COMPUTATIONAL MODELS OF FEATURE REPRESENTATIONS IN THE VENTRAL VISUAL STREAM

by
Shi Pui Donald Li

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
June 2021

ABSTRACT

Understanding vision requires unpacking the representations of the visual processing hierarchy. One major and unresolved challenge is to understand the representations of high-level category-selective areas – areas that respond preferentially to certain semantic categories of stimuli (e.g., scene-selective areas respond more to scenes than objects). Attempts at characterizing the representations of category-selective areas have been hampered by the difficulty of describing their complex perceptual representations in words — these representations exist in an "ineffable valley" between the describable patterns of perceptual features (e.g., edges, colors) and the commonsense concepts of visual cognition (e.g., object categories). Here I developed a novel approach to identify the emergent properties of mid-level representations in purely feedforward deep convolutional neural network (CNN) models of category-selective cortex. Using this approach, CNN models were fit to scene-evoked fMRI responses in both scene-selective cortex and object-selective cortex. This method uses a semantically-guided image-occlusion procedure together with behavioral ratings to systematically characterize the tuning profiles of the category-selective CNNs. I found that while the representations in category-selective CNNs appear complex and difficult to describe at a surface level, large-scale computational analyses can reveal 1) interpretable descriptions of mid-level feature representations and 2) the emergence of semantic selectivity through purely bottom-up perceptual feature tuning. Specifically, these models provide a proof-of-principle demonstration of how the semantic selectivity of category-selective regions could arise through perceptual-feature tuning in a small series of feedforward computations. These effects were robust to variations of model hyperparameters and were reproducible across different CNN architectures and training procedures. Taken together, I demonstrated

how large datasets and *in-silico* computational models can be used to reveal the tuning profiles of category-selective regions and to identify how semantic preferences could emerge through bottom-up processes.

Dissertation Committee

Michael Bonner (primary advisor), Cognitive Science

Michael McCloskey, Cognitive Science

Leyla Isik, Cognitive Science

Charles Edward Connor, Zanvyl Krieger Mind/Brain Institute

Chaz Firestone, Psychological and Brain Sciences

ACKNOWLEDGEMENTS

I am very grateful to all the people who made this work possible. First, I want to express my gratitude to my committee – Michael Bonner, Michael McCloskey, Leyla Isik, Ed Connor and Chaz Firestone – for taking the time to provide me constructive and valuable feedback to my work. Specifically, I am extremely thankful to my dream team of advisors, Michael Bonner, Brenda Rapp and Soojin Park. They have all taught me, guided me, and supported me to be a scientist who can critically think and appreciate science. I always came to them with immature and vague ideas, and they always kindly guided me to think through all the hurdles and refine them into solid scientific arguments. I never would have been made my projects possible without their generous support and guidance.

Next, I would love to thank all the members in the Cognitive Science department. In particular, I truly appreciate Paul Smolensky for offering the Foundation of Cognitive Science course, which helped me to think more like a cognitive scientist and build up my own view toward science. Throughout my graduate school, I enjoyed attending Michael McCloskey's lab meeting a lot as Mike always demonstrated how to think critically. Apart from faculty members, I also want to thank all the staffs in the department, my undergraduate research assistant Jiayu Shao, all the fMRI technicians in Kennedy Krieger Institute, and the participants in my experiments. They together have made my experiments and analyzes possible.

Moreover, I have special thanks to my cohorts, officemates, labmates and friends, including Celia Litovsky, Jane Lukten, Sadwi Srinivas, Jeongho Park, Bob Wiley, Grusha Prasad, Kyriaki Neophytou, Natalia Talmina, Alon Hafri, Caterina Magri, Emalie McMahon, Sherry Chien and Zhengang Lu. They have all been good

listeners and empathizers when I was stressed out. It was so much fun have all of you as friends and colleagues.

Last but not least, I dedicate this dissertation to my parents. My parents are extremely supportive at all the time and allow me to pursue anything that I am interested in, without any worry. Mom and Dad, I would not have been able to take on all the challenges without you, so thank you for all your love and support.

TABLES OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

Vision is a fundamental and essential task for most species. It is defined as the ability to interpret the surrounding environment using light in the visual spectrum. As effortless as it may seem, vision is a series of complex computations which infer the 3-D world from the 2-D retinal input based on a number of assumptions. The goal of vision science is to understand how one can extracts information from the visual input by decomposing vision into functional components (Kriegeskorte & Douglas, 2018).

To understand each functional component, scientists analyze the nature of information processed by each functional component. These components are situated in a hierarchy of visual processing and can be grouped into low-level, mid-level and high-level vision depending on the properties being analyzed (see Figure 1.1). Low-level vision, such as color, motion and edge detection, focuses on analyzing the local perceptual properties of the visual input. Mid-level vision includes the representation of shapes, textures, 3-D depth cues and other complex features that are useful for inferring the structures and content of the environment (Anderson, 2020). There is an agreement in the field that both low-level and mid-level representations in this hierarchy are perceptual in nature, and thus can be computed through a bottom-up feedforward process, and indeed, for many of these low-level and mid-level representations, there exist quantitative models of how the representations could be computed from images. On the top of the hierarchy, high-level vision is involved in interpreting the abstract semantic properties of the visual input, which includes object recognition, face recognition and scene parsing (Cox, 2014). For many of these high-level visual processes, the field lacks explicit quantitative models of how abstract semantic representations arise in the brain. Most theories of semantic representation in high-level vision are descriptive in nature, and there is debate over which descriptive

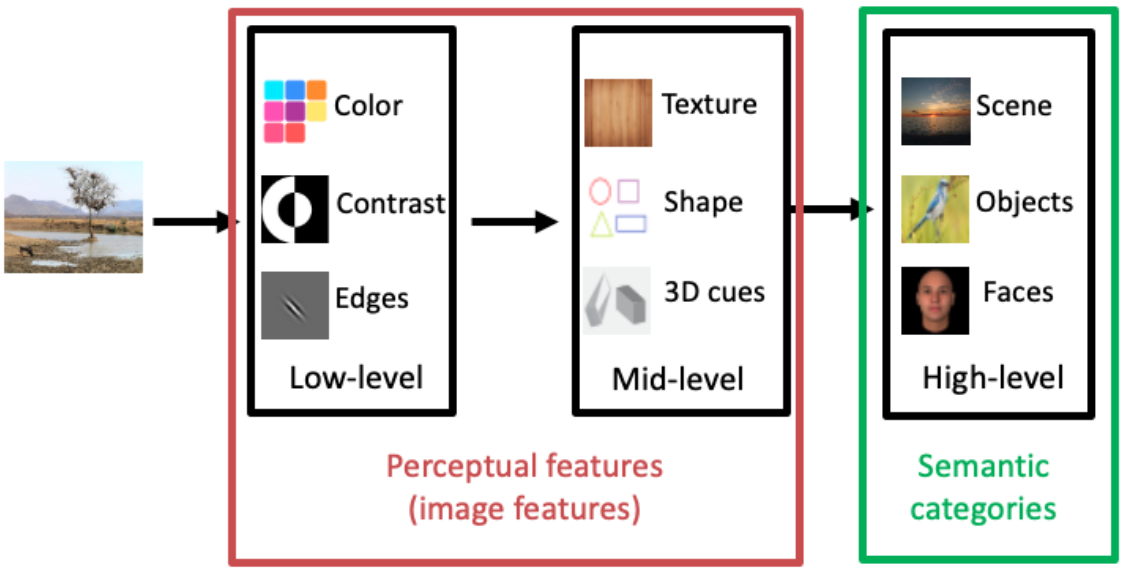
theories best explain the nature of the underlying representations (Cichy & Kaiser, 2019).

There are two main hurdles in understanding the nature of representations in high-level vision. First, a well-established model with explicit descriptions of the representational content is required, but many current models including some of the descriptive theories are lacking an operational description of the representation. For example, some descriptive theories do not explicitly discuss how the high-level semantic information is represented mechanistically (Cichy & Kaiser, 2019). Second, the nature of the representations in some neural substrates remains highly debatable. Specifically, in many neural substrates that are speculated to be related to high-level vision, there is a debate on whether they purely encode mid-level perceptual properties or high-level abstract semantic properties. Within those discussions, some scientists believe that the high-level semantic selectivity observed in those regions are due to the confounds between semantic and perceptual properties. On the other hand, some believe that these neural substrates purely encode abstract semantic properties (e.g., landmark, scene category, object category). While this debate about the nature of selectivity in the visual cortex is an important question, I argue that dichotomizing the interpretation into purely perceptual or purely semantic properties is an over-simplification that arises when attempting to understand visual cortex through descriptive theories that do not seriously grapple with the underlying computational mechanisms. In this dissertation, I will propose a feedforward computational model of category-selective areas – neural substrates that are considered to perform high-level vision – together with an *in-silico* experimental procedure to demonstrate that 1) a feedforward computational model can explain a significant amount of variance in

category-selective representations and 2) the semantic selectivity in high-level vision could be an emergent phenomenon of mid-level feature tuning.

Figure 1.1 ***Illustration of visual hierarchy***

Visual input is first processed by low-level and mid-level vision to extract perceptual features, then high-level vision can infer the semantic information of the visual input.



In this work, I will focus on category-selective regions of high-level visual cortex. These are regions that show selectivity to certain semantic image categories, for example, lateral occipital complex (LOC) shows a higher activation to object images versus scenes (Malach et al., 1995; Grill-Spector et al., 2001) and parahippocampal place area (PPA) shows a higher activation to scenes versus objects (R. Epstein & Kanwisher, 1998).

My work sheds light on the debate over high-level visual representation by first providing a computational model of these regions. Recent advancements of biologically inspired deep convolutional neural networks (CNNs) have yielded image-computable models that can provide insights into the computational basis of visual cognition (Cichy

& Kaiser, 2019). Recent findings have shown that CNNs are the best performing computational models in accounting for neural activity in primate visual cortices (Khaligh-Razavi & Kriegeskorte, 2014; Schrimpf et al., 2018; Yamins et al., 2014). Recent evidence also suggested that CNNs are excellent models in explaining both scene- (Bonner & Epstein, 2018; I. I. Groen et al., 2018) and object-selective areas (Radoslaw Martin Cichy et al., 2016; Khaligh-Razavi & Kriegeskorte, 2014). Therefore, I used CNN encoding models as a tool to understand the selectivity profiles of the high-level category-selective areas.

CNN encoding models are feedforward image-computable models with many simple processing units that extract perceptual features and image statistics from the visual input. The feedforward nature of these models allows me to test whether the model activations are driven by mid-level perceptual tunings properties. In my analyzes, these models showed that the fMRI responses in visual cortex are well explained by CNN features, which suggests that the representation in the category-selective areas could be driven by image-computable perceptual features, as modeled by the CNNs.

I then developed an *in-silico* experimental procedure – semantic preference mapping – to test whether this fully perceptually driven model exhibits the previously identified semantic preferences of the category-selective areas. This method utilizes a large image dataset to identify the selectivity of the models to certain object categories by examining how model activations are affected when a target object is occluded in a natural image. If the model is sensitive to object category X, then when object category X is occluded in the image, the model should show a lower activation compared to the unoccluded image. Using this logic, I characterized the selectivity profiles of CNN models that were fit to category-selective areas. When I correlated the selectivity profiles from the model to human object property ratings, I found that the tuning

profiles can be explained by interpretable object properties, which suggests the models do capture high-level semantic properties of the objects and that this semantic selectivity emerges through tuning to mid-level perceptual features.

While the results suggested that high-level category-selective area models capture the covariance between perceptual features and semantic attributes in the natural statistics of vision, the results also suggests that these models are sensitive to some lower-level perceptual features like curvature and cardinal orientations. These low-level perceptual biases have been previously identified in category-selective regions (Nasr & Tootell, 2012; Yue et al., 2020). To further test the curvilinearity preferences in these regions, I developed an image computable curvature model that can compute a curvature summary statistic from any given image. This curvature model was shown to capture a key representational dimension that differs across category-selective regions.

The dissertation is organized as follows. In Chapter 2, I will review the debate on the level of interpretation in the category-selective areas. In Chapter 3, CNN models will be reviewed, and I will discuss how to build computational models that utilize mid-level perceptual features from CNNs to predict fMRI activation in a large-scale fMRI dataset of scene perception. I will also discuss several experiments to verify and validate our modeling procedures. Building on these backgrounds, in Chapter 4, I will focus on understanding the tuning profiles of the computational models. First, I will explain the novel semantic preference mapping procedure. Second, I will describe a behavioral experiment for collecting object-property ratings to use in combination with the semantic preference mapping procedure. Lastly, I will demonstrate that the tuning profiles of the computational models are closely connected to interpretable object properties. In Chapter 5, I will characterize the low-level perceptual properties that also

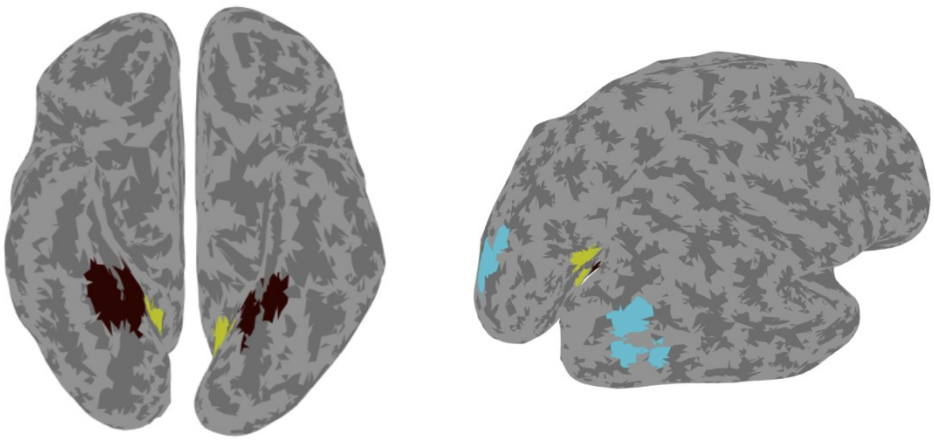
emerge in these computational models of category-selective areas. I also demonstrate an approach to build interpretable image-computable models to explain category-selective areas. By building a curvature model, I will show that category-selective areas are sensitive to the curvilinearity of the visual input. Here, I will argue that curvilinearity and cardinal orientations are important perceptual biases of category-selective regions. Chapter 6 will conclude the dissertation with a discussion of the theocratical implications of both the novel procedures and the findings. In addition, I will suggest further directions for building computational models of category-selective areas.

CHAPTER 2. LEVELS OF INTERPRETATION FOR CATEGORY-SELECTIVE REPRESENTATIONS

Neuroimaging studies have revealed regions along the ventral visual stream that respond preferentially to certain abstract stimulus categories. For example, lateral occipital complex (LOC) responds preferentially to objects. There are also several regions that respond strongly to scenes and landmarks, including parahippocampal place area (PPA), occipital place area (OPA) and retrosplenial cortex (RSC) (see Figure 2.1). Although these areas are functionally identified by their selectivity for categories of visual stimuli, these areas are also shown to be sensitive to low-level and mid-level perceptual features. In natural image statistics, there is an inherent correlation among low- and mid-level perceptual features and high-level semantic properties (R. A. Epstein & Baker, 2019; I. I. A. Groen et al., 2017). Therefore, it is hard to distinguish whether the responses of these regions are driven by the low-level perceptual features or the high-level semantic properties of the preferred stimuli, and there is a debate in the field over whether the selectivity profiles of these regions should be understood in terms of perceptual or semantic properties (I. I. A. Groen et al., 2017; Kim et al., 2009). In this chapter, I will review the current debate about the level of interpretation of the category-selective areas. In section 2.1, I will focus on the level of interpretation debate of the scene-selective ROIs. Section 2.2 will focus on the debate regarding the object-selective area.

Figure 2.1 ***Illustration of scene-selective areas***

The three functionally defined scene-selective areas are parahippocampal place area (PPA) in red, occipital place area (OPA) in blue and retrosplenial cortex (RSC) in yellow.

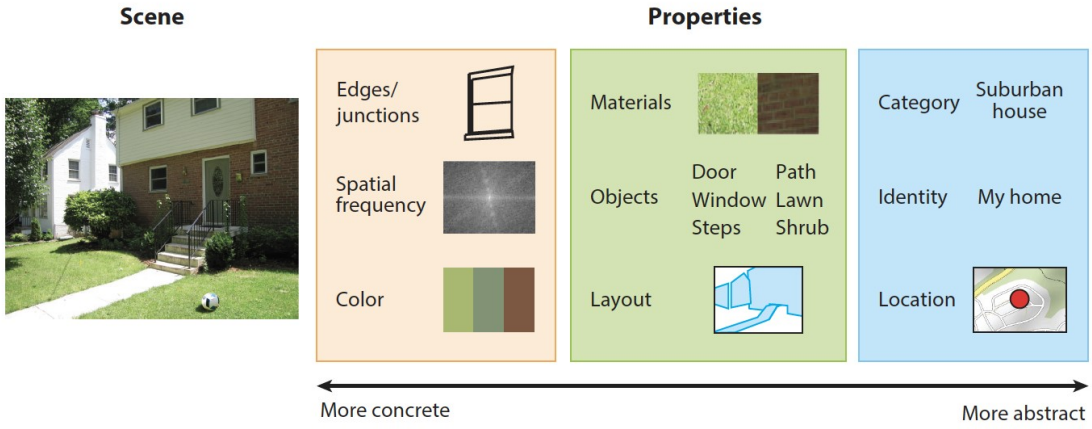


2.1. Scene-selective representations

Scene-selective areas are speculated to be involved in a variety of cognitive functions including landmark recognition and spatial navigation. A lot of different image properties ranging from abstract semantic properties to concrete perceptual features are speculated to support these functions (see Figure 2.2). In this section, I will focus on the representative scene-selective area PPA. PPA is the first identified scene-selective area, and it is defined based on stronger activations to scenes than other non-scene visual stimuli (R. Epstein & Kanwisher, 1998). To better highlight the debate on the level of interpretation in PPA representation, I will focus on the landmark object hypothesis, which suggests that one of the primary characteristics of PPA is its sensitivity to landmark objects in scenes.

Figure 2.2 ***Image properties associated with scene processing.***

Adapted from Epstein & Baker, 2019. The visual system analyzes multilevel properties of scenes. These properties include low-level features like spatial frequency and color, mid-level features like texture and layout and high-level semantic properties like category and geographical locations.



Landmarks are objects that are associated with a specific location in the world. Usually they are large in real-world size and fixed in location (Troiani et al., 2014). Multivoxel activation patterns in PPA are able to classify individual landmarks (R. A. Epstein & Morgan, 2012; Marchette et al., 2015; Morgan et al., 2011). For example, PPA representations can classify familiar buildings from different views, while object-selective areas fail in this classification. These results suggested that abstract landmark objects could be decoded from the PPA, and these PPA representations could carry abstract information about landmark identity.

Landmark objects consist of several different high-level semantic properties, and PPA is sensitive to those abstract semantic dimensions of landmarks. In a study by Troiani et al., 2014, researchers had examined the PPA sensitivity to several semantic object properties that are associated with landmark objects, including real-world size and fixedness. In the real-world size property, they showed that PPA is sensitive to

objects that are large, and it was speculated that large objects tend to be more fixed in location, and thus are more probable to serve as landmarks (Julian et al., 2017). Fixedness is defined as how fixed the object is in the environment. As one of the functions of landmark objects is to be used as a reference to define spatial location, landmark objects should usually be fixed in location. Troiani et al., 2014 found that PPA is also sensitive to fixed objects. The selectivity for large and fixed objects in PPA suggests PPA is tuned to high-level semantic properties of landmarks.

