical areas using magnetic resonance imaging," *Journal of Neuroscience*, vol. 15, no. 4, pp. 3215–3230, 1995.

[199] X. Shi, B. Wang, and J. K. Tsotsos, "Early Recurrence Improves Edge Detection.," in *Proceedings of the British Machine Vision Conference*, 2013.

[200] 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*, vol. 102, no. 4, pp. 2069–2083, 2009.

[201] S. W. Zucker, "Vertical and horizontal processes in low level vision," *Computer vision systems*, pp. 187–195, 1978.

[202] S. W. Zucker, "Relaxation labeling: 25 years and still iterating," in *Foundations of Image Understanding*, pp. 289–321, Springer, 2001.

[203] X. Shi, N. D. Bruce, and J. K. Tsotsos, "Biologically motivated local contextual modulation improves low-level visual feature representations," in *International Conference Image Analysis and Recognition*, pp. 79–88, Springer, 2012.

[204] J. Krüger, "Responses to wavelength contrast in the afferent visual systems of the cat and the rhesus monkey," *Vision research*, vol. 19, no. 12, pp. 1351 – 1358, 1979.

[205] M. Livingstone and D. Hubel, "Segregation of form, color, movement, and depth: anatomy, physiology, and perception," *Science*, vol. 240, no. 4853, pp. 740–749, 1988.

[206] M. J. Hawken and A. J. Parker, "Spatial properties of neurons in the monkey striate cortex," *Proceedings of the Royal society of London. Series B. Biological sciences*, vol. 231, no. 1263, pp. 251–288, 1987.

[207] S. Marĉelja, "Mathematical description of the responses of simple cortical cells," *Journal of the Optical Society of America*, vol. 70, no. 11, pp. 1297–1300, 1980.

[208] A. W. Roe, L. Chelazzi, C. E. Connor, B. R. Conway, I. Fujita, J. L. Gallant, H. Lu, and W. Vanduffel, "Toward a unified theory of visual area V4," *Neuron*, vol. 74, no. 1, pp. 12–29, 2012.

[209] M. Ito and H. Komatsu, "Representation of angles embedded within contour stimuli in area V2 of macaque monkeys," *Journal of Neuroscience*, vol. 24, no. 13, pp. 3313–3324, 2004.

[210] J. Hegdé and D. C. Van Essen, "Selectivity for complex shapes in primate visual area V2," *Journal of Neuroscience*, vol. 20, no. 5, pp. RC61–RC61, 2000.

[211] E. Kobatake and K. Tanaka, "Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex," *Journal of neurophysiology*, vol. 71, no. 3, pp. 856–867, 1994.

[212] W. H. Merigan, "Basic visual capacities and shape discrimination after lesions of extrastriate area V4 in macaques," *Visual neuroscience*, vol. 13, no. 1, pp. 51–60, 1996.

[213] C. R. Ponce, T. S. Hartmann, and M. S. Livingstone, "End-stopping predicts curvature tuning along the ventral stream," *Journal of Neuroscience*, vol. 37, no. 3, pp. 648–659, 2017.

[214] J. Bell, E. Gheorghiu, *et al.*, "Orientation tuning of curvature adaptation reveals both curvature-polarity-selective and non-selective mechanisms," *Journal of Vision*, vol. 9, no. 12, pp. 3–3, 2009.

[215] T. O. Sharpee, M. Kouh, and J. H. Reynolds, "Trade-off between curvature tuning and position invariance in visual area V4," *Proceedings of the National Academy of Sciences, U.S.A.*, vol. 110, no. 28, pp. 11618–11623, 2013.

[216] A. S. Nandy, T. O. Sharpee, J. H. Reynolds, and J. F. Mitchell, "The fine structure of shape tuning in area V4," *Neuron*, vol. 78, no. 6, pp. 1102–1115, 2013.

[217] T. D. Oleskiw, A. Nowack, and A. Pasupathy, "Joint coding of shape and blur in area V4," *Nature communications*, vol. 9, no. 1, pp. 1–13, 2018.

[218] D. V. Popovkina, W. Bair, and A. Pasupathy, "Modeling diverse responses to filled and outline shapes in macaque V4," *Journal of neurophysiology*, vol. 121, no. 3, pp. 1059–1077, 2019.

[219] B. N. Bushnell, P. J. Harding, Y. Kosai, and A. Pasupathy, "Partial occlusion modulates contour-based shape encoding in primate area V4," *Journal of Neuroscience*, vol. 31, no. 11, pp. 4012–4024, 2011.

[220] Y. Kosai, Y. El-Shamayleh, A. M. Fyall, and A. Pasupathy, "The role of visual area V4 in the discrimination of partially occluded shapes," *Journal of Neuroscience*, vol. 34, no. 25, pp. 8570–8584, 2014.

[221] A. M. Fyall, Y. El-Shamayleh, H. Choi, E. Shea-Brown, and A. Pasupathy, "Dynamic representation of partially occluded objects in primate prefrontal and visual cortex," *eLife*, vol. 6, p. e25784, 2017.