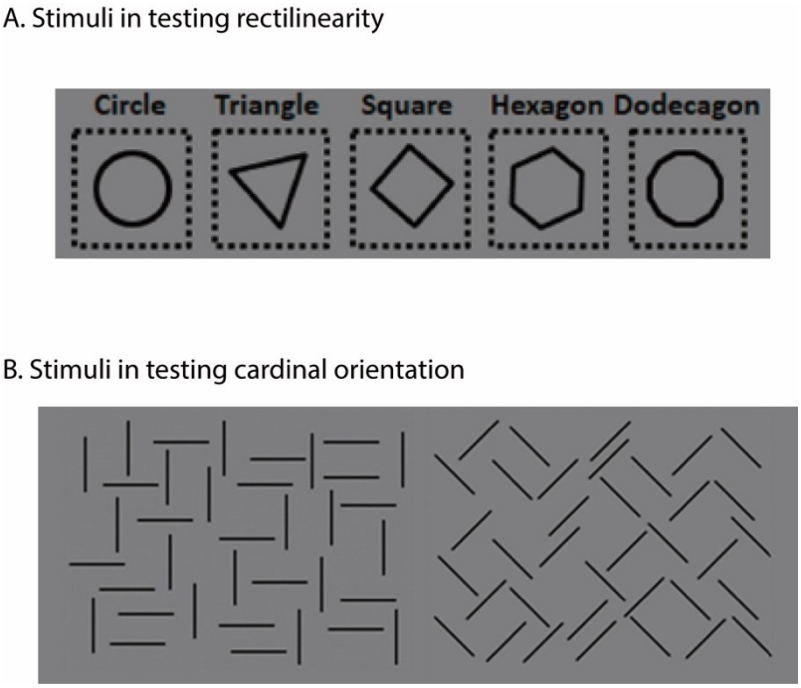
In addition to the evidence linking PPA to high-level landmark processing, there are also findings showing that PPA has low-level feature biases. PPA was shown to have retinotopic biases. The peripheral bias of PPA suggests that PPA tends to respond more strongly to stimuli in the periphery of the visual field (Silson et al., 2016; Silson et al., 2015). In a population receptive field analysis of PPA, PPA was shown to respond more strongly to the upper visual field. Although it has been argued that this retinotopic bias is consistent with a specific role in representing landmark objects, as large fixed objects usually occupy the periphery of the upper visual field (I. I. A. Groen et al., 2017). However, these retinotopic biases suggest that it is unlikely that PPA purely encodes abstract semantic information, otherwise it would not be expected to exhibit a low-level retinal location bias.

As mentioned above, landmark objects tend to have a lot of rectilinear contours and horizontal/vertical contours (i.e. contours at cardinal orientations) (Nasr et al., 2014; Nasr & Tootell, 2012). Therefore, scene-selective areas could be tuned to the low-level features of cardinal orientations and mid-level features of rectilinearity. In fact, a study by Nasr et al., 2014 shows that when scene-selective areas were presented with rectilinear stimuli compared to rounded stimuli, even when the stimuli were just simple shapes, these areas responded much more than when they were presented with

stimuli containing rounded shapes (see Figure 2.3 panel A). Given that these stimuli are not meaningful scenes, and do not define any spatial-layout information, these results suggested that scene-selective areas are highly tuned to the mid-level feature of rectilinear contours. Another study by Nasr & Tootell, 2012 demonstrated that scene-selective areas are tuned to the low-level feature of cardinal orientations. In this study, participants were presented with arrays of lines, each array contains lines in either cardinal or oblique orientations, and these stimuli again do not form any meaningful visual objects or scenes (see Figure 2.3 panel B). They observed that PPA is more activated to stimuli with cardinal orientations compared to oblique orientations, suggesting that PPA is selective to the low-level perceptual feature of cardinal orientations. Apart from cardinal orientations and rectilinearity, the representation in PPA is also modulated by the low-level perceptual feature of high spatial frequency (Rajimehr et al., 2011). Altogether, there is conflicting evidence on whether the responses of the scene-selective areas are driven by the abstract properties of landmarks or lower-level perceptual features.

Figure 2.3 ***Stimuli used in testing low- and mid-level features preferences in scene-selective areas.***

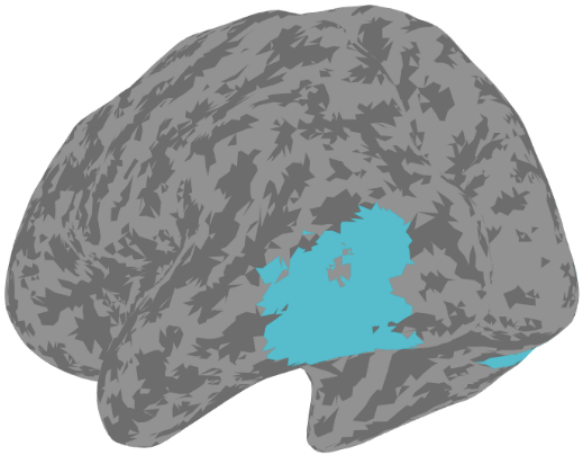
Adapted from Nasr et al., 2014 and Nasr & Tootell, 2012. A: stimuli used in Nasr et al., 2014 to test the rectilinear preferences in scene-selective areas. B: stimuli used in Nasr & Tootell, 2012 to test the cardinal orientation preferences in scene-selective areas.



2.2. Object-selective representations

As the name suggested, object-selective regions respond strongly when pictures of objects are shown compared to pictures of textures or scrambled objects. One of the most well-studied object-selective areas in humans is LOC (see Figure 2.4). Currently, the debate on the representation in LOC focuses on whether it encodes semantic object identity or perceptual object shape.

Figure 2.4 ***Illustration of object-selective area lateral occipital complex (blue).***



Object-selective areas like LOC were first thought to encode the abstract semantics of visual objects (Grill-Spector, 2003; Grill-Spector et al., 2001). LOC has long been speculated to encode object identity and object category. For example, Naselaris et al., 2009 were able to use LOC to reconstruct natural images from brain activity. In their study, they used the LOC representation as one of the semantic dimensions in their encoding model to encode the semantic category of the visual stimuli. The success of the reconstruction using the semantic encoding in LOC demonstrates that LOC representation contains information about the semantic categories of the visual input. Moreover, researchers found that LOC responses could be used to classify different object categories (e.g. chair vs. teapot) while it is insensitive to lower-level image features, such as the retinal size of the object (big teapot vs. small teapot) (Eger et al., 2008). This study suggested that LOC represents the abstract semantic category of objects in a manner that is invariant to view-specific perceptual information.

Although there is strong evidence that LOC encodes object identity and category, other studies suggest the LOC is representing perceptual features like object

shapes and object parts rather than object semantics (Hayworth & Biederman, 2006; Kim et al., 2009; Shpaner et al., 2013). In an fMRI adaptation study by Kim et al., 2009, they found no adaptation effects in LOC when the two objects only share object category but not physical shape; however, adaptation effects were observed when the two stimuli shared similar physical shapes, suggesting LOC is sensitive to shape rather than the semantic category of an object. This result suggests that LOC is sensitive to perceptual features like object shape that happens to be confounded with categorical information.

While the debate continues, Cichy et al., 2011 argues that the representation in LOC contains both high-level identity information and low-level location information. In their study, they tried to decode both object identity and object location from the LOC signal, while areas like early visual cortex (EVC) can only decode location, but not object category, LOC shows above chance classification performance in decoding both object identity and object location using images across different exemplars. This result indicated that the representation in LOC could be more complicated than containing purely semantic or purely perceptual information. Rather, it can represent both types of information.

More recent studies suggest that LOC representations are organized along continuous dimensions for high-level semantic properties, including animacy and real-world object size (Konkle & Caramazza, 2013; Konkle & Oliva, 2012). More interestingly, these dimensions are speculated to be correlated with the shape of objects. For example, small and animate objects tend to have more curvy shapes, while large, inanimate objects have more rectilinear contours (Konkle & Caramazza, 2013; Konkle & Oliva, 2012; B. Long et al., 2018; Torralba & Oliva, 2003). This speculation

suggested that the perceptual shape of objects covaries with the semantic categories of objects. Thus, LOC could be capturing such covariance in its representation.

In this chapter, I have introduced the debate on the level of interpretation of the representation in category-selective areas. While there is scientific evidence to support both sides of the argument, it is unclear whether category-selective areas are better understood in terms of tuning to perceptual or semantic properties. As the existence of inherent correlations among low- and mid-level perceptual features and high-level abstract semantic features make it hard to attribute the representation to a particular level of interpretation, building explicit computational models may help us in understanding the underlying representations. By having an explicit computational model, scientists can gain insight into the computations involved in transforming from perceptual into high-level semantic representations, and scientists can use large-scale experiments to understand the covariance between perceptual features and semantic properties in the natural statistics of images. In the following chapters, I will use computational models to address some of the key issues in this debate.

CHAPTER 3. COMPUTATIONAL MODELS OF THE CATEGORY-SELECTIVE AREAS

Chapter 2 reviewed the debate on the level of interpretation of both scene- and object-selective areas. Modern computational approaches like CNNs were shown to be promising in exploring human vision and high-level visual cortices as they outperform other computational models in explaining the visual cortex. In this chapter, I will focus on the development of the CNN encoding models of the category-selective cortices, these image-computable models first extract mid-level perceptual features of the input images, then map these features onto the neural representation of category-selective areas. These computational models serve as 1) a proof-of-principle that representation in high-level visual cortices can be predicted from a linear model applied to mid-level perceptual features and 2) a tool for the investigation of the semantic selectivity in high-level visual cortices. First, I will introduce the computations involved in deep convolutional neural network, and the principle of how it extracts perceptual features from input images. I will then introduce the dataset used in developing the models, and focus on the model architecture and training procedures. Lastly, I will present several *in-silico* experiments that demonstrate the predictive power of the CNN encoding models.

3.1. Deep Convolutional Neural Networks

Deep convolutional neural networks (CNNs) are a class of computational models that can perform a range of computer vision tasks, including challenging high-level tasks, like object recognition (Szegedy et al., n.d.), semantic segmentation (J. Long et al., 2015) and scene reconstruction (Aäron van den Oord & Kalchbrenner, 2016). There are a variety of specific classes of deep convolutional neural network (e.g.,

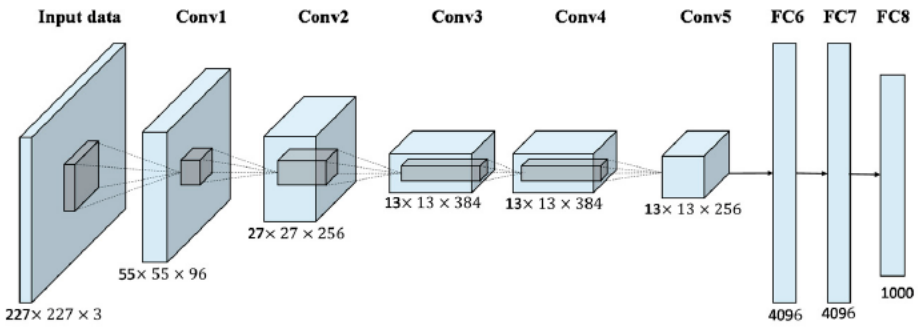
feedforward CNN, recurrent CNN, generative-adversarial CNN, etc.) that are designed for different tasks; however, since the goal in this study is to use CNNs to capture bottom-up mid-level perceptual features from input images, the discussion will focus on the class of feedforward CNNs. In particular, I will discuss a specific CNN, called AlexNet to demonstrate some characteristics of this class of network.

3.1.1. AlexNet

AlexNet was one of the first feedforward CNNs that was built to successfully perform image classification in natural images (Krizhevsky et al., 2012). This network consists of five convolutional layers and three fully connected layers. One of the important characteristics of this network is that it is able to perform semantic categorization of images without any feedback or recurrent connections. In other words, the network purely relies on feedforward feature extraction to accomplish the categorization task. The architecture of the network is shown in Figure 3.1.

Figure 3.1 ***AlexNet illustration from Han et al., 2017.***

There are seven layers in the network where the first five layers are convolutional layers, and the last two layers are fully connected layers.



In each convolutional layer, first, convolution operations are applied to the input feature map/image, which could be thought of as a similarity measure between the local feature map patches and the convolutional kernels. Each layer has a different number of convolutional kernels to capture different image features. Second, a non-linear ReLu activation function is applied to the output of the convolution. Third, a spatial max-pooling operation is applied to reduce the spatial dimension of the output feature maps. The convolutional layer was thought to capture important perceptual features and image statistics that are relevant to the task.

In the fully connected layer, all units are fully connected to all the units in the next layer, and the non-linear ReLu activation function is implemented after the linearly connected layer.

This model architecture is trained using the backpropagation algorithm and is shown to be powerful in both object classification (Krizhevsky et al., 2012) and scene classification (Zhou et al., 2017). In the field of cognitive neuroscience, using representational similarity analysis (RSA), Khaligh-Razavi and Kriegeskorte (2014) showed that the CNN representation in the fully connected layer correlated best with

the representation in IT, which is an object-selective area, while the CNN representation in the early convolutional layers correlate best with V1. The correlation between the representation of CNN and IT approaches the noise ceiling, which demonstrates CNN as the first computational model to explain almost all of the explainable variance in the IT representation. This result suggests that the features extracted from the CNN inner representation could well predict the representation in category-selective areas. Jozwik et al., 2017 also showed that CNN representation in the later layers outperforms perceptual feature-based model in predicting human object similarity judgement, which strengthen the evidence that CNN can be used to model human judgments of object similarity. Similar findings on neural and behavioral experiments are observed when using scene images (I. I. Groen et al., 2018), suggesting that the nature of CNN representation is consistent across different stimuli categories.

While earlier studies focus on mapping the earlier and later layer representation onto the brain, more recent findings suggest that the intermediate layers of the CNN could be informative in studying mid-level vision. For example, B. Long et al., 2018 demonstrated that the mid-level representation encoded in the intermediate layers of CNN consists of both texture and shape information, and such representation could be served as an organization principle in the ventral stream. Mid-level features can also be extracted in complex scenes by intermediate CNN layers, Bonner & Epstein, 2018 showed that the perceptual features extracted by intermediate CNN layers in scenes include information like cardinal orientations and boundary-defining junctions, which is important to the affordance properties of visual scenes.

Recently, Cichy and Kaiser (2019) argued that CNNs have the potential to help scientists generate new hypothesis and serve as a proof-of-principle demonstration of how perceptual and cognitive functions could be implemented in biologically plausible

computational models. In this study, I used CNN encoding models to model fMRI responses to a large number of images from image-computable perceptual features. This model serves two purposes. First, these feedforward models demonstrate the representations of category-selective cortex can be predicted from a small series of non-linear computations performed on image inputs. Second, through understanding the how these computational models react to different visual stimuli in a large-scale *in-silico* experiment, one can characterize the features that the model is sensitive to, thus leading to a better understanding of the tuning properties of these models.

3.2. Deep Convolutional Neural Network Encoding Models

CNNs have been shown to be powerful in predicting responses in the human and non-human primate visual system. A recent study by Schrimpf et al., 2018 revealed that the features extracted from AlexNet is remarkably similar to the neural representation along the ventral visual stream. In this analysis, I constructed a class of computational models called CNN encoding models that could relate neural representation to the intermediate AlexNet layer features. AlexNet was trained on a large object image dataset (~1M images) – the ImageNet dataset. These models can then be examined through *in-silico* experiments to understand their underlying tuning preferences.

3.2.1. BOLD5000

In order to build CNN encoding models, a large-scale fMRI dataset is needed. In particular, I used the BOLD5000 dataset to train the encoding models. Chang et al., 2019 collected slow event-related fMRI signal from four neurologically normal subjects (age: 24-27; 1 male; all right-handed) while they viewed images of scenes.

Each subject (except subject 4) underwent 143 experimental runs and 1 localizer run over 15 scanning sessions. Each session was 1.5 hours long. Subject 4 only finished 9 sessions out of the 15 fMRI sessions. Each participant also conducted an additional MRI scanning session to collect anatomical and diffusion imaging data. In each experimental run, 37 images were shown sequentially (375X375 pixels within 4.6 degrees of visual angle) for 1 second followed by 9 seconds fixation cross. When each image was shown, a valence judgement task was performed to indicate how much the participant liked the image by pressing “like”, “neutral” or “dislike”. In each localizer run, 60 images from each category of scene, object and scrambled image were used. The stimuli were presented in a block design format. Each block had 16 trials, with stimulus duration of 800ms and a 200ms ISI. Within the 16 trials, 14 unique images and 2 repeated images were shown, participants were asked to perform a one-back task. Between task blocks there were 6 seconds of fixation. There were 12 blocks per run, and 4 blocks per condition.

4916 unique images were selected as experimental stimuli. Images were drawn from three different computer vision datasets to represent image diversity across image categories (see Figure 3.2 for examples). In particular, 1000 indoor and outdoor scene images with over 250 categories were selected from the SUN dataset. Images were chosen to be scenic, depicting both outdoor and indoor scenes. 2000 images of multiple objects were chosen from the COCO dataset, with objects in a realistic context interacting with other objects. 1916 images with mostly singular objects were chosen from the ImageNet dataset. These images depicted a single object as the focus of the picture. Within these 4916 images, 112 images were shown 4 times and 1 image was shown three times across sessions, the remaining images were presented once to each

participant. In each experimental run, roughly 1/5 of Scene images, 2/5 of COCO images and 2/5 of ImageNet images were presented in a random order.

Figure 3.2 ***Sample images from the BOLD5000 dataset.***

Adapted from Chang et al., 2019. BOLD5000 dataset consists of experimental stimuli selected from three computer vision datasets.

All functional data were preprocessed by fMRIPrep (Esteban et al., 2019), where 3D motion correction, distortion correction and co-registration to the corresponding T1 anatomical image was performed. A general linear model with three conditions (scenes, objects and scrambled images) using a canonical hemodynamic response function was implemented in AFNI (R. W. Cox, 1996) for all the localizer runs. Scene-selective ROIs (PPA, OPA and RSC) were defined by using the contrast of scenes compared with objects together with an anatomical constraint, the top 200 voxels in each hemisphere that had the highest contrast within each anatomical ROI parcel were selected. The same procedure was used to define object-selective ROI (LOC) by using the contrast of objects compared with scrambled objects. Finally, early visual cortex (EVC) was defined using the same procedure with the contrast of scrambled objects compared with objects.

Experimental runs after preprocessing were modeled through a general linear model including a regressor for each trial compared with all other trials using the function 3DLSS (Mumford et al., 2012) to obtain an activation estimate for each trial. This way of modelling was shown to be more representative of the true activation magnitudes in event-related designs with lower signal to noise (Mumford et al., 2012).

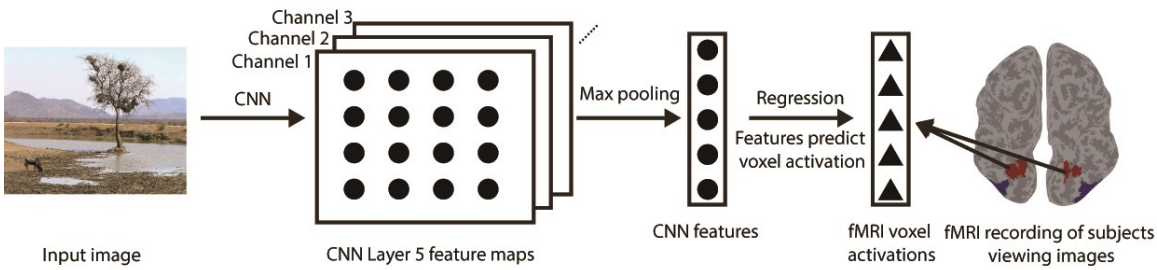
3.2.2. Encoding model architecture and training

Voxel-wise encoding models were built to model the category-selective area activations. The encoding models could be understood as a computational model to compute fMRI activation from perceptual image features extracted by the CNN. The model takes AlexNet intermediate layer feature maps of the image as input, then max-pools over the whole image for every convolutional kernel, which results in an AlexNet feature vector. Such max-pooling is helpful to prevent overfitting to the data while having a tradeoff of not preserving spatial information. This operation throws away all spatial information, so neural substrates which are highly sensitive to local position such as EVC would not perform well in this kind of model. AlexNet feature vectors can then be fitted through regression to the voxel-wise fMRI activation to learn the weights of connection between CNN features and fMRI activation (see Figure 3.3). All regressions had no bias term, which is necessary for regularized regression. In particular, I performed three regressions, including ordinary least square (OLS) regression, LASSO regression (L1 penalized regression) and ridge regression (L2 penalized regression) because adding regularization term was shown to be beneficial for models with collinearity between predictors (Tibshirani, 1996). For LASSO and ridge regression, a cross-validation is conducted to choose the penalty weight from the log scale space for each individual voxel to maximize performance. An independent model

was trained for each ROI with each subject using the same procedure. For the purpose of the follow-up *in silico* experiments, I will refer to these encoding models as a simulated model of the ROI it was fit to. For example, I will refer to the PPA encoding model as simPPA and the LOC encoding model as simLOC.

Figure 3.3 ***Illustration of the CNN encoding model architecture.***

The models took an image and generated model activations for a neural substrate. They were trained on BOLD5000 dataset using LASSO regression to model fMRI responses



3.3. CNN encoding model performance

To evaluate the performance of the CNN encoding models, I have conducted several validation experiments. These experiments and results are described in detail below, and I provide a brief overview here. First, I evaluated the regression methods used to fit the linear weights for simPPA and simLOC, and LASSO regression was shown to outperform other regression methods in modelling. Second, I used 10-fold cross-validation to assess the best layer of the CNN for explaining category-selective area fMRI activations. Layer 5 of AlexNet performed the best in this analysis, so the remaining analyzes were all based on the AlexNet layer 5 models. Third, I tested whether he encoding models demonstrated the classic category-selective responses in when shown a new set of images from an fMRI localizer experiment. Lastly, I

performed a strong test of generalization performance using a completely novel fMRI dataset with different images and different subjects. These generalization experiments were successful, suggesting that the trained encoding models were able to accurately predict activations to novel stimuli based on mid-level perceptual feature representations.

It is worth noting that when testing model performance, the accuracy of the models is bounded by the proportion of the variance of the fMRI data that is related to the stimuli, as opposed to noise or other unknown trial-specific or subject-specific effects (Lage-Castellanos et al., 2019). The bound on model performance has been referred to as the noise ceiling. In the following analysis, the noise ceiling of the dataset is calculated through measuring the across-subject reliability of the dataset. First, each participant's data is correlated with the mean data from the rest of the participants using the leave-one-out approach. The mean correlation of this leave-one-out procedure is the noise ceiling of the dataset.

3.3.1. Regression methods comparison

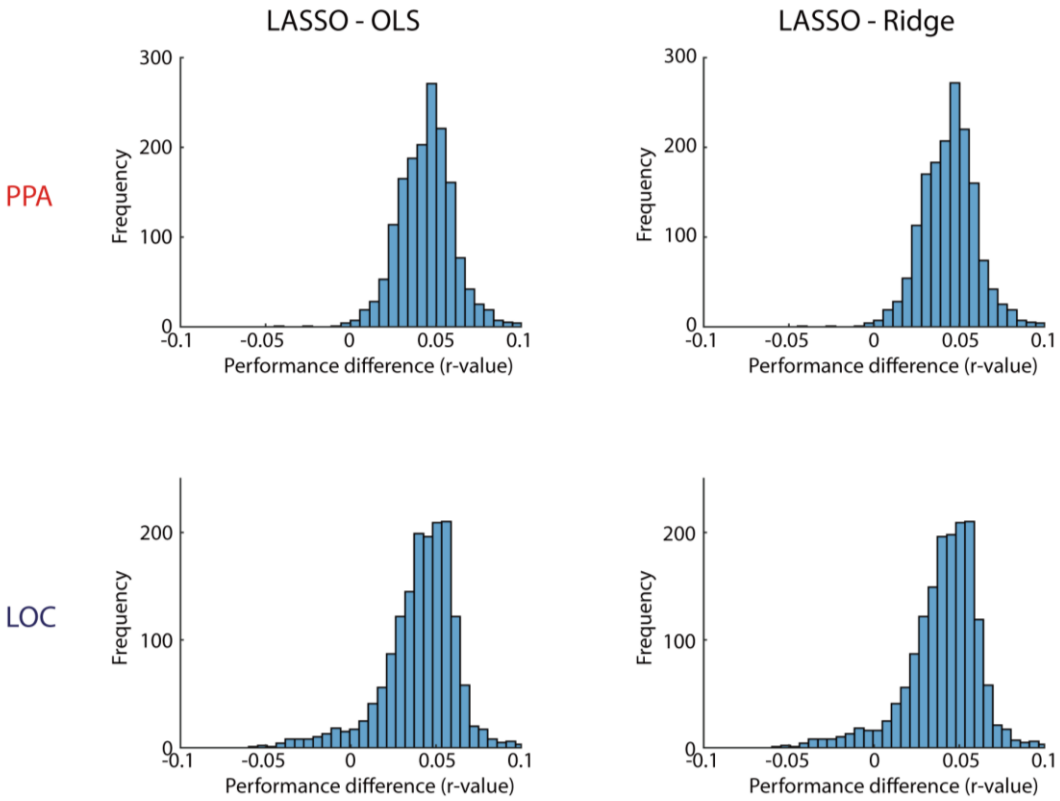
I am interested in L1 regularization as a potential means of learning sparse encoding models that emphasize the CNN features that are most important for each ROI. I evaluated the performance of different regression methods by running encoding-model analyses on the BOLD5000 dataset with 10-fold cross-validation using ordinary least squares (OLS) regression (without regularization), LASSO regression (L1 regularization) and ridge regression (L2 regularization). For LASSO and ridge regression, a separate 10-fold cross-validation was performed before assessing performance to determine the best penalty parameters. Because the penalty parameters for LASSO and ridge are learned on the same data that we use for quantifying model

performance on the BOLD5000 dataset (using a different cross-validation design), the performance estimates for the regularized models may be slightly biased upwards. However, this is not problematic for my follow-up analyses for three reasons. The first reason is that the encoding models perform well even when using OLS regression without regularization, which means that regularization is not required to achieve statistically significant performance. The second reason is that the results and conclusions and conclusions I will discuss do not depend on the specific values of the performance estimates. The third reason is that these models were shown to have good prediction accuracy when predicting responses to a completely different dataset of novel images and novel subjects—thus, any concerns that these models are overfit to noise in the BOLD5000 dataset are mitigated by this strong test of generalization performance.

LASSO regression had the best performance (10-fold cross-validation within BOLD5000) in both scene-selective and object-selective areas (see Figure 3.4). LASSO regression performs both feature selection and regression in one model and forces the weights of potentially irrelevant features to zero; therefore, this result suggested there were irrelevant features in the CNN to the neural representation and regularized regression helped the encoding model training to prevent overfitting. In the following, we performed follow-up analyses using models fit with LASSO regression.

Figure 3.4 ***Distribution of prediction score differences between different regression methods.***

The left figure shows the distribution of prediction score differences between LASSO and ridge regression. Most voxels show a higher prediction score for LASSO regression. The right figure shows the distribution of prediction score differences between LASSO and ordinary least squares (OLS) regression. LASSO regression has a better performance for most voxels.



3.3.2. CNN layer performance comparisons

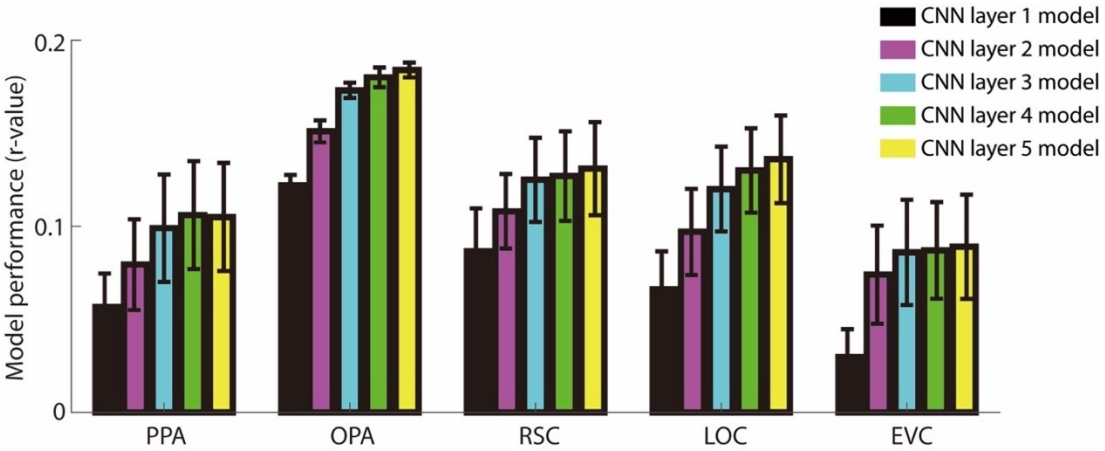
The performance of the encoding model is assessed through the Pearson correlation between the model activation and the actual fMRI activation recorded. A 10-fold cross validation was used to examine the performance of the model in the BOLD5000 data. Figure 3.5 shows the performance of the CNN encoding models using different layer feature maps. Our results align with previous findings (Khaligh-Razavi & Kriegeskorte, 2014) that among convolutional layers, activation of layer 5 best

predicted fMRI representation of mid-level visual cortices. On the other hand, simEVC showed a different pattern that deeper layer did not perform better in explaining neural representation. This could be attributed to the fact that EVC is sensitive to location information, however, the max-pooling operation in the CNN encoding model discards spatial information, which likely dampens the performance of simEVC. In all the analysis below, I used the encoding models built using AlexNet layer 5.

Both scene-selective and object-selective areas achieved reasonable performance given the noise ceiling of the dataset (i.e., which is likely due to the lack of stimulus repetitions in the dataset). Indeed, the CNN encoding model performance exceeds the noise ceilings in all ROIs (PPA: 0.04, OPA: 0.08, RSC: 0.04, LOC: 0.05, EVC: 0.04) suggesting our models were able to capture all explainable variance in the dataset.

Figure 3.5 ***CNN encoding model cross-validation performance.***

10-fold cross validation on AlexNet encoding model performance on BOLD5000 dataset for all ROI. Performance is quantified using the Pearson correlation between the model and actual fMRI activations. Results indicate layer 5 of AlexNet has the best encoding performance for all ROIs. Error bars indicate +/-1 SD across subjects.



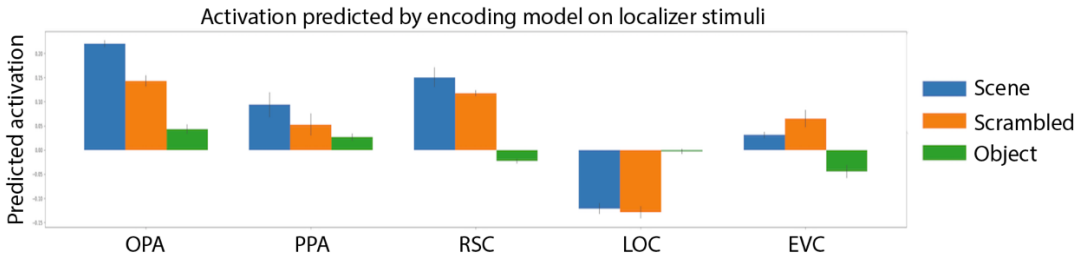
3.3.3. Generalizability of CNN encoding models

In order to test whether these models could generalize model activation to novel images, localizer images were passed to the CNN encoding models to examine whether its activation matches with expectation. In the first analysis, localizer images from the BOLD5000 dataset were processed through the CNN encoding model and the model activation was shown in Figure 3.6. For all scene-selective areas, scene images produced a higher model activation than scrambled and object images. The simPPA model showed the classic scene-selective response profiles that is the defining characteristic of the actual PPA. Conversely, simLOC showed a classic object-selective preference that is used to define the actual LOC. Lastly, simEVC showed a higher activation to scrambled objects than other images which is exactly how we define EVC in fMRI data. These results suggested that the CNN encoding models were able to

generalize its activation to novel images that were not included in the BOLD5000 dataset.

Figure 3.6 ***CNN encoding model activation on localizer stimuli.***

Model activations were averaged across voxels and subjects within an ROI. Error bars indicate +/- 2 SD of the mean activations.



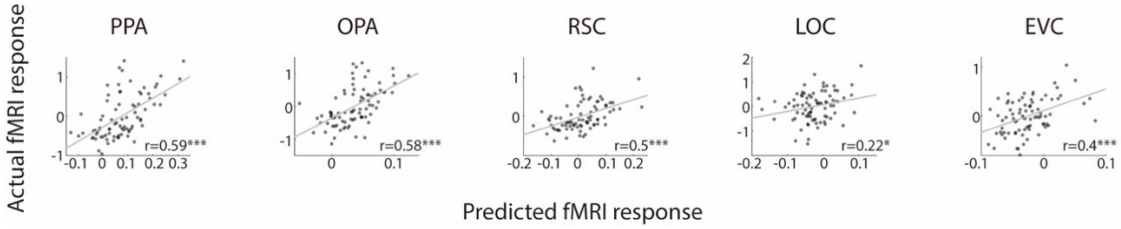
In the second analysis, BOLD5000 trained models were used to predict fMRI activation in the object2vec dataset (Bonner & Epstein, 2020), which has a different set of images and different subjects (see Figure 3.7 for example images). I ran the CNN encoding models on the 810 images across 81 object categories used in the object2vec dataset, then the model activations were averaged over each category and voxels in each ROI. Unlike the BOLD5000 dataset, object2vec used a block-design, which was more reliable and had a higher noise ceiling. The object2vec activations of each object category were averaged across subjects and across voxels within each ROI. The observed activations and the model activations were highly correlated (see Figure 3.8) and approaching the between-subjects noise ceiling (PPA: 0.75, OPA: 0.76, RSC: 0.51, LOC: 0.77, EVC: 0.45). These results suggest that the BOLD5000 trained model can generalize to completely different subjects and stimuli.

Figure 3.7 ***Example images from the object2vec dataset.***

Adapted from Bonner & Epstein, 2020. This dataset contains 810 object images across 81 different object categories.

Figure 3.8 ***CNN encoding model performance on the object2vec dataset.***

*Across-subject validation using CNN encoding models trained on BOLD5000 to predict activation for different groups of subjects with different stimuli in the object2vec dataset. Significant correlations between the model responses and the observed fMRI responses indicate that the CNN encoding models were able to generalize to novel subjects and novel images. * indicates p<0.05, *** indicates p<0.001*



To conclude, these findings discussed in this chapter demonstrate that feedforward CNN encoding models can reliably predict fMRI activations in the category-selective areas through a simple linear re-weighting of mid-level perceptual features computed from image inputs. Understanding the selectivity profiles of these models using a large image dataset can potentially provide insight into the nature of perceptual and semantic representations. In the following chapters, I will introduce a series of *in-silico* experiments that use the encoding models to study the semantic selectivity of different ROIs.

CHAPTER 4. UNDERSTANDING MID-LEVEL TUNING PROFILES THROUGH SEMANTIC-PREFERENCE MAPPING

In Chapter 3, I built image-computable models of the category-selective cortex, which were shown to explain a high amount of variance even in a novel dataset. In this chapter, I will discuss experiments that characterize the semantic selectivity of computational models for these category-selective areas.

First, I will discuss an existing technique –network dissection– for characterizing the selectivity profiles of computational models. Second, I will introduce a new computational technique called semantic preference mapping to characterize the selectivity profiles of CNNs. Third, I will discuss a human behavioral experiment that is used to collect object property ratings. Lastly, using both semantic preference mapping and the object property ratings, I will discuss insights into the tuning profiles of computational models for category-selective cortex.

4.1. Characterizing the selectivity profiles using network dissection

Current CNNs yield surprisingly good performance on predicting the neural representations of visual cortex (Schrimpf et al., 2018). However, the internal representations of CNNs are difficult to interpret, given the many nonlinear operations in a CNN; therefore, scientists need some method to help characterize the internal representations (Montavon et al., 2018). One prominent method to characterize the internal representation of computational model is called network dissection.

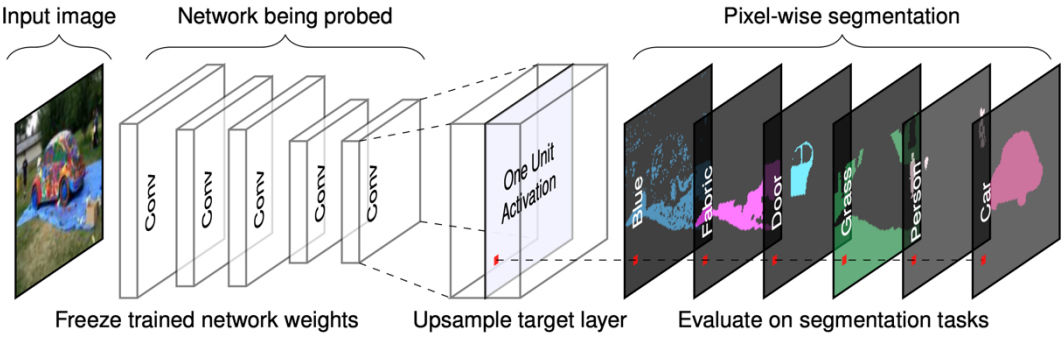
In any given image, it contains hundreds of thousands of pixels, and the CNN activates based on these hundreds of thousands of pixels. However, not every pixel contributes the same to the activation of a given unit in the CNN model. A unit may only be sensitive to a limited portion of the image and only to a particular perceptual

pattern in that portion of the image. Nevertheless, given the number of parameters in the model, it is hard to characterize the selectivity profiles of a given unit just from inspecting the parameters. The goal of network dissection is to solve this problem by characterizing the visual inputs that cause a unit to activate. The advantage of using this method is that this can be applied to any unit in any image-computable CNN model (Bau et al., 2017).

To perform network dissection, the model is first fed with some input images and the activation of a target unit from a particular layer will be recorded. The image that maximizes the unit activation will be discovered through this process. Second, since the dimension of the target layer may not be the same as the input image, the target layer activation will be scaled up into the original input space to allow proper visualization. Third, the upsample target activation will then be segmented to show only regions corresponding to the highest activation of the target unit. This segmentation mask indicates the visual region that has a high activation of a CNN unit. When the segmentation mask is applied to the input image, human labeling can characterize the corresponding property encoded in the unit. See figure 4.1 for a pictorial description of the algorithm.