[222] R. Jiang, I. M. Andolina, M. Li, and S. Tang, "Clustered functional domains for curves and corners in cortical area V4," *eLife*, vol. 10, p. e63798, 2021.

[223] R. Tang, Q. Song, Y. Li, R. Zhang, X. Cai, and H. D. Lu, "Curvature-processing domains in primate v4," *eLife*, vol. 9, p. e57502, 2020.

[224] "Conformation," in *Merriam-Webster.com 2021*, https://www.merriam-webster.com/dictionary/conformation, (2021-07-30).

[225] S. V. David, B. Y. Hayden, and J. L. Gallant, "Spectral receptive field properties explain shape selectivity in area V4," *Journal of neurophysiology*, vol. 96, no. 6, pp. 3492–3505, 2006.

[226] T. D. Oleskiw, A. Pasupathy, and W. Bair, "Spectral receptive fields do not explain tuning for boundary curvature in V4," *Journal of neurophysiology*, vol. 112, no. 9, pp. 2114–2122, 2014.

[227] X. Hu, J. Zhang, J. Li, and B. Zhang, "Sparsity-regularized hmax for visual recognition," *PLoS ONE*, vol. 9, no. 1, p. e81813, 2014.

[228] Y. Wang and L. Deng, "Modeling object recognition in visual cortex using multiple firing k-means and non-negative sparse coding," *Signal Processing*, vol. 124, pp. 198–209, 2016.

[229] A. Eguchi, B. M. Mender, B. Evans, G. Humphreys, and S. Stringer, "Computational modelling of the neural representation of object shape in the primate ventral visual system," *Frontiers in computational neuroscience*, vol. 9, p. 100, 2015.

[230] H. Wei and Z. Dong, "V4 neural network model for shape-based feature extraction and object discrimination," *Cognitive Computation*, vol. 7, no. 6, pp. 753–762, 2015.

[231] A. Rodríguez-Sánchez, S. Oberleiter, H. Xiong, and J. Piater, "Learning V4 curvature cell populations from sparse endstopped cells," in *International Conference on Artificial Neural Networks*, pp. 463–471, Springer, 2016.

[232] Y. Hatori, T. Mashita, and K. Sakai, "Sparse coding generates curvature selectivity in V4 neurons," *Journal of the Optical Society of America A*, vol. 33, no. 4, pp. 527–537, 2016.

[233] S. Kawakami, T. Ito, Y. Makino, M. Hashimoto, and M. Yano, "A cell model in the ventral visual pathway for the detection of circles of curvature constituting figures," *Heliyon*, vol. 6, no. 11, p. e05397, 2020.

[234] R. Abbasi-Asl, Y. Chen, A. Bloniarz, M. Oliver, B. D. Willmore, J. L. Gallant, and B. Yu, "The deeptune framework for modeling and characterizing neurons in visual cortex area V4," *bioRxiv*, p. 465534, 2018.

[235] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Robust Object Recognition with Cortex-Like Mechanisms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 411–426, 2007.

[236] T. Serre, A. Oliva, and T. Poggio, "A feedforward architecture accounts for rapid categorization," *Proceedings of the National Academy of Sciences, U.S.A.*, vol. 104, no. 15, pp. 6424–6429, 2007.

[237] T. T. Wu, K. Lange, *et al.*, "Coordinate descent algorithms for lasso penalized regression," *The Annals of Applied Statistics*, vol. 2, no. 1, pp. 224–244, 2008.

[238] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[239] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, 1996.

[240] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Advances in Neural Information Processing Systems*, pp. 801–808, 2007.

[241] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the International Conference on Machine Learning*, pp. 689–696, 2009.

[242] M. Yaghoobi, L. Daudet, and M. E. Davies, "Parametric dictionary design for sparse coding," *IEEE Transactions on Signal Processing*, vol. 57, no. 12, pp. 4800–4810, 2009.

[243] M. Ataee, H. Zayyani, M. Babaie-Zadeh, and C. Jutten, "Parametric dictionary learning using steepest descent," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1978–1981, 2010.

[244] M. Beyeler, E. L. Rounds, K. D. Carlson, N. Dutt, and J. L. Krichmar, "Neural correlates of sparse coding and dimensionality reduction," *PLoS computational biology*, vol. 15, no. 6, p. e1006908, 2019.

[245] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[246] P. O. Hoyer and A. Hyvärinen, "A multi-layer sparse coding network learns contour coding from natural images," *Vision research*, vol. 42, no. 12, pp. 1593–1605, 2002.

[247] P. O. Hoyer, "Modeling receptive fields with non-negative sparse coding," *Neurocomputing*, vol. 52, pp. 547–552, 2003.

[248] K. Tsunoda, Y. Yamane, M. Nishizaki, and M. Tanifuji, "Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns," *Nature neuroscience*, vol. 4, no. 8, pp. 832–838, 2001.

[249] W. Bair, D. Popovkina, A. De, and A. Pasupathy, "Modeling shape representation in area V4," *MODVIS*, 2015.

[250] R. Tibshirani, "The lasso method for variable selection in the cox model," *Statistics in medicine*, vol. 16, no. 4, pp. 385–395, 1997.