Figure 4.1 ***Network dissection illustration from Zhou et al. (2018).***

In this example, one unit of convolutional layer 5 in a given CNN is probed by network dissection to evaluate its match on various segmentation maps.



This method has shed light on the interpretation of CNN representation. Given the high correspondence between CNN and the visual cortex, cognitive neuroscientists have used this method to understand the representations of high-level visual cortex. For example, Bonner & Epstein, 2018 developed a receptive field mapping technique, which is similar to network dissection, to visualize units in a CNN that show a high correspondence to OPA voxels. The goal of this analysis was to find regions of an image that CNN units are sensitive to. This method discovered that the CNN units that best matched the OPA representations responded most strongly to image regions containing boundary-defining junctions and large extended surfaces.

Network dissection uses the segmentation masks created by a CNN unit and then performs post-hoc interpretations of these segmentation masks. However, if one is specifically interested in understanding selectivity to object classes (or any other scene element), then it is possible to directly assess this by leveraging existing segmented image databases and performed targeted semantic occlusions. This is the approach used in the semantic preference mapping procedure.

4.2. Semantic preference mapping

Semantic preference mapping examines how the activations of a CNN encoding model are affected by the object categories present in an image by systematically occluding a specific object category from each image in a large set of samples of natural scene images. If the model was sensitive to a specific object category, then the model activations should decrease as a result of occlusion of that object category. Using this logic, we compared the activations between the original image and the occluded image, and the resulting difference ($act_{original}$ - $act_{occluded}$) was recorded for each pair of images. This procedure was then repeated in a large number of images for each object category, and the averaged difference across images for each object category was assigned as the selectivity index of the object category.

4.2.1. The ADE20K dataset

This analysis used a separate image dataset from the one used in the BOLD5000 fMRI experiment. I specifically used the ADE20K dataset (Bolei Zhou et al., 2017), which did not intersect with images used in the BOLD5000 dataset. The ADE20K dataset consisted of more than 22,000 natural images with fully annotated object segmentation maps, which made it possible to perform targeted occlusions of specific object categories in each image. Figure 4.2 shows example images from the ADE20K dataset. The use of ADE20K also allowed me to examine the semantic selectivity of the CNN encoding models in the context of a large and diverse sample of natural images. This is important because it ensures that the semantic-selectivity findings are broadly representative of natural image statistics rather than being an idiosyncratic confound of the fMRI stimulus set.

Figure 4.2 ***Example images from the ADE20K dataset.***

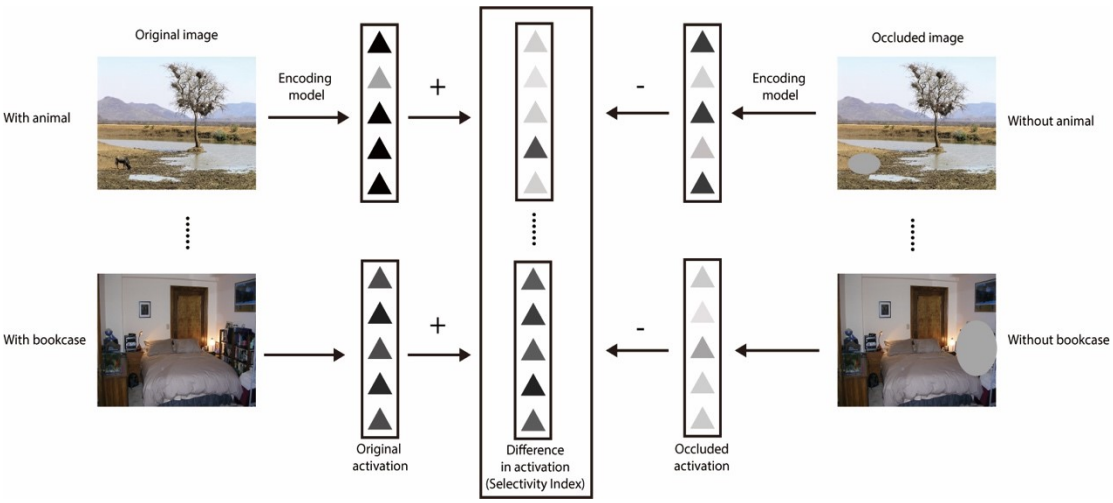
The first row is the original images, the second row indicates the object segmentation map.

I first chose object categories with more than 500 instances in the ADE20K dataset (Bolei Zhou et al., 2017), which resulted in 85 categories. In the following analysis, I focused on understanding the selectivity for these 85 object categories. For each object category, we examined all images that contained that object. I then used the segmentation mask to locate the object(s) and created the smallest oval mask(s) that covered the target object(s). The pixels in the oval mask were assigned random RGB values. The oval occluder was the minimum possible size that fully occluded the object (i.e., the occluder covered the entire object segmentation mask), and the edges of the occluder were smoothed to avoid adding high-frequency noise to the image. An oval shape was used, rather than the object segmentation mask itself, to avoid including shape information in the mask; therefore, no information from the occluded objects remained in the occluded image. Both the original image and the occluded image were fed into the CNN encoding models to generate activations. For every CNN unit, we subtracted the activation to the occluded image from the activation to the original image to obtain the selectivity index of the particular occluded object. After repeating this procedure for all images containing the target object categories, we calculated the mean selectivity index across images, which captures the degree to which the responses of

the unit are sensitive to the target object category. We repeated this procedure for all 85 object categories.

Figure 4.3 ***Semantic preference mapping procedure.***

Model activations to an image with an occluded object are compared with the model activations to the original image for all instances of an object categories in the ADE20K dataset to produce the selectivity index.



4.2.2. Selectivity indices

Using the semantic-preference mapping approach, I characterized the selectivity profiles of the CNN encoding models of the category-selective areas. Given the similar observation across scene-selective models, I will focus the discussion on the simPPA model, which is the most representative scene-selective model. simPPA showed a high selectivity index for skyscrapers, houses and bookcases and a low index for animals and balls. On the contrary, simLOC showed an opposite pattern, where it showed a high selectivity index to balls and animals and a low selectivity index to skyscrapers, houses and bookcases (see Figure 4.4). We ranked the occlusion indices of all 85 categories of objects and observed that many of the top-ranked object

categories in simPPA tended to be more rectilinear (e.g., skyscraper, house, bookcase) and large in size, while the top-ranked object categories in simLOC tended to be curvy and small in size. This suggested the possibility that curvilinearity and real-world size could be important latent dimensions in the selectivity profiles underlying the category-selective areas. An alternative possibility was that these results simply reflected the occluder size (i.e., simPPA was selective for objects that are larger in the image and thus require larger occluders, and simLOC was selective for objects that are small and thus require smaller occluders). However, any potential effects of occluder size were likely minimized by our use of global max-pooling, which discards spatial information from each feature channel. Furthermore, we performed analyses to specifically address this possibility. In the following analyzes, occluder size (i.e., the number of pixels in the occluder) was fully regressed out from the selectivity indices, so that any observed effect could not be explained by occluder size. We systematically explored the factors that relate to these semantic-selectivity profiles in the follow-up analyses.

Figure 4.4 ***Ranking of the selectivity indices. Representative categories are shown here for demonstrative purpose.***

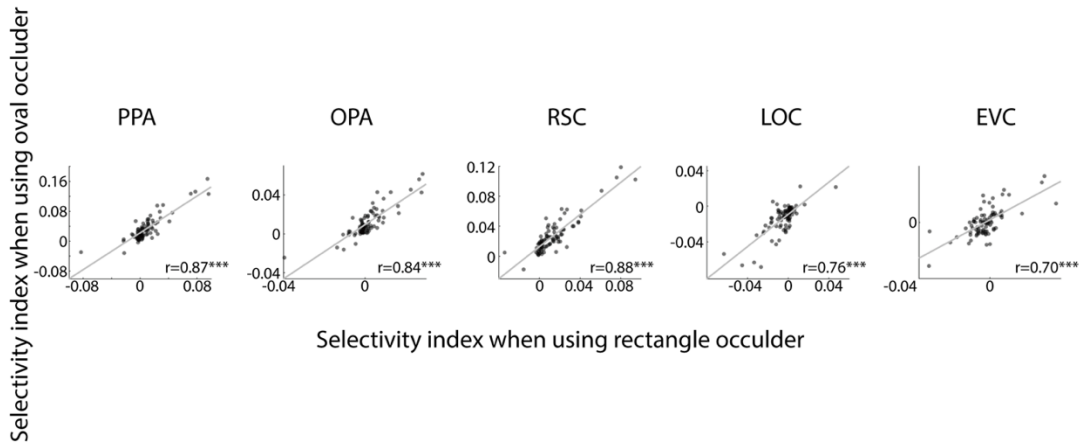
Full results are shown in the Appendix A1. simPPA demonstrated a preference towards fixed, large objects. simLOC showed a preference towards animate, small objects.

4.2.3. Validation of the semantic preference mapping results

These results showed high between-subject correlations of the selectivity profiles in these areas and this suggested that the tuning profiles discovered were consistent across subjects (mean across-subject correlations of selectivity profiles of simPPA: 0.97 and simLOC: 0.9). To further validate the semantic preference mapping procedure, I conducted four separate experiments. First, to ensure that the shape of the occluders did not influence the selectivity indices, I repeated the semantic preference mapping procedure using a rectangular occluder instead of an oval occluder. Instead of creating the smallest oval mask(s) over the target objects, I used the smallest rectangular mask. As shown in Figure 4.5 the results using the rectangular occluders were highly correlated with the results from the oval occluders (all $r>0.7$, $p<0.0001$).

Figure 4.5 ***Validation that the semantic preference mapping results were not highly sensitive to the occluder shape.***

*Semantic preference mapping was conducted using both oval occluders and rectangular occluders. These scatter plots show that the selectivity indices generated from the oval occluders and rectangular occluders were highly correlated. *** indicates p<0.001*



Second, I lowered the requirement that each object category must have at least 500 instances in the dataset in order to include a larger number of object categories and determined how this affected the results. In particular, I repeated the semantic preference mapping procedure in the ADE20K dataset with more object categories (155 object categories with at least 200 instances in the ADE20K dataset). As a way examining whether our findings diverged when using 155 categories instead of 85 categories, I performed the semantic preference mapping procedure and compared how the selectivity indices correlated with the image-computable model of curvature summary statistics (described in Chapter 5.2). This allowed us to use an automated procedure to characterize the results from both versions of the semantic-preference mapping experiment. In both versions, I observed the same pattern of curvilinearity preferences: namely, that simPPA preferred boxy objects and simLOC preferred curvy objects. In simPPA, selectivity indices were negatively correlated with the curvature

indices in both the 85-categories and 155-categories versions of the procedure, suggesting selectivity to rectilinear objects (85 categories: $r=-0.6$, $p<0.0001$; 155 categories: $r=-0.45$, $p<0.0001$). In simLOC, the selectivity indices were positively correlated with the curvature preferences (85 categories: $r=0.55$, $p<0.0001$; 155 categories: $r=0.46$, $p<0.0001$), suggesting selectivity to curvy objects. This result suggests that our findings are not highly contingent on the parameter that determines the number of object categories examined.

Third, I examined if the semantic-preference mapping procedure was sensitive to the random initialization of the CNN parameters. To do this, I used 10 AlexNets with different randomization parameters to train the fMRI encoding models, and then performed the same semantic occlusion procedure. We obtained similar results from different randomizations suggesting that the results were robust to different initializations of parameters. Specifically, we adapted the AlexNets published by (Mehrer et al., 2020), which included 10 different AlexNets trained on the CIFAR dataset using different initial randomizations. We trained LASSO encoding models with these AlexNet layer 5 activations and performed univariate semantic preference mapping using the same procedure described above. Pairwise correlations of selectivity indices between different randomizations were obtained (mean correlation >0.99 across simPPA and simLOC). These correlations suggested that the parameter initialization of the CNN did not have an effect on the results of the semantic preference mapping procedure.

Lastly, I investigated whether the resulting selectivity indices were largely dependent on the CNN architecture and the CNN training set by repeating the procedure with different CNNs. I repeated the same LASSO encoding model training and univariate semantic preference mapping procedure on three different CNNs, including

AlexNet trained on the Places365 dataset (B. Zhou et al., 2018), Resnet 18 trained on the ImageNet dataset (He et al., 2015) and Resnet 18 trained on the Places365 dataset (B. Zhou et al., 2018). For AlexNet trained on Places365, I used the layer 5 activations to feed into the encoding models, and for both Resnet 18 models, I took the output activations of the fourth block of convolution layers to feed into the encoding models. In the table below, the correlation of the selectivity indices using and the other CNNs are reported. For both simPPA and simLOC, the results showed a robust result between CNN architecture and the training image set. This result indicated that the selectivity indices obtained from the semantic-preference mapping procedure were robust to variations in CNN architectures and training sets. Taken together, these results suggest that semantic-preference mapping is a robust and reliable procedure for examining how CNN activations are affected by the presence of specific object categories in images.

Table 4.1 Correlation of selectivity indices between AlexNet trained on ImageNet and other CNNs.

** indicates p<0.05, *** indicates p<0.001*

CNN	simPPA	simLOC	simEVC
AlexNet trained on Places 365	0.64***	0.22*	0.1
Reset-18 trained on ImageNet	0.54***	0.5***	-0.33
Reset-18 trained on Places 365	0.72***	0.65***	0.15

This section demonstrated that semantic preference mapping is a powerful *in-silico* experiment that examines the selectivity profiles using a large dataset. This procedure can be widely applied to many types of encoding model, and is not limited to CNN encoding models. In addition, the semantic preference mapping procedure can be generalized to understand other visual features such as color and texture by occluding a particular color or occluding a particular texture.

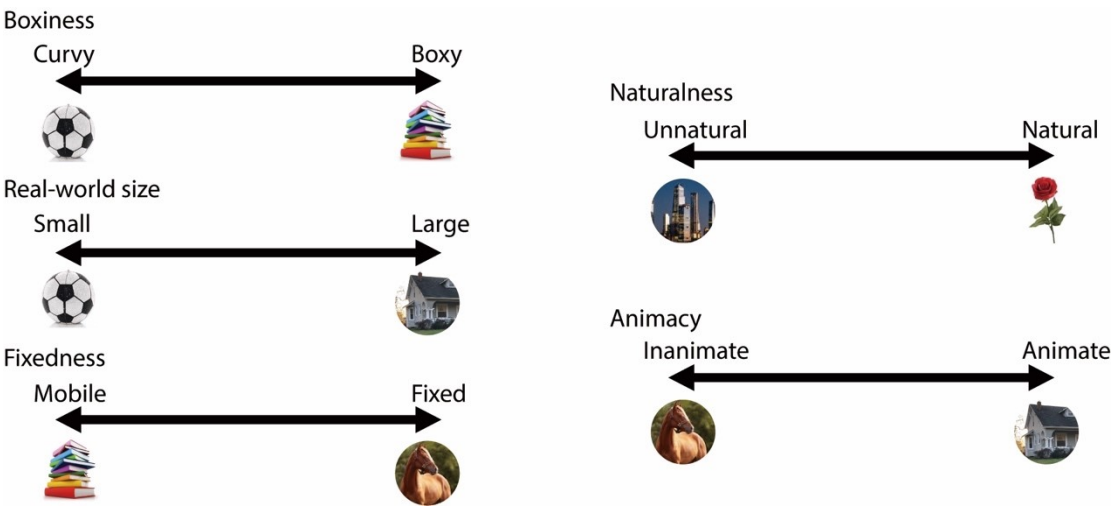
Precisely characterizing tuning preferences has been one of the hurdles in understanding the representation in the category-selective cortex. While the category-selective areas contain complex representation, a good computational model of these areas can help better understand the nature of the representation. In the following sections, I will demonstrate using the tuning profiles from the semantic preference mapping analysis to address outstanding questions on the nature of the mid-level representation. I will show that the category-selective area models capture selectivity to some high-level semantic properties of objects, suggesting that selectivity to these object properties could be an emergent phenomenon of the bottom-up feedforward computations of mid-level visual features.

4.3. Object property ratings

To systematically investigate the object properties that correlated with the selectivity indices, we collected human ratings for the 85 object categories used in the semantic preference mapping on different object properties, including curvature, real-world size, fixedness, animacy and naturalness, which have all been previously shown to be important organizational principles of the representations in category-selective areas (Konkle & Oliva, 2012; B. Long et al., 2018; Long Sha et al., 2015; Troiani et al., 2014; Yue et al., 2020), as discussed in Chapter 2 (see Figure 4.6). For the curvature

rating, I asked whether the highlighted object(s) have straight lines and sharp corners or curvy and rounded contours. This rating captured the perceptual curvilinear information of the target objects. For the real-world size rating, subjects were asked to judge how big the highlighted object is in the real-world. Fixedness is a property that measures how fixed an object is in the environment—an object that is easily transportable would not be considered as fixed. The fixedness rating gets at how spatially fixed an object is, so the question asked how often you would expect the highlighted object to change position. For the animacy rating, we adopted a continuum dimension to ask the subject to judge how "alive" the highlighted object is (Long Sha et al., 2015). The naturalness ratings asked subjects to judge whether the highlighted objects look manmade or neutral.

Figure 4.6 ***Object properties that were previously shown to be important dimensions in the ventral stream.***



Fifty subjects were recruited online through the Prolific platform. This online experiment was in compliance with the procedures approved by the Johns Hopkins University Institutional Review Board.

Each subject was asked to judge five object properties of the highlighted object(s) in the image using a 7-point scale. The judged object properties included curvature, real-world size, animacy, naturalness and fixedness. After each stimulus, there was a 300ms inter-trial-interval before the start of the next trial. Each subject was presented with one image per each of the 85 object categories.

Stimuli were randomly chosen from the images used in the semantic preference mapping procedure. 50 images were drawn from the each of the 85 object categories. The target object is highlighted by an opaque mask of the background (see Figure 4.7). The experiment also included a magnifying glass, which subjects could freely move using the mouse to enlarge any part of the image.

The distributions of the object property ratings are reported in Figure 4.8. To evaluate the relationship between different object properties, I took the mean of each rating across all object instances within an object category. In Figure 4.9, the correlation between each pair of object properties is reported. Because animacy was highly correlated with naturalness ($r=0.91$, $p<0.001$), we merged the animacy and naturalness ratings as a single dimension by averaging the two ratings.

Figure 4.7 ***Human object property rating procedure.***

This shows an example of the webpage interface used for the online human object property rating experiment. The target object was highlighted with a red oval, and the background context was made transparent. The magnifying glass could be moved around to enlarge the image. Subjects were then to asked to judge multiple properties of the highlighted objects.

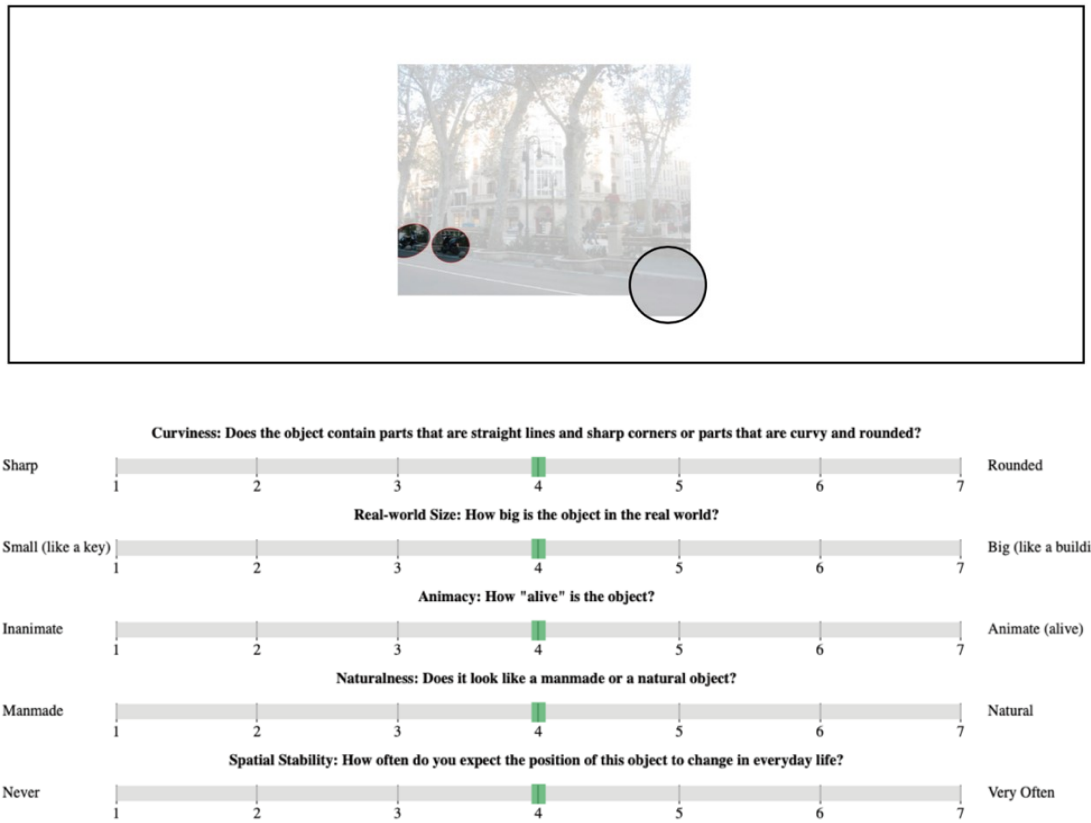


Figure 4.8 ***Distributions of object property ratings.***

Each histogram shows the distribution of an object property rating.

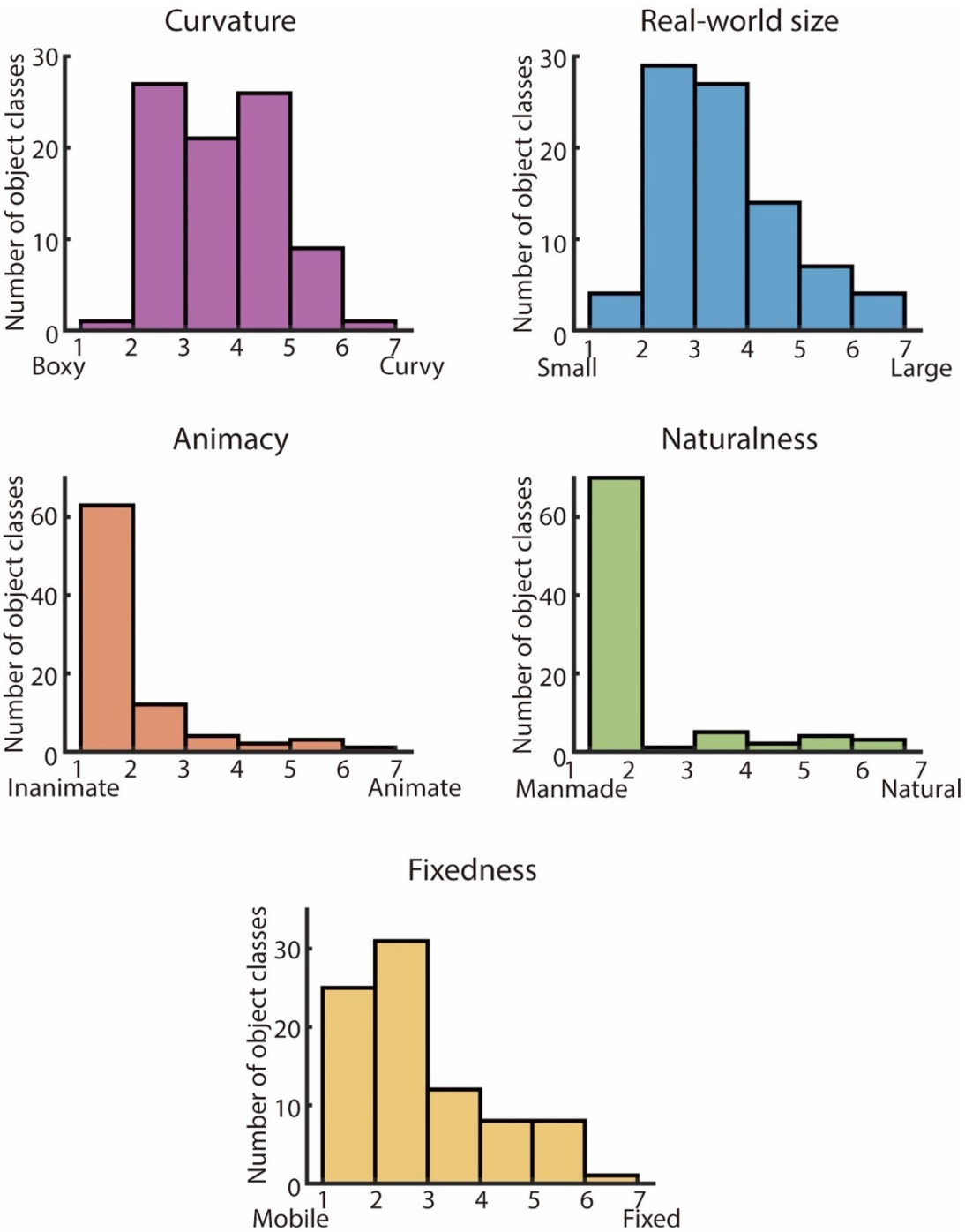
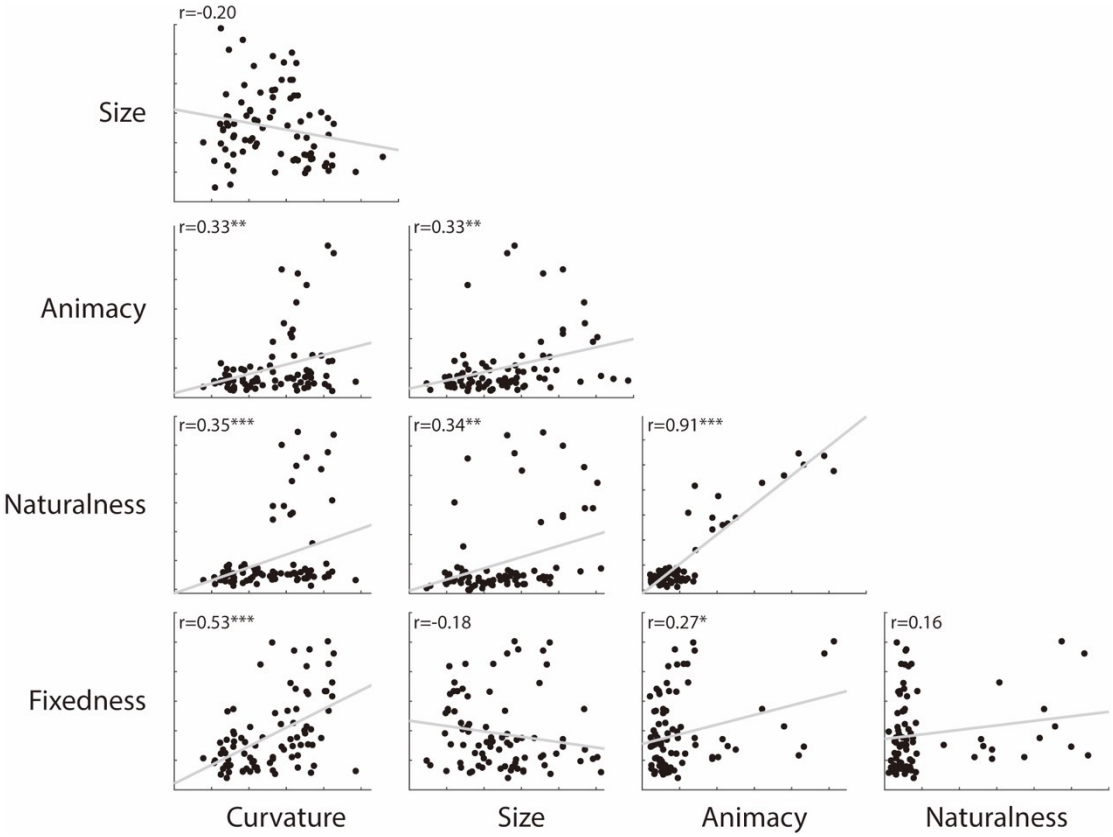


Figure 4.9 ***Covariance of object property ratings.***

*Each scatter plot shows the correlation between two object properties collected by the human rating experiment. This result indicated that many object properties covaried. * indicates p<0.05, ** indicates p<0.01, *** indicates p<0.001.*



4.4. Univariate selectivity index analysis

To investigate whether the selectivity indices are tuned to different object categories, a correlation analysis was performed. Here, I focus the discussion on the scene-selective simPPA and object-selective simLOC. Results from other ROIs can be found in Figure 4.11. The pattern of results observed from this analysis suggested the simulated models exhibit a similar high-level semantic tuning preference as previously reported in fMRI experiments of these category-selective areas. Specifically, simPPA was selective to boxy, large, fixed and inanimate/unnatural objects. These properties, which are highly correlated with landmark object features, have also been observed in

the actual PPA tuning profiles (I. I. A. Groen et al., 2017; Nasr et al., 2014; Troiani et al., 2014; Yue et al., 2020). On the other hand, simLOC was selective to curvy, small and mobile objects, which has also been observed in LOC fMRI responses from previous studies (I. I. A. Groen et al., 2017; Konkle & Oliva, 2012; B. Long et al., 2018; Long Sha et al., 2015).

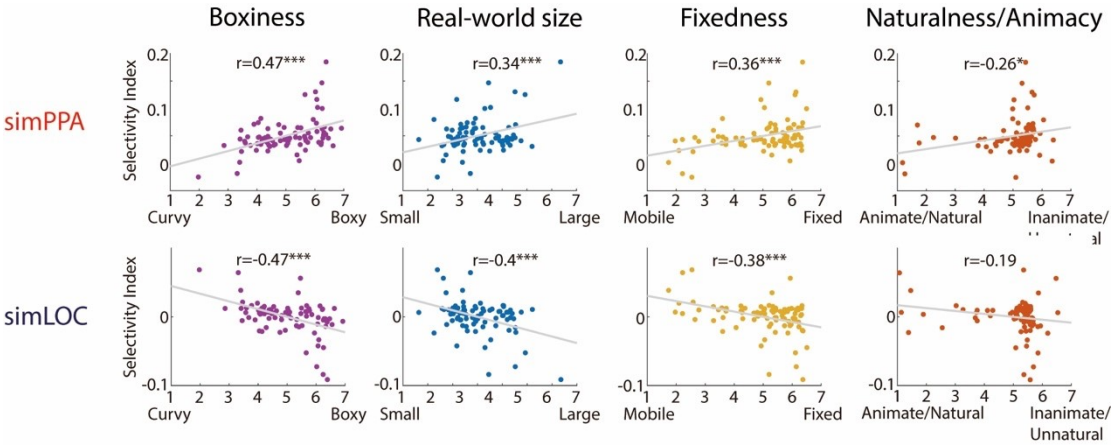
One surprising finding from this analysis was that selectivity to high-level semantic features were able to emerge from a model of mid-level perceptual feature representations. In particular, this result suggests that abstract semantic properties including real-world size, fixedness, naturalness and animacy could emerge through bottom-up feedforward computations in the CNN encoding models.

Rectilinearity/curvature has been proposed to be an important mid-level property of objects along the ventral visual stream (El-Shamayleh & Pasupathy, 2016; B. Long et al., 2018; Yue et al., 2020). Curvilinear preference tuning was speculated to be an important mid-level dimension in inferring high-level semantic properties of objects (B. Long et al., 2018). In Figure 4.9, I confirmed this speculation that the curvilinear property of objects correlated with high-level properties like animacy and fixedness of objects. This further suggests that the mid-level tuning preferences observed in these areas could directly support their hypothesized role in representing the high-level semantic properties of visual stimuli.

Real-world size has also been argued to be a feature that organizes object responses in the occipitotemporal cortex (Coutanche & Koch, 2018; Konkle & Oliva, 2012). In this analysis, real-world size was shown to be an important object property in both object- and scene-selective areas. However, one speculation suggested that the real-world size preferences could be explained by lower-level preferences for curvature

and shape (B. Long et al., 2018). To test this hypothesis, in the next analysis, I focused on understanding the unique contribution of each object property.

Figure 4.10 ***Scatter plots showing the correlation between different object properties and selectivity indices in simPPA and simLOC after regressing out occluder size.***



To understand the unique contribution of each object property, I ran a partial correlation analysis for each property after regressing out the contribution of all other object properties (see Figure 4.11). In simPPA, only curvature, real-world size and animacy/naturalness showed significant partial correlations with the selectivity indices. In simLOC, only curvature and real-world size remained as the significant unique dimensions to contribute to the selectivity indices.

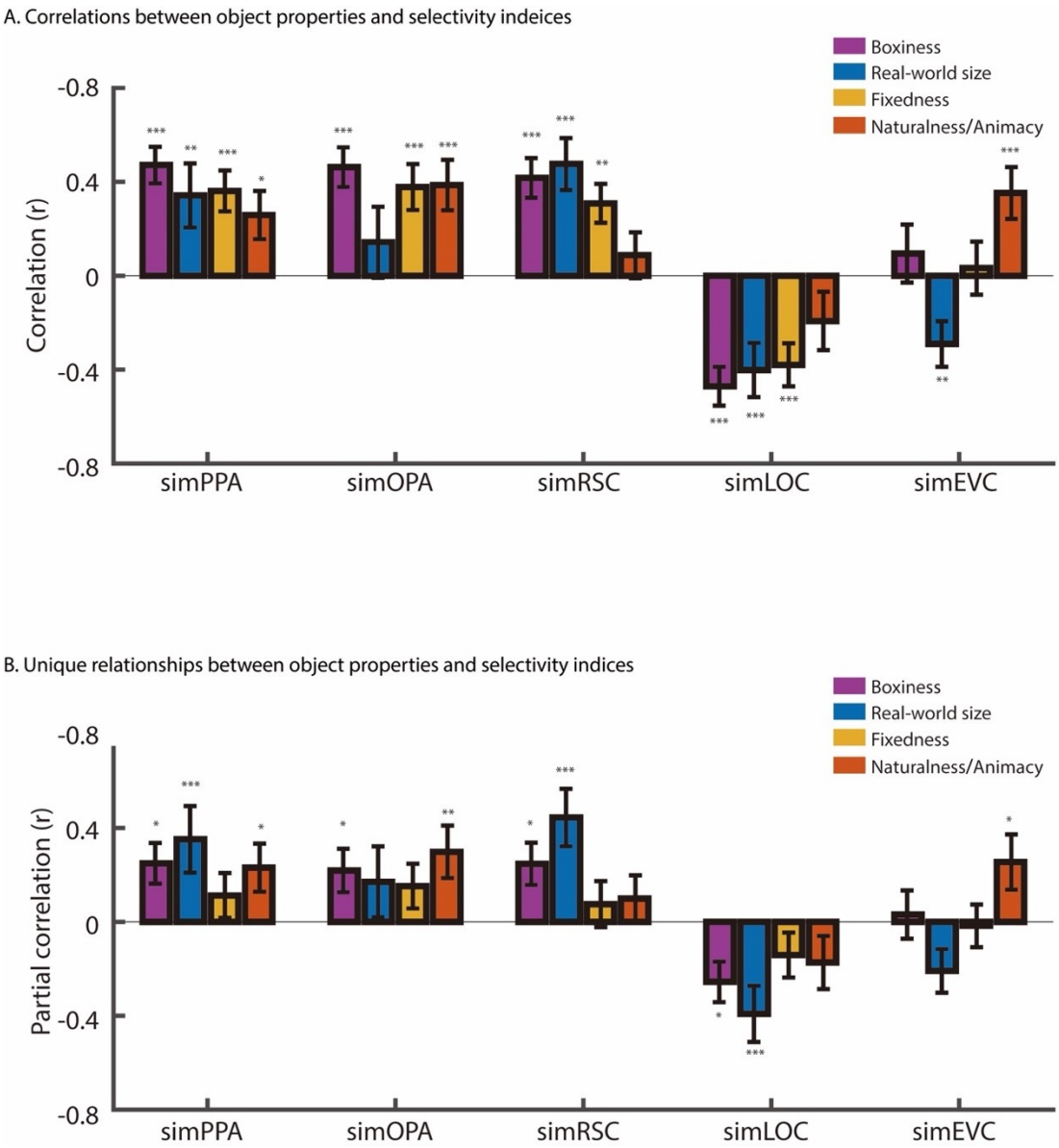
Previous studies had speculated that cortical preference to real-world size could be reduced to curvature encoding (B. Long et al., 2018). In this analysis, the unique contribution of real-world size when curvature is regressed out, demonstrate that preferences for real-world size cannot be solely explained by selectivity to curvature alone.

We found that fixedness did not have a unique contribution to the univariate selectivity indices of our models. Fixedness was shown to be an important object

property in previous studies of the PPA (R. A. Epstein & Baker, 2019; Troiani et al., 2014). One speculation from our analysis is that fixedness could be explained by other object properties such as animacy, since the animacy dimension is correlated with fixedness as shown in Figure 4.9.

The univariate analysis revealed the information that was encoded in the mean overall response in each of the simulated models. However, characterizing the representation this way may have lost some of the important dimensions represented in the multivariate activation patterns. Therefore, the next analysis focused on determining whether there was additional information contained in the multivariate selectivity index pattern in each of the category-selective area models.

Figure 4.11 ***Correlation between object properties and univariate selectivity indices.*** *Error bars indicate +/-1 SD using bootstrapping (N=10,000). * indicates p<0.05, ** indicates p<0.01, *** indicates p<0.001. A: Pearson correlation between object properties and univariate selectivity indices while regressing out occluder size is shown for all ROIs. B: Unique contribution (partial correlation) between object properties and univariate selectivity indices are shown for all ROIs while accounting for the covariance of different object properties.*



4.5. Multivariate selectivity index analysis

In fMRI studies, MVPA is a powerful tool to explore representational dimensions that requires multiple voxels to represent the information (Haxby, 2012, p.; Norman et al., 2006). We aimed at developing a selectivity index pattern analysis that could reveal multivariate encoding dimensions. Instead of averaging the selectivity indices across units in a model, the principal component analysis (PCA) took the selectivity index pattern and found the principal components of the selectivity patterns. In simPPA, the first PC already accounted for 82% variance, and the second PC accounted for 7% variance, and other scene-selective models showed a similar result (1^{st} PC – simOPA: 87%; simRSC: 93%/2^{nd} PC – simOPA: 8%; simRSC: 2). In simLOC, the first PC accounted for 71% of the variance while the second PC accounted for 16% of the variance. Similar to the univariate analysis, we correlated the PC scores with different object properties. The first PC revealed similar findings as the univariate analysis, where curvature and real-world size were the most important dimensions in simPPA and all four properties were important to simLOC. Moreover, the second PC revealed that real-world size and animacy/naturalness were important dimensions in both simPPA and simLOC (see Figure 4.12 panel A).