[251] S. J. Wright, "Coordinate descent algorithms," *Mathematical Programming*, vol. 151, no. 1, pp. 3–34, 2015.

[252] K. B. Petersen and M. S. Pedersen, "The matrix cookbook," October 2008.

Appendix A

The RBO Network Neuron Responses

Responses of all mBO neurons in the RBO are depicted. In total, there are 16 mBO cells of type 1 and type 3 as classified by Zhou *et al*. [11].

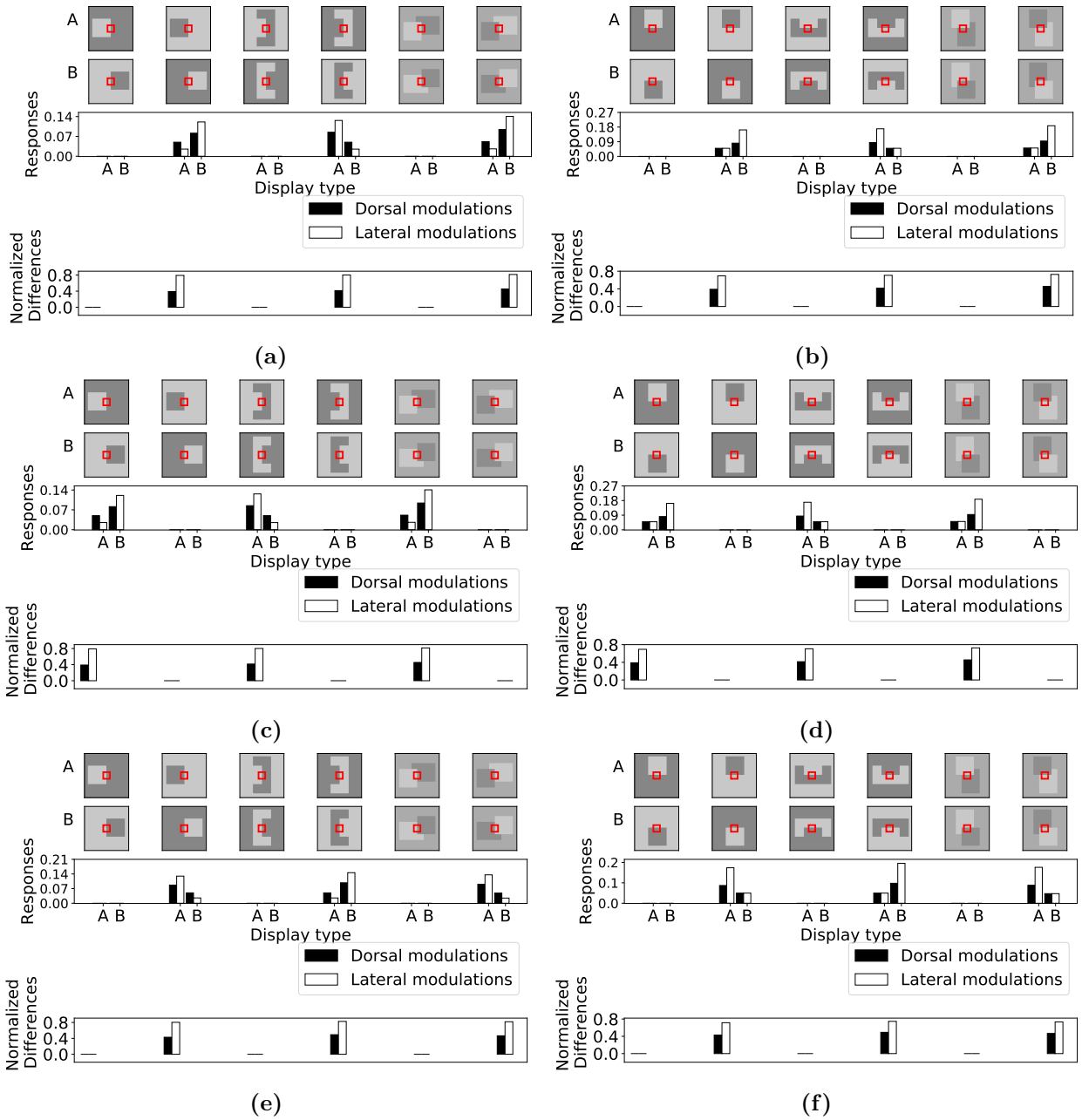


Figure A.1: Example of model border ownership responses to stimuli used in the neurophysiological experiments by [11]. Each subplot is separated from that of others by vertical and horizontal lines. In each subplot, the set of stimuli images are shown in two rows. Responses of the neurons to each pair of stimuli in each column are plotted below the column in the middle bar plot. The bottom plot represents the normalized differences for before and after RL. In both plots in each sub-figure, dark bars are responses after early recurrence and light bars are responses after relaxation labeling.