The partial correlation analysis was similar to the univariate partial correlation analysis while the PC scores were used instead of the univariate selectivity indices. The results (see Figure 4.12 panel B) from the first PC illustrated the same pattern as the univariate analysis in simPPA, which suggested there was one representational dimension that represented curvature and real-world size while explaining the most variance. Interestingly, this first PC of simPPA reflects a preference to large, manmade objects, while the second PC of simPPA reflects a preference towards large and natural objects. In simLOC, the pattern suggested that apart from curvature and real-world size,

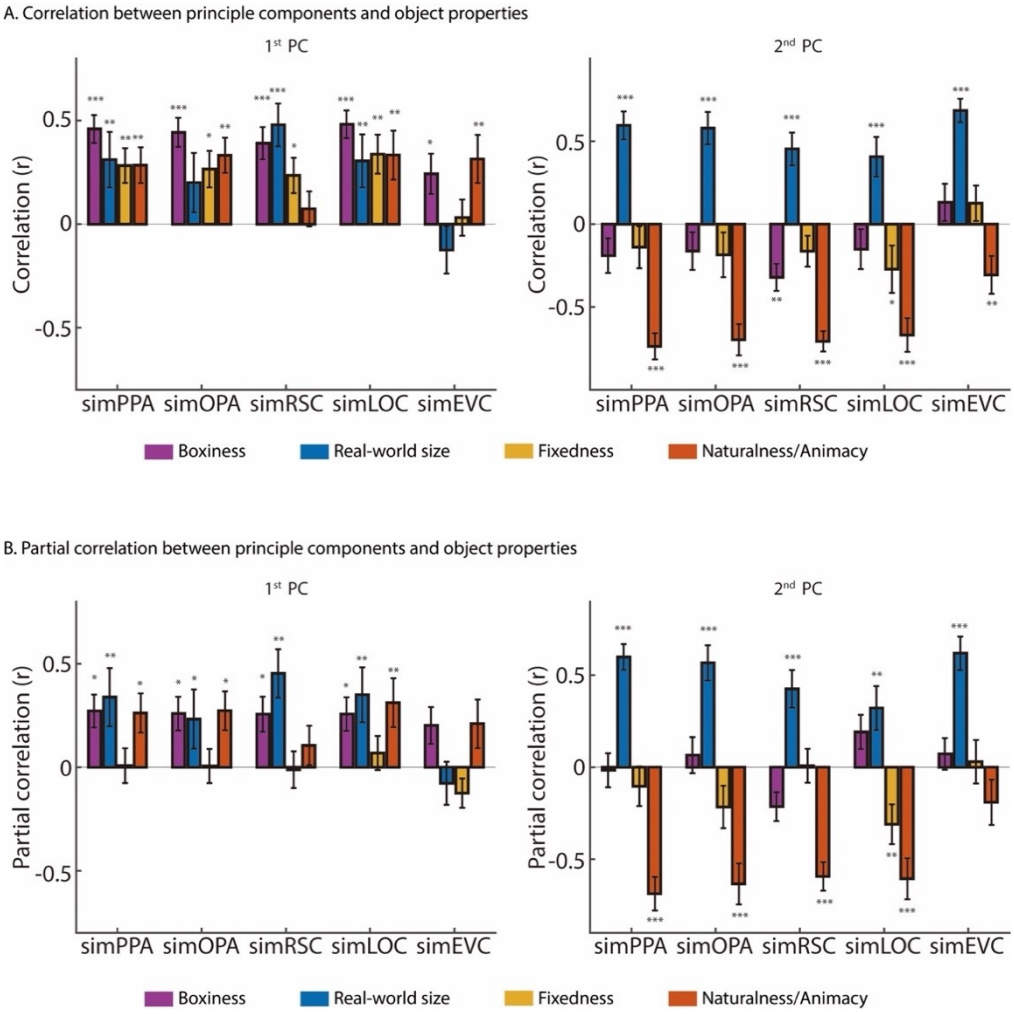
naturalness/animacy is an important object property contributing to the first PC. This aligned with previous findings that LOC is sensitive to animate objects. Furthermore, real-world size and animacy/naturalness contributed uniquely in the simPPA second PC, but not fixedness. These results suggested that the simPPA representational space contained representational dimensions of rectilinearity, real-world size and animacy/naturalness. However, the second PC of simLOC is uniquely sensitive to real-world size, animacy/naturalness and also fixedness while explaining more amount of variance. In this analysis, we have revealed that curvature, real-world size and naturalness/animacy as the primary dimension of the selectivity preference in both simPPA and simLOC.

4.6. Summary

In this chapter, I revealed a novel method, semantic preference mapping, to understand the semantic selectivity of the category-selective area models. I then used human behavioral object property ratings to characterize the tuning profiles in these models. First, from both the univariate and multivariate analyzes, the results suggested that the encoding models of category-selective areas showed classic tuning preferences to abstract semantic dimensions like real-world size, animacy, fixedness and naturalness and the perceptual dimension of curvilinearity. Second, since these models only carry out bottom-up feedforward processes, the success of characterizing high-level object properties in these models suggested that those properties could be computed through a series of bottom-up feedforward processes.

Figure 4.12 ***Principal component analysis (PCA) of selectivity indices in all ROIs.***

*The left panel shows the first PCs and the right panel shows the second PCs in different simROIs. Error bars indicate the +/-1 SD using bootstrapping (N=10,000). * indicates p<0.05, ** indicates p<0.01, *** indicates p<0.001. A: Bar plots show the correlation between object properties and the PCs of selectivity indices while regressing out occluder size. B: Unique contribution (partial correlation) between objects properties and the PCs of the selectivity indices in all ROIs are shown while taking the covariance of other object properties into account.*



CHAPTER 5. MID-LEVEL PERCEPTUAL FEATURE TUNINGS

In Chapter 4, I argued that semantic selectivity in the category-selective area models can emerge from mid-level features tuning. While these models show a classic semantic selectivity to high-level perceptual features, one remaining question is whether these models also exhibit selectivity to low- and mid-level perceptual features. In the study of category-selective areas, scientists found that these areas can be tuned to both high-level semantic and mid-level perceptual properties (Radoslaw Martin Cichy et al., 2011). Therefore, it is important to investigate whether such property is also observed in the computational model in order to understand whether the proposed computational architecture demonstrated a selectivity profile similar to the brain. In this chapter, I will explore whether these feedforward models exhibit low- and mid-level perceptual properties that have been previously found to be associated with the category-selective areas. In section 5.1, I will illustrate the importance of cardinal orientations in the representation in scene-selective cortex; in section 5.2, I will examine the importance of curvilinearity encoding in category-selective cortex through modelling curvature using a hand-engineered model.

5.1. Cardinal orientations

Orientation, as a low-level image property, has been shown to predict neural response across category-selective areas along the ventral stream (Rice et al., 2014). In particular, there is a substantial amount of evidence that orientations of contours can predict the image category of scene and object images (Olshausen & Field, 2000; Simoncelli & Olshausen, 2001). While recent fMRI studies reported low-level visual regions like V1 does not bias towards certain orientations (Freeman et al., 2011; Swisher et al., 2010), scene-selective area PPA shows higher sensitivity towards

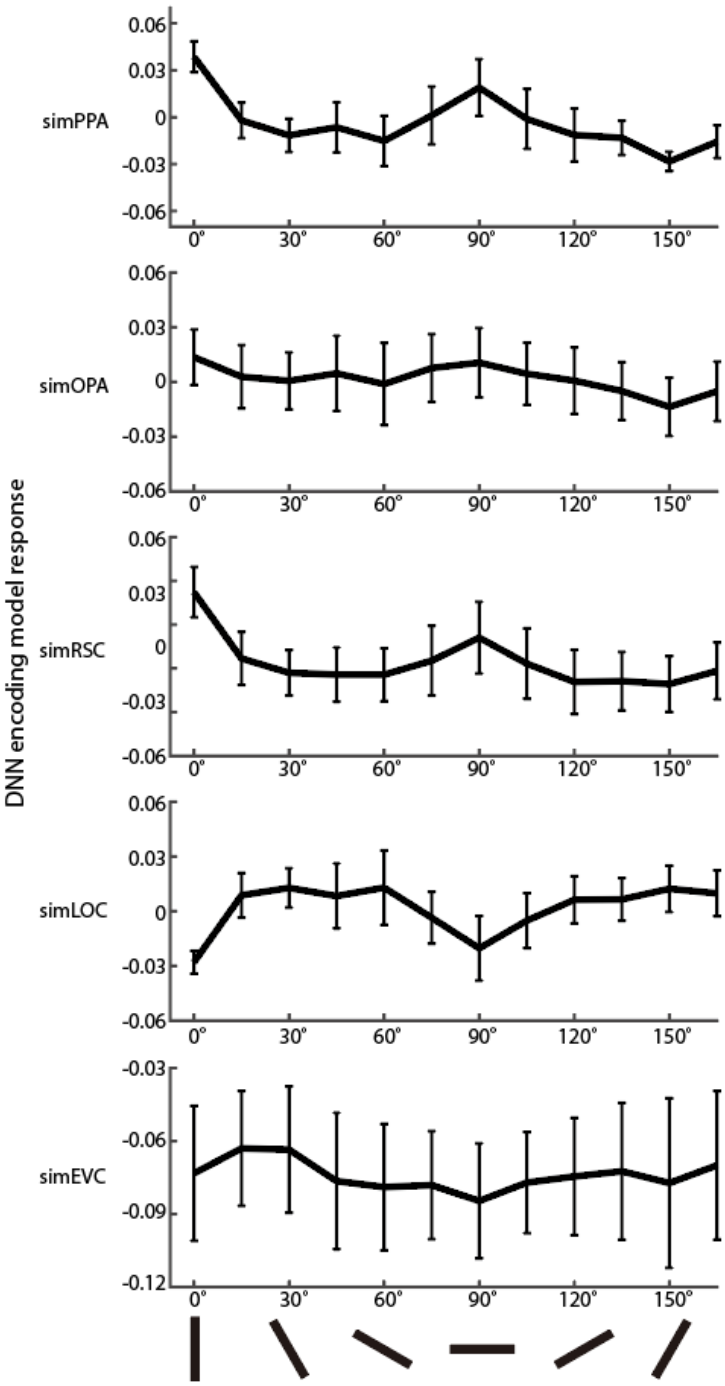
horizontal and vertical contours (i.e., cardinal contours) than oblique contours (Nasr & Tootell, 2012). Image statistics analysis of scene images confirms that scenes are dominated by cardinal orientations compared to oblique orientations (Torralba & Oliva, 2003), so the cardinal orientation selectivity in PPA is believed to be linked to the natural image statistics of scenes.

A visual inspection of the selectivity profile of simPPA suggests that this model prefers object categories that contain a lot of cardinal contours like buildings, skyscrapers and bookcases. In this analysis, I rigorously tested whether simPPA and other models exhibit such tuning preferences by passing simple Gabor stimuli with different orientations to the models. Results are shown in Figure 5.1, simPPA shows a higher response to cardinal orientations compared to oblique orientations. This result aligns with previous findings of cardinal preferences of PPA. As discussed in Chapter 4, simPPA is selective to landmark objects, which likely contain a lot of vertical and horizontal contours; therefore, the observed effect here suggests the semantic selectivity for fixedness and real-world size in simPPA may rely on a preference for mid-level features whose contours are predominantly at cardinal orientations. On the other hand, another scene-selective ROI model, simOPA, does not exhibit such cardinal orientation preference. Nasr & Tootell, 2012 also did not observe cardinal orientation preference in OPA although OPA is sensitive to scene images. Current findings from Bonner & Epstein, 2017 suggests OPA is sensitive to the navigational affordances of scenes, such as paths. Navigable trajectories span a large range of orientations, not only the cardinal orientations, which may explain why simOPA does not exhibit such cardinal orientation preferences. To serve as a control, simEVC shows no orientation preferences which is also observed in other studies (Freeman et al., 2011; Swisher et

al., 2010). To conclude, this analysis suggests that the representations encoded in simPPA exhibit a perceptual bias towards contours at cardinal orientations.

Figure 5.1 ***Selectivity preferences to different orientations for simulated ROIs.***

simPPA shows a higher activation to vertical and horizontal orientations compared to oblique orientations. Error bars indicate +/-1 SD across subjects.



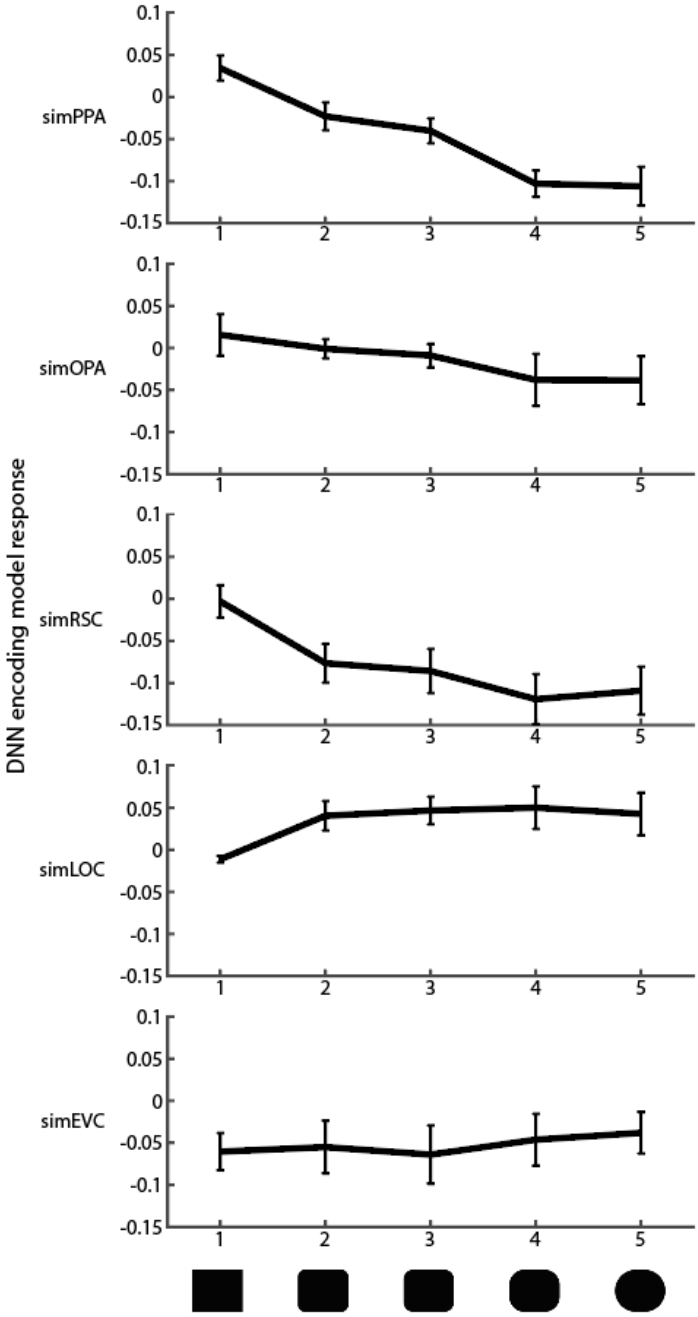
5.2. Curvature

Apart from cardinal orientations, curvilinearity has been shown to be an important perceptual feature encoded in the category-selective areas (B. Long et al., 2018). Curvature and rectilinearity are considered as a second-order mid-level perceptual feature. In scene-selective areas, PPA, OPA and RSC are shown to be sensitive to rectilinear shapes compared to curvy shapes (Nasr et al., 2014; Yue et al., 2020). One of the explanation is that scenes contain many manmade objects like buildings and skyscrapers which contain a lot of rectilinear contours (Chao et al., 1999). In object-selective areas, IT and LOC are shown to prefer curvy shapes compared to rectilinear stimuli (Yue et al., 2014, 2020). The curvature preference is speculated to be related to the animate object preferences in these areas as most animals and animate objects contain a lot of curvy contours (Konkle & Caramazza, 2013).

Do the simulated ROI models exhibit curvilinearity preferences? In this analysis, several simple shapes were passed through the simulated models to evaluate whether such preference can be observed in these models. In Figure 5.2, scene-selective models – simPPA and simRSC show a clear preference to the boxy shape, their responses are higher to the square stimulus than to all other stimuli with curved corners. On the contrary, simLOC exhibits a reversed pattern, with higher responses to all stimuli containing curved corners compared to the square stimulus. These results are consistent with previous findings that scene-selective areas are sensitive to boxy shapes while object-selective areas are sensitive to curvy shapes (B. Long et al., 2018; Nasr et al., 2014; Yue et al., 2020). Furthermore, these results suggest the curvilinearity preferences in these models could relate to the semantic-selectivity observed in Chapter 4.

Figure 5.2 ***Selectivity preferences to curvilinearity for simulated ROIs.***

Scene-selective models – simPPA and simRSC show a higher activation to the boxy shape compared to shapes with curved corners. Object-selective model – simLOC shows a higher activation to curvy shapes compared to boxy shape. Error bars indicate +/-1 SD across subjects.



In the following, I will demonstrate an approach to build an image-computable model of curvature summary statistics for any input image. This is one of the first hand-crafted models for computing summary statistics of mid-level curvature features for any natural image. This model takes the input image and convolves with a curvature filter bank, then finds the curvature level of each edge pixel to create a curvature distribution. In this analysis, I calculated the mean curvature level of the distribution as the curvature index (i.e., curvy-boxy index).

5.2.1. Curvature filter bank

Gabor filters are used extensively in both computer vision applications and neural signal modelling. These filters have been shown to capture oriented bars and edges in both biological vision model and computer vision model (Mehrotra et al., 1992). Here, I will introduce a technique called curvature wavelet which extends Gabor filters to not only detect straight edges, but also curved edges (Ibrahim, n.d.). A single curvature wavelet can be thought of as a feature detector that detect a contour with a particular orientation and curvature. In this model, a number of these wavelets were created to form a filter bank which can detect different oriented curves.

Each filter is built by combining a rotated and curved complex wave function (F) and a rotated and curved Gaussian function (G). A bias term is also added to make sure the whole wavelet has a sum of zero. This curvature wavelet idea is borrowed from the construction of the Gabor wavelet filter, which consists of a sinusoid function and a gaussian function. A single curvature wavelet filter is parameterized by six variables, including frequency (f), orientation (θ), curvature (c) and size (s), and scale of the gaussian filter in x (σ_x) and y (σ_y) direction. Each filter can be composed by the following mathematical formulas:

$$B(x,y) = G(x,y) \cdot (F(x,y) - bias)$$

$$G(x,y) = \exp\left\{-\frac{f^2}{2} \cdot \left[\frac{(x_c + c \cdot x_s^2)^2}{\sigma_x^2} + \frac{x_s^2}{s^2 \cdot \sigma_y^2}\right]\right\}$$

$$F(x,y) = \exp(i \cdot f \cdot (x_c + c \cdot x_s^2))$$

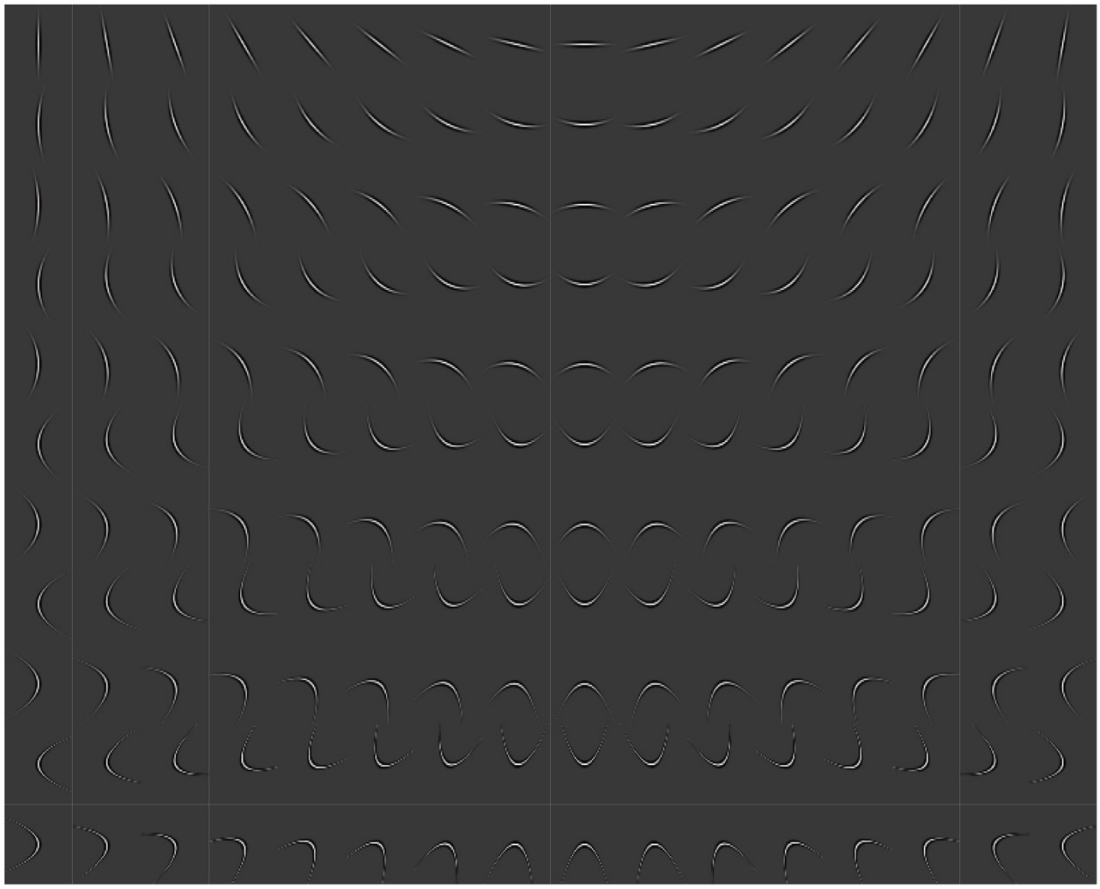
$$x_c = x \cdot cos\theta + y \cdot sin\theta$$

$$x_s = -x \cdot sin\theta + y \cdot cos\theta$$

$$bias = e^{-\frac{\sigma_x}{2}}$$

To ensure that I sampled a sufficient number of orientations and curvature levels, I have created the curvature filter bank with a wide range of parameter values in both the curvature level (6 levels were sampled in log scale from 0 to 1/12 and each level except 0 has a concave [positive curvature] and a convex level [negative curvature] to create 11 curvature levels in total) and orientation (16 orientations evenly sampled from 0 to 180 degrees), with fixed frequency (1.2), size (50 pixels*50 pixels) and scale of gaussian filter (1 in both directions). Parameters were chosen empirically to best capture natural image statistics. This filter bank consists of a total of 176 curvature wavelet in total. A subset of the wavelet filters is shown in figure 5.3.

Figure 5.3 ***Subset of the curvature filter bank to illustrate the sampling space of curvatures and orientations.***



5.2.2. Curvature model

The curvature model consists of four steps to compute the curvature index of a give image. Figure 5.4 illustrates how the model computes curvature rating of a given image.

Step 1. Compute curvature feature map: The curvature model starts with convolving the curvature filter bank to the grayscale input image with paddings to keep the output feature map in the same size as the input image, this results at 176 feature maps.

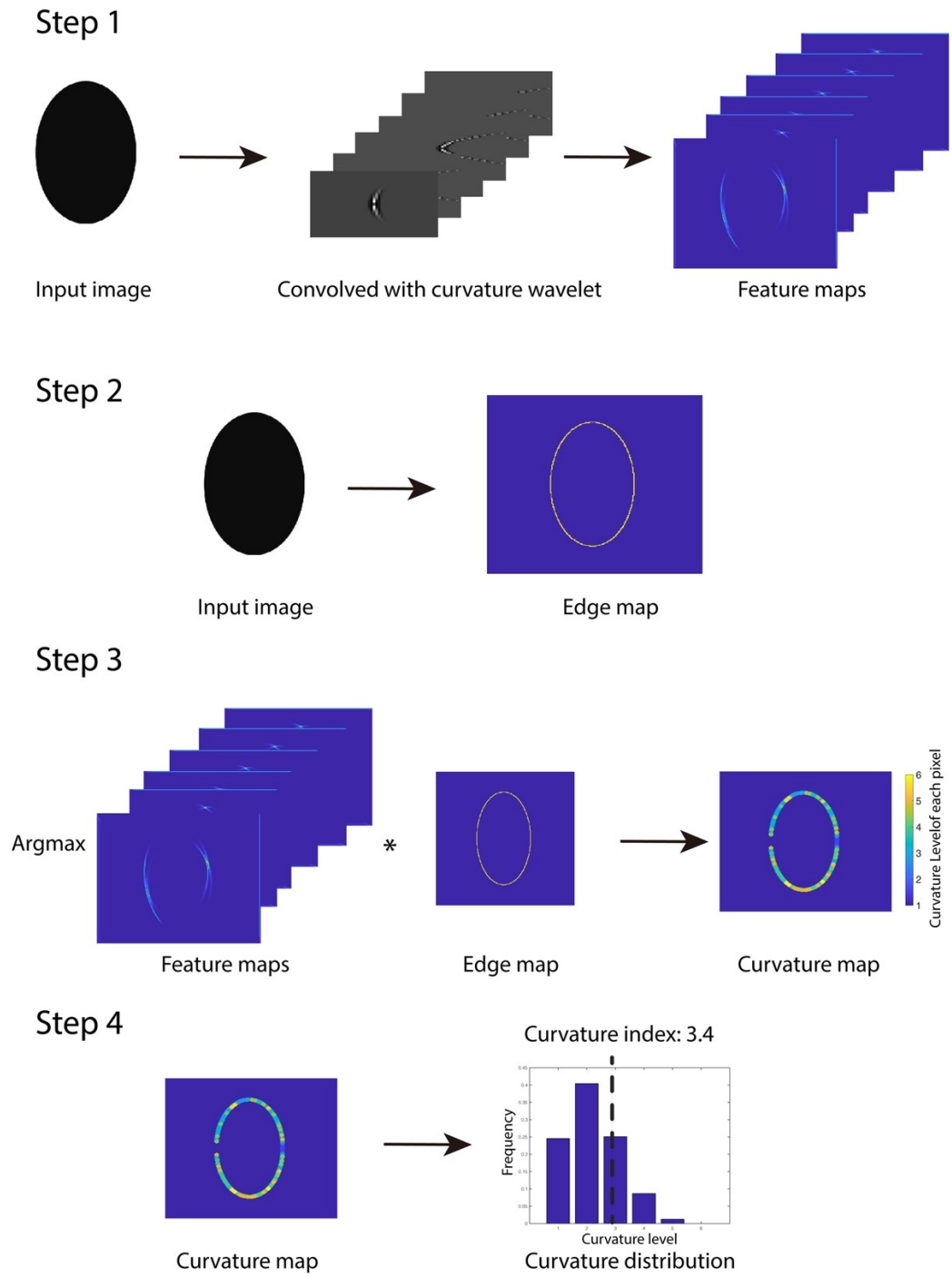
Step 2. Edge detection: The purpose of this step is to identify edge pixels. The grayscale input image is fed into an edge detection algorithm to separate edge pixels from other pixels. A Roberts edge detector is used, and this results an edge map of the input image.

Step 3. Pixel-wise curvature level: The goal of this step is to compute the curvature level of each edge pixel. A higher value of the feature map suggests a higher similarity between the corresponding curvature wavelet and the local image patch, so the highest value across feature maps suggests that the corresponding curvature wavelet is the most similar to the local image patch. The model assigns the curvature level of each edge pixel by finding the corresponding curvature wavelet that maximizes the feature maps.

Step 4. Curvature distribution: For each of the 6 curvature levels, compute the percentage of edge pixels that has the corresponding curvature level. The resulting distribution represents the percentage of edge pixels in each curvature level. Mean curvature of the whole image can then be calculated from this distribution. The mean curvature level across edge pixels is considered as the curvature index of the input image.

Figure 5.4 ***Illustration of the curvature model.***

The model first convolves the input image with a bank of curvature filters and then identifies the best-fitted curvature level of each edge pixel. Finally, it obtains a global curvy-boxy index by averaging the curvature level across all edge pixels.



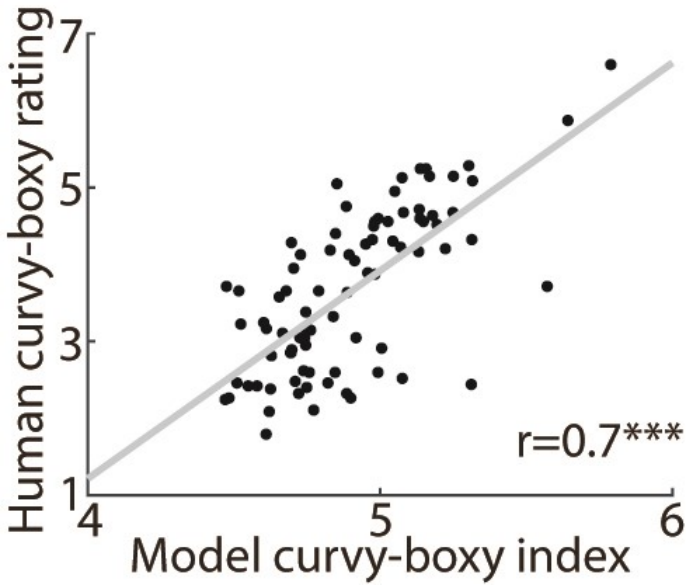
5.2.3. Curvature index

In this section, I will demonstrate the use of the curvature model in understanding the contribution of curved contours to the mid-level representations of the simulated ROIs. First, to test whether the curvature model captures curvature summary statistics that match human judgments, I used the model to generate curvy-boxy indices of object images, and compared with human curvature ratings. Second, to characterize the contribution of curvilinearity to the tuning profiles of the simulated ROIs, I generated the curvature indices of the occluded objects from the semantic preference mapping procedure and correlated the curvy-boxy indices with the selectivity indices.

In the first analysis, I used the human curvature ratings discussed in Chapter 4 to understand whether the curvature model rated curvilinearity in the way that humans do. To obtain the model curvy-boxy index, I cropped the occluded objects used in the semantic-preference mapping procedure and passed them into the curvature model to obtain a curvy-boxy index. Then I averaged the curvy-boxy index across each object category and correlated them with the human curvy-boxy ratings. In Figure 5.5, the result indicates a strong correlation between the human curvature ratings and the model curvy-boxy index ($r=0.7$, $p<0.0001$). Such strong correlation suggests that the curvature model is able to capture the human ratings. This suggests that the curvature model successfully captures important summary statistics about the presence of curved features in natural images.

Figure 5.5 ***Correlation between model*** **curvy-boxy** ***index and human curvature ratings.***

*Each dot represents curvature ratings of an object category. The x-axis shows the mean curvy-boxy index from the curvature model, and the y-axis shows the human curvy-boxy ratings. *** indicates p<0.001*



In the second analysis, I focused on understanding whether curvilinearity as a mid-level perceptual feature contributes to the semantic tuning profiles in the simulated models. The curvy-boxy index of the occluded parts of those images were computed using the curvature model. These ratings were then correlated with the selectivity indices using Pearson correlation. Figure 5.6 shows the result of this analysis.

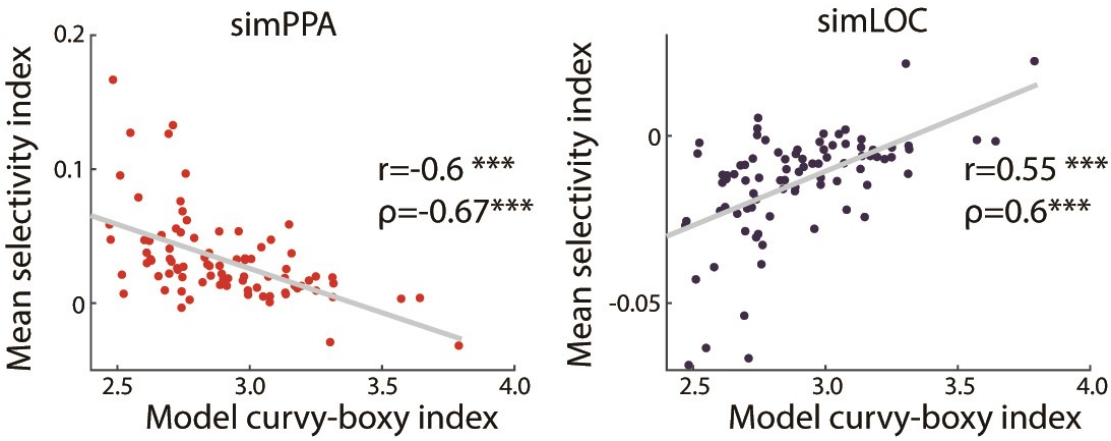
Results indicated that the selectivity index result from simLOC is positively correlated with the curvy-boxy index ($r=0.55$, $p<0.001$). This shows that curvier objects led to a greater selectivity index for simLOC. This suggests that object-selective area is sensitive to mid-level features with a predominance of curved contours. This results aligned with previous studies showing that curvature preference is an important mid-level feature tuning of the object-selective areas (B. Long et al., 2018).

On the contrary, the selectivity index from simPPA showed a negative correlation with the curvy-boxy index (PPA: $r=-0.6$, $p<0.001$). This shows that rectilinear objects elicit stronger responses in simPPA. This provides evidence that the mid-level feature preferences of simPPA give rise to not only high-level semantic preferences for landmark objects but also lower-level perceptual preferences for rectilinear contours. This may reflect the fact that landmark objects tend to have a predominance of rectilinear contours.

Taken together, these results suggest that the selectivity indices of category-selective areas are modulated by the curvilinearity of the visual input, thus the curvilinearity preferences in the category-selective areas suggested the mid-level features encoded within these areas may covary with the visual shape structure.

Figure 5.6 ***Curvature model correlation with selectivity index.***

*Each dot represents curvature index of an object category. The model curvy-boxy indices are negatively correlated in simPPA (Pearson's r=-0.6, p<0.0001) and positively correlated in simLOC (Pearson's r=0.55, p<0.0001). *** indicates p<0.001.*



In conclusion, this chapter has illustrated two important perceptual features – cardinal orientation and curvilinearity – are contributing to the tuning profiles in the computational models of category-selective areas. In particular, simPPA prefers vertical and horizontal contours and boxy shapes, while simLOC is more sensitive to oblique contours and curvy shapes. These results imply that feedforward models with high-level semantic selectivity for scenes and objects also exhibit characteristic patterns of lower-level perceptual biases. A thorough understanding of these low- and mid-level perceptual features can help us recognize how the feedforward process of vision can make use of these perceptual features to obtain abstract properties of the visual input. Last but not least, a computational model like the proposed curvature model could be used as a demonstration of an interpretable computational model of perceptual features.

CHAPTER 6. GENERAL DISCUSSION

In this work, I employed an *in-silico* experimental approach to address the debate on the level-of-interpretation for representations in category-selective areas in the ventral stream. While many argue that the activations in these areas are either better explained by mid-level perceptual or high-level semantic properties, I hypothesized that dichotomizing the representation into purely perceptual or purely semantics could be oversimplified. The current results revealed that the CNN encoding models of category-selective areas exhibited tuning preferences to both mid-level perceptual features and high-level semantic properties. Thus, although these models were not explicitly trained to represent specific high-level object properties, semantic selectivity was nonetheless evident in the responses of these models to a large and diverse sample of natural images. I briefly summarize the key findings below.

First, through building CNN encoding models of the category-selective cortex, I found that the representations in both scene-selective and object-selective areas could be well explained from a series of purely feedforward computational processes. The explanatory power of the models suggests that feedforward processes capture a lot of explainable variance in the category-selective areas.

Second, by running large-scale *in-silico* experiments on the category-selective computational models, I discovered that the tuning profiles of these models can be explained by interpretable object properties such as real-world size, fixedness, animacy and naturalness. These effects were robust to variations of model parameters and experimental procedures, and both univariate and multivariate analyzes revealed the same selectivity pattern of the simulated ROIs. In particular, scene-selective simPPA showed a preference toward fixed, large in real-world size, inanimate and unnatural objects; on the other hand, object-selective simLOC preferred mobile, small in real-

world size and animate objects. Since the models are feedforward in nature, these results demonstrated that the semantic selectivity profiles in these regions could emerge through bottom-up perceptual processes.

Third, these feedforward computational models not only showed the classic semantic selectivity profiles, but they also illustrated low- and mid- level perceptual biases of cardinal orientations and curvilinearity. These results suggested that a feedforward architecture of the category-selective areas would give rise to both high-level semantic preferences and a characteristic pattern of lower-level perceptual biases. Below, we discuss the implications of these findings in understanding the visual cortex.

6.1. The role of computational models in understanding vision

Vision is a complicated cognitive process which involves a hierarchy of non-linear computations. Given the complexity of vision, a multilayer system with non-linear components like a CNN is necessary to model its underlying mechanisms. Although the internal representations of complex non-linear models like CNNs are hard to interpret from merely inspecting the model's parameters, there are methods for studying these models that can provide insight into the nature of their internal representations and their relationship to the natural statistics of scenes. For example, the CNN encoding models of category-selective areas are hard to interpret on the surface, but in Chapter 4 and Chapter 5, with the help of *in-silico* experiments, I characterize some perceptual and semantic properties that these models are sensitive to. Without a proper computational model, it would be hard to test the idea that a unified model can exhibit such multifaceted selectivity profiles with preferences for both low-level perceptual and high-level semantic image attributes.

A unified computation model of the category-selective areas can inform us on how the category-selective representations could arise through mid-level feature tuning. The current CNN encoding models of category-selective areas serve as a proposal to understand the necessary computations to achieve the representations of category-selective cortex. One of the advantages of the current DNN encoding models is that CNNs provide explicit models of the required computations that underlie a cognitive function, including convolution and all the non-linear mappings. While this class of models does not characterize representations descriptively, I argue that in order to fully understand how category-selective areas work, different classes of models, including computational models are essential. The coexistence of diverse classes of models can help address different aspects of the category-selective areas. Another class of cognitive model, the cognitive-architecture (or "box-and-arrow") model is widely used to decompose a complex cognitive process into constituent cognitive functions that can be described in words and mapped onto intuitive concepts. In the box-and-arrow cognitive model of vision, more often it is used to address the "what" question, that is what representation is being characterized in each cognitive function, and what process or what neural substrates is involved in such representation. On the other hand, the computational model focuses on the "how" question, that is how such process can be achieved mechanistically. Ultimately, a holistic understanding of these areas requires the answers to both types of questions; therefore, a unified computation model for each category-selective area is essential to help us answer the "how" question.