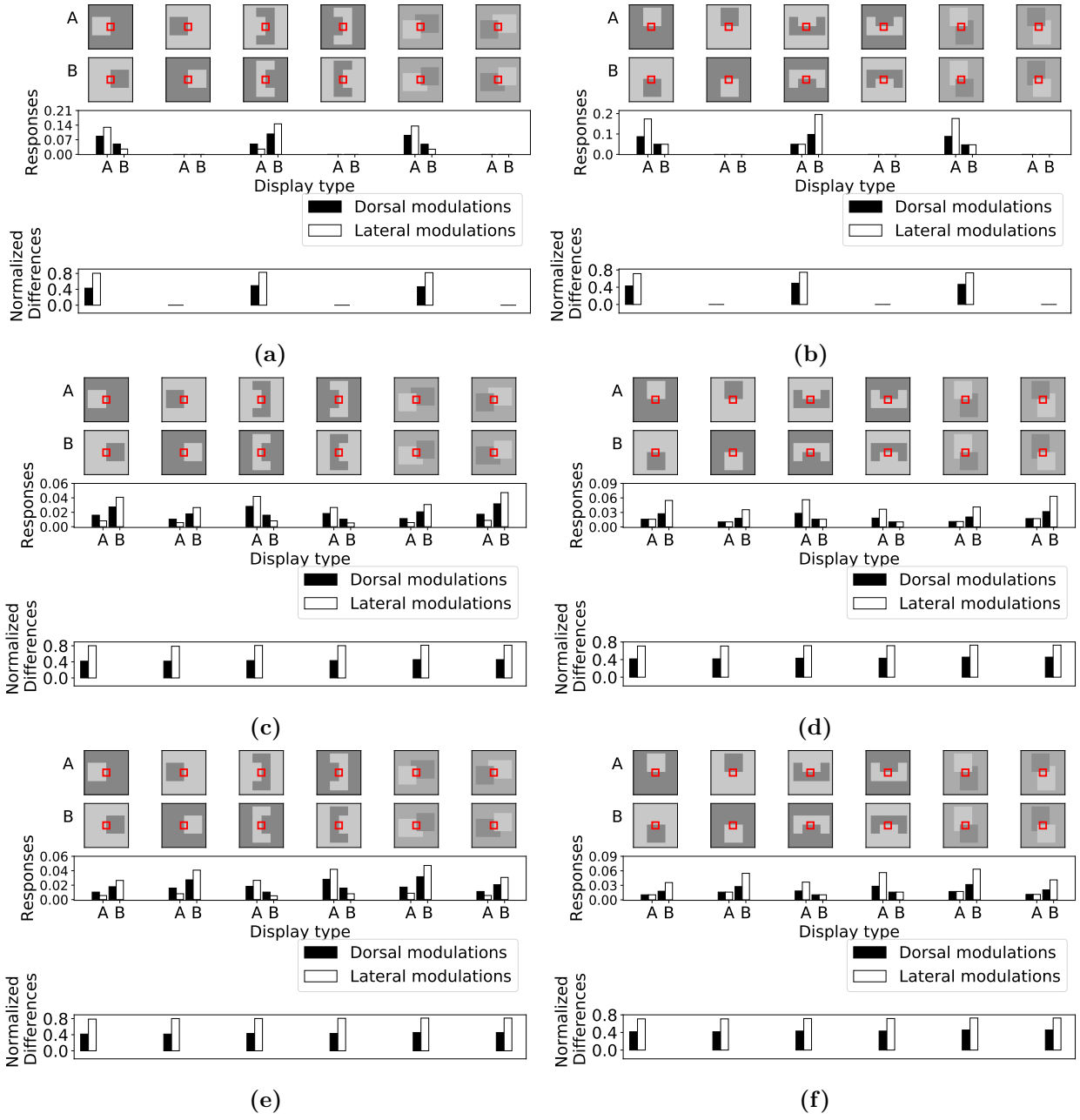


Figure A.2: Example of model border ownership responses to stimuli used in the neurophysiological experiments by [11]. Each subplot is separated from that of others by vertical and horizontal lines. In each subplot, the set of stimuli images are shown in two rows. Responses of the neurons to each pair of stimuli in each column are plotted below the column in the middle bar plot. The bottom plot represents the normalized differences for before and after RL. In both plots in each sub-figure, dark bars are responses after early recurrence and light bars are responses after relaxation labeling.

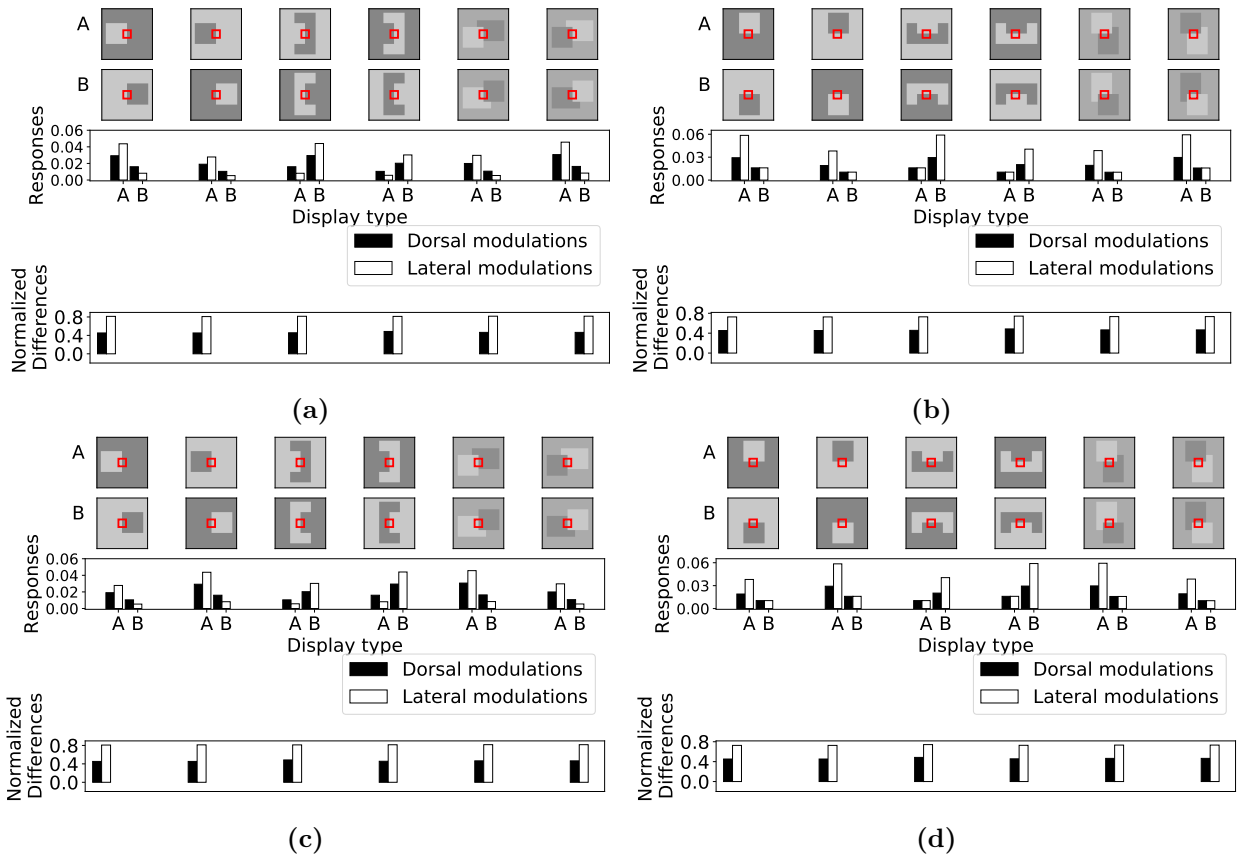


Figure A.3: Example of model border ownership responses to stimuli used in the neurophysiological experiments by [11]. Each subplot is separated from that of others by vertical and horizontal lines. In each subplot, the set of stimuli images are shown in two rows. Responses of the neurons to each pair of stimuli in each column are plotted below the column in the middle bar plot. The bottom plot represents the normalized differences for before and after RL. In both plots in each sub-figure, dark bars are responses after early recurrence and light bars are responses after relaxation labeling.

Appendix B

Sparse Coding Optimization

In order to learn V4 receptive fields from the stimulus-response pairs obtained from cell recordings, we solve the optimization problem in Equation 4.12 by solving for γ and $\mathbf{V}$, which i-th column forms the parameters of the i-th Gaussian. That is, the i-th column of $\mathbf{V}$ will be $\mathbf{v}_i = [\mu_{ix}, \mu_{iy}, \sigma_{ixx}, \sigma_{ixy}, \sigma_{iyy}]^T$, where $\mu_i = (\mu_{ix}, \mu_{iy})^T$ and $\Sigma_i = \begin{bmatrix} \sigma_{ixx} & \sigma_{ixy} \\ \sigma_{ixy} & \sigma_{iyy} \end{bmatrix}$ are the mean and covariance of the i-th Gaussian. Let

$$E = \sum_{t=1}^{\tau} \frac{1}{2} \|R_t - D_t^T \gamma\|_2 + \lambda \|\gamma\|_1. \tag{B.1}$$

In minimizing E with respect to γ and the elements of $\mathbf{V}$, we iteratively alternate between optimizing for γ and the parameters of Gaussians since the derivate of one is a function of the other. Optimizing E with respect to γ is the lasso problem [250] and we can use coordinate descent [251] for this subproblem. In order to solve for the parameters of the Gaussians, our approach is similar to that of parametric dictionary learning methods such as that of Ataee *et al.* [243]. The derivative with respect to v_{ij} is:

$$\frac{\partial E}{\partial v_{jk}.} = \sum_{t=1}^{\tau} \frac{1}{2} \frac{\partial \|R_t - D_t^T \gamma\|_2}{\partial v_{jk}}. \tag{B.2}$$

We can rewrite

$$\|R_t - D_t^T \gamma\|_2 = (R_t - D_t^T \gamma)^T (R_t - D_t^T \gamma),$$

where we have used short notations $C(t, i)$ and $G(i)$ for brevity. As a result, we have

$$\begin{aligned} \frac{\partial \|R_t - D_t^T \gamma\|_2}{\partial v_{jk}} &= \frac{\partial}{\partial v_{jk}} (-2 R_t D_t^T \gamma) + \\ &\quad \frac{\partial}{\partial v_{jk}} (\gamma^T D_t)(D_t^T \gamma). \end{aligned} \tag{B.3}$$

Note that in these equations R_t is a scalar while D_t and γ are 72×1 vectors. We will compute the derivatives in the first and second terms of the equation above separately. For the first term, we have:

$$\frac{\partial}{\partial v_{jk}} (-2 R_t D_t^T \gamma) = \frac{\partial}{\partial v_{jk}} (-2 R_t \sum_{i=1}^{72} d_t(i) \gamma(i)),$$

where we have replaced the dot product of D_t and γ. Note that the derivatives of R_t, $\gamma(i), \forall i$, and

$d_t(l), \forall l \neq k$ with respect to v_{jk} are zero. Therefore,

$$\frac{\partial}{\partial v_{jk}}(-2R_t D_t^T \gamma) = -2R_t \gamma(k) \frac{\partial d_t(k)}{\partial v_{jk}}.$$

Replacing $d_t(k)$ with Equation 4.11 gives

$$\frac{\partial}{\partial v_{jk}}(-2R_t D_t^T \gamma) = -2R_t \gamma(k) \frac{\partial}{\partial v_{jk}}(\sum_{x,y} C(t,k) \cdot G(k)),$$

where we have used short notations for brevity. Similarly, the derivative of $C(t,k)$ with respect to v_{jk} is zero since the curvature maps are not functions of the Gaussian kernels, resulting in

$$\frac{\partial}{\partial v_{jk}}(-2R_t D_t^T \gamma) = -2R_t \gamma(k)(\sum_{x,y} C(t,k) \cdot \frac{\partial G(k)}{\partial v_{jk}}), \tag{B.4}$$

where $\frac{\partial G(k)}{\partial v_{jk}}$ is the derivative of the k-th Gaussian kernel with respect to its j-th parameter. We will compute the derivative of a Gaussian with respect to its parameters later. Now, we can move to computing the derivative in the second term of Equation B.3. In each line, the rule we have applied from the line above will be mentioned in brackets. In what follows, **tr** denotes the trace of a matrix. For the second term we have

$$
\begin{aligned}
\frac{\partial}{\partial v_{jk}}(\gamma^T D_t D_t^T \gamma) = & \quad \frac{\partial}{\partial v_{jk}}(\mathbf{tr}(\gamma^T D_t D_t^T \gamma)) && [a = \mathbf{tr}(a) \text{ for scalar } a] \\
= & \quad \frac{\partial}{\partial v_{jk}}(\mathbf{tr}(\gamma\gamma^T D_t D_t^T)) && [\mathbf{tr}(AB) = \mathbf{tr}(BA)] \\
= & \quad \frac{\partial}{\partial v_{jk}}(\mathbf{tr}(PH)) && [\text{Let } P = \gamma\gamma^T, \\
& & & H = D_t D_t^T] \\
= & \quad \frac{\partial}{\partial v_{jk}} \sum_m \sum_n P_{mn} H_{nm} && [\text{Definition of } \mathbf{tr} \\
& & & \text{of a product}] \\
= & \quad \sum_{m \neq k} \frac{\partial}{\partial v_{jk}}(P_{mk} H_{km}) + \sum_n \frac{\partial}{\partial v_{jk}}(P_{kn} H_{nk}) && [\text{Only the } k\text{-th row and} \\
& & & \text{column of } H \text{ are} \\
& & & \text{functions of } v_{jk}] \\
= & \quad 2 \sum_{m \neq k} \frac{\partial}{\partial v_{jk}}(P_{mk} H_{km}) + && [\text{Both } P, H \text{ are} \\
& \quad \frac{\partial}{\partial v_{jk}}(P_{kk} H_{kk}) && \text{symmetric matrices}] \\
= & \quad 2 \sum_{m \neq k} P_{mk} \frac{\partial}{\partial v_{jk}}(H_{km}) + P_{kk} \frac{\partial}{\partial v_{jk}}(H_{kk}) && [P \text{ is not a function of} \\
& & & v_{jk}] \\
= & \quad 2 \sum_{m \neq k} P_{mk} \frac{\partial}{\partial v_{jk}}(d_t(k) d_t(m)) + && [\text{Since } H = D_t D_t^T] \\
& \quad P_{kk} \frac{\partial}{\partial v_{jk}}(d_t(k) d_t(k)) & & \\
= & \quad 2 \sum_{m \neq k} P_{mk} d_t(m) \frac{\partial}{\partial v_{jk}}(d_t(k)) + && [d_t(m) \text{ is not a function} \\
& & & \text{of } v_{jk}] \\
& \quad 2 P_{kk} d_t(k) \frac{\partial}{\partial v_{jk}}(d_t(k)) && [\text{Taking the derivative of} \\
& & & d_t^2(k)] \\
= & \quad 2 \sum_m P_{mk} d_t(m) \frac{\partial}{\partial v_{jk}}(d_t(k)) && [\text{Gathering sums}] \\
= & \quad 2 \sum_m P_{mk} d_t(m) \sum_{x,y} C(t,k) \cdot \frac{\partial G(k)}{\partial v_{jk}}. && [\text{Replacing } d_t(k)] \qquad \text{(B.5)}
\end{aligned}
$$

Substituting Equations B.4 and B.5 into Equation B.3 gives

$$\frac{\partial \|R_t - D_t^T\gamma\|_2}{\partial v_{jk}} = -2R_t\gamma(k)(\sum_{x,y} C(t,k)\cdot\frac{\partial G(k)}{\partial v_{jk}}) + 2\sum_m P_{mk}d_t(m)\sum_{x,y} C(t,k)\cdot\frac{\partial G(k)}{\partial v_{jk}}.$$

To further simplify this equation, note that $P_{mk} = \gamma(m)\gamma(k)$. Therefore,

$$\begin{aligned}\frac{\partial \|R_t - D_t^T\gamma\|_2}{\partial v_{jk}} = & \quad 2\gamma(k)(-R_t + \sum_m \gamma(m)d_t(m))\sum_{x,y} C(t,k)\cdot\frac{\partial G(k)}{\partial v_{jk}} \\ = & \quad 2\gamma(k)(D_t^T\gamma - R_t)\sum_{x,y} C(t,k)\cdot\frac{\partial G(k)}{\partial v_{jk}} \end{aligned} \tag{B.6}$$

Now, we compute $\frac{\partial G(k)}{\partial v_{jk}}$ below and substitute those in the equation above. We will skip the index k for the k-th Gaussian kernel for simplicity. Each Gaussian kernel is

$$G(x,y) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left\{\frac{-1}{2}((x,y)^T-\mu)^T\Sigma^{-1}((x,y)^T-\mu)\right\},$$

where $\mu = (\mu_x, \mu_y)^T$ and $\Sigma = \begin{bmatrix}\sigma_{xx} & \sigma_{xy}\\ \sigma_{xy} & \sigma_{yy}\end{bmatrix}$. Our space is two dimensional and so, $n = 2$. Therefore,

$$\begin{aligned}\frac{\partial G(x,y)}{\partial \mu} = & \quad \exp\left\{\frac{-1}{2}((x,y)^T-\mu)^T\Sigma^{-1}((x,y)^T-\mu)\right\}\frac{\partial}{\partial\mu}\frac{1}{(2\pi)^{\frac{1}{2}}|\Sigma|^{\frac{1}{2}}} + \\ & \frac{1}{(2\pi)^{\frac{1}{2}}|\Sigma|^{\frac{1}{2}}}\frac{\partial}{\partial\mu}\exp\left\{\frac{-1}{2}((x,y)^T-\mu)^T\Sigma^{-1}((x,y)^T-\mu)\right\} \\ = & \quad \frac{1}{(2\pi)^{\frac{1}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left\{\frac{-1}{2}((x,y)^T-\mu)^T\Sigma^{-1}((x,y)^T-\mu)\right\}\times \\ & \frac{\partial}{\partial\mu}\left\{\frac{-1}{2}((x,y)^T-\mu)^T\Sigma^{-1}((x,y)^T-\mu)\right\} \\ = & \quad G(x,y)\times\frac{\partial}{\partial\mu}\left\{\frac{-1}{2}((x,y)^T-\mu)^T\Sigma^{-1}((x,y)^T-\mu)\right\} \\ = & \quad G(x,y)\times((x,y)^T-\mu)^T\Sigma^{-1}\end{aligned} \tag{B.7}$$

Before jumping to computing the derivative with respect to the covariance, as a reminder, we have [252]:

$$\frac{\partial|X|}{\partial X} = |X|(X^{-1})^T, \tag{B.8}$$

where $|X|$ is the determinant of the matrix X, and also for vectors a, b and matrix X we have,

$$\frac{\partial a^T X^{-1} b}{\partial X} = -X^{-T} a b^T X^{-T}. \tag{B.9}$$

Now, taking the derivative with respect to the covariance matrix, we have:

$$\begin{aligned}\frac{\partial}{\partial \Sigma} G(x,y) = \quad & \exp\left\{\frac{-1}{2}((x,y)^T - \mu)^T \Sigma^{-1}((x,y)^T - \mu)\right\} \frac{\partial}{\partial \Sigma} \frac{1}{(2\pi)^{\frac{1}{2}}|\Sigma|^{\frac{1}{2}}} + \\ & \frac{1}{(2\pi)^{\frac{1}{2}}|\Sigma|^{\frac{1}{2}}} \frac{\partial}{\partial \Sigma} \exp\left\{\frac{-1}{2}((x,y)^T - \mu)^T \Sigma^{-1}((x,y)^T - \mu)\right\}. \end{aligned} \tag{B.10}$$

We will work on each of the terms in Equation B.10 separately below. For the first term, we have:

$$\begin{aligned}\frac{\partial}{\partial \Sigma} \frac{1}{(2\pi)^{\frac{1}{2}}|\Sigma|^{\frac{1}{2}}} &= \frac{-1}{((2\pi)^{\frac{1}{2}}|\Sigma|^{\frac{1}{2}})^2} \times \frac{\partial}{\partial \Sigma}((2\pi)^{\frac{1}{2}}|\Sigma|^{\frac{1}{2}}) \\ &= \frac{-1}{((2\pi)^{\frac{1}{2}}|\Sigma|^{\frac{1}{2}})^2} \times (2\pi)^{\frac{1}{2}} \frac{\partial}{\partial \Sigma}(|\Sigma|^{\frac{1}{2}}) \\ &= \frac{-1}{((2\pi)^{\frac{1}{2}}|\Sigma|^{\frac{1}{2}})^2} \times (2\pi)^{\frac{1}{2}} \times \frac{1}{2}|\Sigma|^{-\frac{1}{2}} \frac{\partial}{\partial \Sigma}(|\Sigma|) \\ &= \frac{-1}{((2\pi)^{\frac{1}{2}}|\Sigma|^{\frac{1}{2}})^2} \times (2\pi)^{\frac{1}{2}} \times \frac{1}{2}|\Sigma|^{-\frac{1}{2}}|\Sigma|(\Sigma^{-1})^T \quad \text{[Due to Equation B.8]} \\ &= \frac{-1}{2(2\pi)^{\frac{1}{2}}|\Sigma|^{\frac{1}{2}}} \times (\Sigma^{-1})^T. \end{aligned} \tag{B.11}$$

For the second term, let $B = \exp\left\{\frac{-1}{2}(a^T \Sigma^{-1} a\right\}$, where we have used $a = ((x,y)^T - \mu)$. Then:

$$\begin{aligned}\frac{\partial B}{\partial \Sigma} &= B \times \frac{\partial}{\partial \Sigma}\left\{\frac{-1}{2}(a^T \Sigma^{-1} a\right\} \\ &= B \times \frac{-1}{2}(-\Sigma^{-T} a a^T \Sigma^{-T}). \quad \text{[Due to Equation B.9]} \end{aligned} \tag{B.12}$$

Replacing Equations B.11 and B.12 back into Equation B.10 gives:

$$\begin{aligned}
\frac{\partial}{\partial\Sigma}G(x,y) = & \exp\left\{\frac{-1}{2}(((x,y)^T-\mu)^T\Sigma^{-1}((x,y)^T-\mu)\right\}\times \\
& \frac{-1}{2(2\pi)^{\frac{1}{2}}|\Sigma|^{\frac{1}{2}}}\times(\Sigma^{-1})^T + \\
& \frac{1}{(2\pi)^{\frac{1}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left\{\frac{-1}{2}(((x,y)^T-\mu)^T\Sigma^{-1}((x,y)^T-\mu)\right\}\times \\
& \frac{-1}{2}(-\Sigma^{-T}((x,y)^T-\mu)((x,y)^T-\mu)^T\Sigma^{-T}) \\
= & \frac{-1}{2(2\pi)^{\frac{1}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left\{\frac{-1}{2}(((x,y)^T-\mu)^T\Sigma^{-1}((x,y)^T-\mu)\right\}\times \\
& \Sigma^{-1}(\Sigma-(x,y)^T-\mu)((x,y)^T-\mu)^T)\Sigma^{-1} \quad [\Sigma \text{ is symmetric}] \\
= & \frac{-1}{2}G(x,y)\times\Sigma^{-1}(\Sigma-(x,y)^T-\mu)((x,y)^T-\mu)^T)\Sigma^{-1}. \qquad \text{(B.13)}
\end{aligned}$$

Algorithm 1 V4 RF Learning

Input:

* Sets of eight curvature maps, $C(t,i,x,y)$,
* V4 responses, R_t, to stimulus t for all t in the training set,
* Sparsity weight γ,
* Threshold ξ,
* Learning rate α.

Output: Eight Gaussians, $G(t,i,x,y;\mu_i,\Sigma_i)$ and the sparse code vector γ

μ_i = visual field center $\forall i$
$\Sigma_i = \mathbf{I}\ \forall i$
$\gamma = \mathbf{1}$
$E_{\text{old}} = \infty$
Compute D_t from Equation 4.11
$E_{\text{new}} = \sum_{t=1}^{\tau}\frac{1}{2}\|R_t - D_t^T\gamma\|_2 + \lambda\|\gamma\|_1$
while $\|E_{\text{old}} - E_{\text{new}}\|_2 > \xi$ **do**
    for $i \in \{1,2,...,72\}$ **do**
        $\mu_i^{\text{new}} \leftarrow \mu_i^{\text{old}} - \alpha\frac{\partial E}{\partial\mu_i}$ from Equations B.2, B.6, B.7
        $\Sigma_i^{\text{new}} \leftarrow \Sigma_i^{\text{old}} - \alpha\frac{\partial E}{\partial\Sigma_i}$ from Equations B.2, B.6, B.13
    end for
    Update D_t from Equation 4.11
    Learn γ with coordinate descent
    $E_{\text{old}} \leftarrow E_{\text{new}}$
    $E_{\text{new}} \leftarrow \sum_{t=1}^{\tau}\frac{1}{2}\|R_t - D_t^T\gamma\|_2 + \lambda\|\gamma\|_1$
end while