Computational models can help us test the level of complexity of the computation. In Chapter 3, I compared the predictive power of encoding models using different CNN layers, and I found that the later convolutional layer (i.e., convolutional layer 5 in AlexNet) provides better predictive power in the category-selective regions

compared to earlier layer (i.e., convolutional layer 1 in AlexNet). Although this results itself does not inform us about what is represented in these layers or models, the result of this comparison suggested that the representation of category-selective areas may require multiple non-linear computational operations to achieve, and a single stage of non-linear computation may not be sufficient.

6.2. The nature of representation of the category-selective areas

As discussed in Chapter 2, the debate on the level of interpretation of the representation in the category-selective areas focuses on whether the tuning profiles of these areas are perceptual in nature or semantic in nature. In Chapter 4 and 5, I examine both hypotheses in the CNN encoding models of category-selective areas. Surprisingly, these models not only exhibited emergent semantic selectivity that is consistent with previous findings, but these models also showed previously identified lower-level perceptual biases. This suggests two levels of interpretation for these models. On the one hand, these models respond preferentially to the semantic attributes of objects like real-world size, animacy and fixedness. Alternatively, these models are also tuned to image-computable perceptual features of cardinal orientations and curvilinearity. These two levels of interpretations are not mutually exclusive, and they both provide useful and accurate descriptions of these representations. A plausible explanation for the observed result is that given the natural image statistics of covariance between perceptual features and semantic properties (as shown in Chapter 4 that curvilinearity covaries with fixedness, animacy and naturalness), mid-level feature tuning may be sufficient for rapidly embedding visual inputs into a representational space that is organized along meaningful semantic dimensions for all image that do not drastically diverge from statistical regularities of natural images. All in all, the current modeling

approach is a proof of principle that there exist mid-level representations that exhibit the classic category-selective effects across a large sample of natural images, which suggests that it is possible for the semantic-selectivity profiles of these models to emerge from mid-level perceptual tuning.

6.3. Organizational principle of scene-selective areas

In the scene-selective areas, there are two main observations to note about the representational structure of PPA, each associate with the implications for the function of PPA. The first observation is that PPA representation could be largely explained by properties related to landmark objects. In Chapter 4, I showed that simPPA is sensitive to large and fixed objects. These high-level semantic features are speculated to be related to the landmark object features (Julian et al., 2017; Troiani et al., 2014). In Chapter 5, simPPA was shown to be sensitive to cardinal orientations and rectilinear shapes, which are possible perceptual features linked to landmark objects. Taken together, simPPA is sensitive to both high-level semantic and low-level perceptual features associated with landmark objects, which align with previous findings (R. A. Epstein & Baker, 2019; I. I. A. Groen et al., 2017; Marchette et al., 2015; Troiani et al., 2014).

The second observation is that bottom-up mid-level features explain a large amount of variance in the PPA representations. simPPA has the best predictive power among different ROIs, which suggests that simPPA heavily relies on bottom-up image computable perceptual information. In Chapter 5, simPPA was shown to exhibit strong low-level and mid-level feature preferences. One speculation is that information encoded in simPPA contains a lot of texture information which relies on low- and mid-level perceptual features. In previous work, PPA has been shown to be sensitive to

texture (J. Park & Park, 2017), which is similar to intermediate CNN representations (B. Long et al., 2018). Further work is needed to determine whether simPPA is particularly sensitive to texture information.

6.4. Organizational principle of object-selective areas

In the object-selective area, the selectivity profile in simLOC is almost a direct opposite of the scene-selective areas. In previous studies, LOC is argued to be a neural substrate that encodes the shape and category of objects (Kim et al., 2009; Shpaner et al., 2013). We found that simLOC showed a unique strong preference towards small and curvy objects which is consistent with previous findings. Previous studies suggested that shape may be intrinsically correlated with object size based on gravitational and physical constraints; therefore, smaller objects have a higher probability of being curvy in shape (Konkle & Oliva, 2012). Although I did observe a weak correlation between real-world size and curvature in the human object rating, the partial correlation analysis revealed that there are unique contributions for real-world size and curvature, thus this suggests the selectivity to real-world size cannot be solely explained by selectivity to curvature in the simLOC tuning profiles.

The PCA result suggested simLOC requires 2PCs to explain most of its variance. The first PC of simLOC shows a similar tuning profile as the selectivity profiles in the univariate analysis, which is uniquely sensitive to curvilinearity, real-world, naturalness and animacy. In the second PC of simLOC, it is strongly sensitive to the naturalness and animacy of objects, while also moderately sensitive to fixedness and real-world size in the partial correlation analysis. LOC's sensitivity to animate objects was observed in an earlier study, and I speculate that the secondary dimension encoded in simLOC is strongly related to the animacy preference of LOC (B. Long et

al., 2018; Sha et al., 2015). All in all, the two separate dimensions observed in simLOC suggest that multiple, orthogonal object properties may drive the responses of LOC.

6.5. Organizational principles in the ventral stream

Curvilinear tuning has been shown to be an important property in primate and human mid-level visual area V4 (El-Shamayleh & Pasupathy, 2016; Habak et al., 2004; Yau et al., 2013). Curvature information was shown to be an important feature that can be used as an organizing principle along the ventral visual stream (B. Long et al., 2018; Yue et al., 2020). There is a systematic preference for curvilinear versus rectilinear stimuli in different category-selective regions. In this study, I systematically investigate the unique effect on curvature while taking the covariance of other object properties into account, and I found that such unique contribution of the curvy-boxy index in the category-selective regions could not be solely explained by other high-level semantic properties such as real-world size, animacy/naturalness, and fixedness. In the curvature model, I demonstrated that the summary statistics of curvilinearity is highly correlated with the selectivity profiles in the simulated ROIs. Therefore, curvilinear information is an important bottom-up perceptual feature that may act as an organizing principle in the ventral visual stream, and such information contributes as a unique dimension different from other object/scene properties.

In particular, my results suggest that the scene-selective PPA which is located in the medial part of the ventral stream, tends to prefer rectilinear information, while the object-selective LOC which is located in the lateral part of the ventral stream prefers curvy information. This medial-lateral anatomical organization of the curvy-boxy preferences was also found in previous studies (B. Long et al., 2018; Yue et al., 2020). In addition, the medial-lateral organization also has an implication on the receptive field

biases. The retinotopic biases of medial LOC shows that it responds preferentially to the center of the visual field, while the retinotopic biases of lateral PPA shows a peripheral preference. These results support some earlier findings suggesting that curvilinearity preferences interact with central-peripheral biases (Yue et al., 2020). The central-periphery organization is consistent with the previous speculation that curvilinear objects are more frequently foveated in the central visual field while rectilinear objects are more frequently processed by peripheral vision (Ponce et al., 2017; Yue et al., 2020).

6.6. Does CNN explain everything?

My results, together with many others (Bonner & Epstein, 2018; Henriksson et al., 2019; Khaligh-Razavi & Kriegeskorte, 2014; Schrimpf et al., 2018; Yamins et al., 2014; Zhuang et al., 2021) have shown that CNNs and encoding models were able to account for nearly all explainable variance in the responses of high-level visual cortex to naturalistic visual stimuli. However, cautious interpretation of these results is necessary. In this study, I applied LASSO regression in the CNN encoding models, and the resulting models relied on a small number of CNN units to predict neural representation, and we also found that such regression performed better than using OLS regression, which utilized all units. Similar regularization techniques (e.g. ridge regression, elastic net) were used in many other studies (Nunez-Elizalde et al., 2019). Therefore, the similarity between CNNs and brains could be driven by a small subset of units in the CNN rather than the whole CNN feature representational space.

One current hurdle in the field is that many state-of-the-art CNNs do not have interpretable internal representations. Therefore, even if scientists find that CNN representations are similar to the human brain or behavioral representations, it does not

inform us on the nature of the representations. Development of *in-silico* experiments are important to understand the CNN models. In this study, I proposed the semantic-preference mapping approach to explore the tuning profiles of the CNN encoding models. The results yielded better understanding of the semantic selectivity in the category-selective area models. Apart from semantic preference mapping, currently there are different *in-silico* experiments conducted to yield a better understanding of the internal representation, including examining the input that drives the responses of a particular neuron/unit (Bashivan et al., 2019; Bau et al., 2020; Bonner & Epstein, 2018; Srinath et al., 2021; Walker et al., 2019). Although CNNs differ in many ways from human brains and they can be hard to interpret, they are nonetheless powerful computational tools that can be leveraged to understand the possible mechanisms and representations that underlie human cognition (Radoslaw M. Cichy & Kaiser, 2019).

6.7. Modelling image computable summary statistics of mid-level features

Mid-level features such as curvilinearity can be abstract and hard to quantify. In particular, it is hard to quantify such dimension in natural images. Previous studies relied on experimenter's intuitions or human ratings to quantify these dimensions (Hebart et al., 2019, 2020; Konkle & Caramazza, 2013; Konkle & Oliva, 2012; B. Long et al., 2018). While in our study, we also adopted a similar approach to obtain behavioral ratings to quantify the perceptual dimension of curvilinearity, we also took another pathway to quantify this dimension by developing the curvature model. A computational model of a mid-level feature can be thought as a formal system to quantify perceptual summary statistics. Such formal system can help researchers to understand its function in vision and how it can be derived from low-level features. One

advantage of using an objective computational model instead of human ratings in quantifying mid-level features is to generate an unbiased bottom-up summary statistic and avoid the problem of (i) potentially biased subjective ratings; (ii) ambiguous instructions which participants interpret differently; and (iii) using context instead of the dimension itself in judging the features. Another advantage of using the objective computational model is it allows for the scaling up of large-scale image analyses because the computational model does not require any human ratings.

6.8. Future directions

We proposed a novel semantic-preference mapping approach to explore the selectivity profiles in the category-selective ROIs. While the current occlusion procedure is a reliable computational method, the occluded images are not considered as natural images. In order to better simulate natural occluded images, generative networks like PixGan, PixRNN, Generative Adversarial Network (GAN) and Progressive Generative Network (PGN) (Cai & Wei, 2020; Dolhansky & Ferrer, 2018; Aaron van den Oord et al., 2016; Yeh et al., 2017; Zhang et al., 2018) could be used to generate the natural occluded images without the random pixel occluder for future experiments. Figure 6.1 illustrated some inpainting images generated by current state-of-the-art inpainting algorithms.

In this study, I focused on the feedforward processes of the category-selective areas. However, recent studies suggest that recurrent connections and long-short-term-memory (LSTM) models can improve the explanatory power of the visual cortex (Kietzmann et al., 2019; Nayebi et al., 2021; Schrimpf et al., 2018). While this study showed that feedforward processes alone could explain a large amount of variance in the category-selective areas. From my other study (S. Park et al., 2020), I showed that

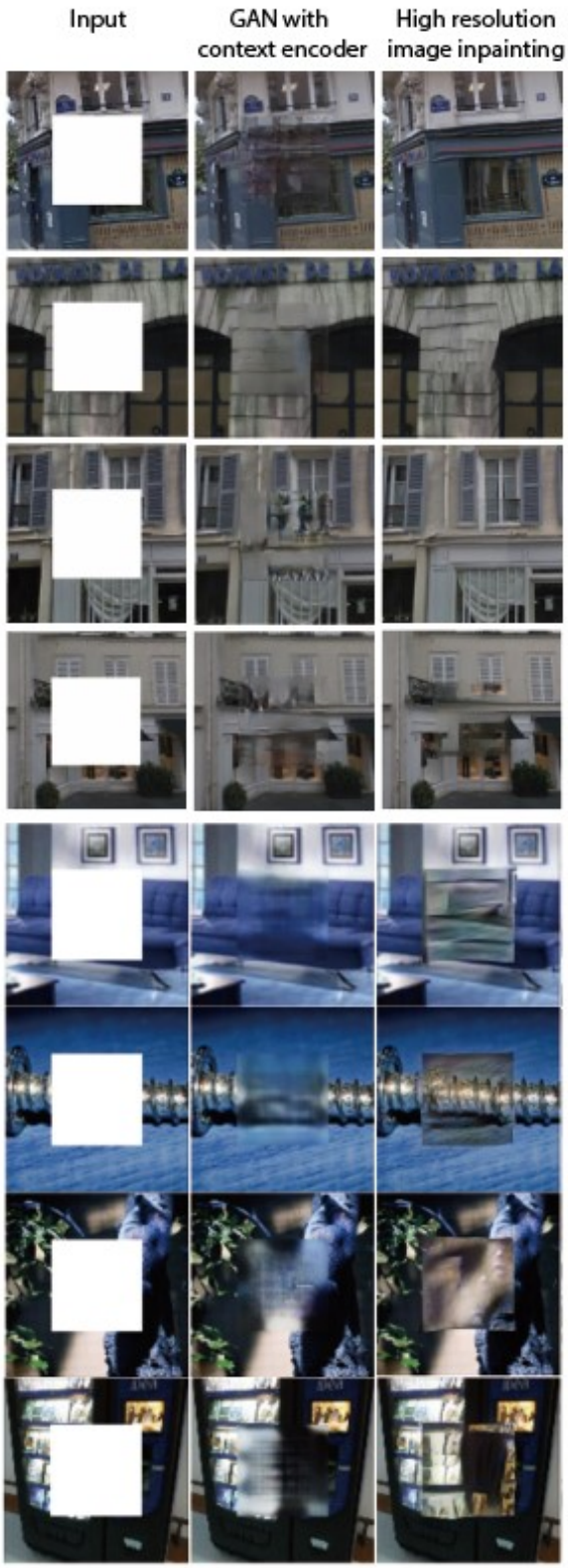
the scene-selective PPA responded differently to scenes which are associated with different navigational experiences while controlling for perceptual features. This result suggested that PPA response could be affected by the top-down effect of the memory associated with scenes. In object-selective area, Large et al., 2007 observed that the LOC responses could be modulated by the task given even when the visual input is the same. All these results point to the fact that category-selective regions are influenced by top-down information, which is not captured in the current study. While the current study suggests the category-selective areas capture the covariance between semantic-selectivity and perceptual feature preferences through bottom-up feedforward processes, one of the unanswered questions is how much the feedforward process contributes to understand the visual world. While it is very hard to gain insight on the causality of cognitive processes from human neuroimaging experiments, computational models like ours allow researchers to make predictions and generate causal image manipulations to test the contributions of specific bottom-up features in scene and object processing. Furthermore, our model could be built upon and compared with other architectures to test theories about the role of feedback and recurrent processes.

In the semantic preference mapping approach, the occlusion procedure allows us to determine how the representations change when an object category is occluded. This demonstration provides a proof-of-principle method on understanding the causality between the visual input and representations in the computational models. This technique can be extended to investigate the causality relationship between other image features and model activations. For example, this could be applied to study the effect of spatial frequency on the representation by filtering a specific range of spatial frequencies. By applying these techniques to mid-level perceptual features in follow-up fMRI experiments, scientists can further understand the causal relationship between

perceptual features and high-level semantic tuning in visual cortex. For example, if occluding certain mid-level features does cause an effect on semantic selectivity in the category-selective areas, then one may conclude that the semantic selectivity observed in these regions could be casually affected by this particular mid-level feature. On the contrary, if occluding a certain mid-level feature does not affect the selectivity profiles of such regions, then that perceptual feature may be independent of the semantic selectivity. Revealing the causal relationship between perceptual features and semantic selectivity could be important for understanding the relevant perceptual features for a cognitive function.

Figure 6.1 ***Images generated by inpainting generative networks.***

Adapted from (Zhang et al., 2018). Current GANs could in-paint images with high-fidelity.



6.9. Conclusion

The present work investigated the tuning preferences of the category-selective areas using CNN encoding models and *in-silico* experiments. I found that a purely feedforward CNN architecture was able to explain most of the explainable variance in the fMRI activations in these areas. Our semantic preference mapping procedure demonstrated that these models are selective to some high-level semantic properties including real-world size, fixedness, animacy and naturalness of objects and that such preference could emerge through bottom-up processes. Apart from abstract semantic properties, these simulated ROIs were also tuned to low- and mid-level perceptual features like cardinal orientations and curvilinearity. This study, for the first time, showed a unified mechanistic model of category-selective areas that captures both the perceptual and semantic feature preferences of these areas. These results suggest that the mid-level tuning in category-selective visual cortex may be shaped by the covariance between image-computable perceptual features and high-level semantic properties. To conclude, this work provides evidence to situate the level of representation in category-selective areas, showing that the semantic selectivity in high-level vision could be an emergent phenomenon of mid-level feature tuning.

APPENDIX

Table A.1 Semantic preference mapping results.

Ranking from objects causing the biggest activation decrease to objects causing the lowest activation decrease.

simPPA	simOPA	simRSC	simLOC	simEVC
skyscraper	skyscraper	skyscraper	ball	bookcase
house	bookcase	house	animal	food
bookcase	house	building	floor	fireplace
building	computer	bookcase	rug	ball
computer	windowpane	computer	person	sofa
windowpane	building	windowpane	magazine	windowpane
fireplace	fireplace	base	rock	computer
base	curtain	fireplace	trade name	curtain
curtain	base	road	figurine	stove
road	sink	palm	telephone	building
swivel chair	chest of drawers	swivel chair	switch	armchair
blind	stove	sea	light	stool
chest of drawers	swivel chair	sky	ashcan	house
palm	blind	curtain	fluorescent	coffee table
sink	armchair	blind	wall socket	railing
desk	desk	field	shoe	desk
stove	food	desk	spotlight	base
sky	chandelier	chandelier	pot	floor
column	column	painting	bicycle	truck
chandelier	railing	chest of drawers	plaything	bicycle
railing	coffee table	column	glass	chest of drawers
painting	painting	earth	boat	rug
armchair	road	railing	minibike	blind
sea	sofa	sink	stairs	skyscraper
poster	palm	stove	path	minibike
coffee table	stool	armchair	van	car
food	towel	poster	jar	plant
earth	table	sidewalk	pillow	chandelier
stool	plant	grass	candlestick	fence
sidewalk	sky	coffee table	bucket	stairway
table	television receiver	television receiver	table	signboard
television receiver	poster	food	awning	palm
towel	floor	towel	umbrella	book
plant	pillow	fence	basket	sink
fence	signboard	plant	sofa	column
field	truck	path	car	candlestick
bannister	stairway	bannister	towel	basket
truck	bannister	shrub	shrub	umbrella
signboard	sidewalk	table	air conditioner	bannister

sofa	sea	mountain	streetlight	pillow
stairway	boat	stool	flag	telephone
awning	can	truck	mountain	towel
pillow	earth	awning	can	pot
grass	ashcan	signboard	monitor	can
air conditioner	fence	stairway	book	plaything
boat	bucket	traffic light	traffic light	wall socket
path	air conditioner	person	bannister	light
can	basket	car	fence	switch
traffic light	book	sofa	plant	jar
bucket	rug	boat	stairway	spotlight
ashcan	awning	pillow	television receiver	swivel chair
car	flag	bucket	coffee table	glass
floor	plaything	ashcan	earth	painting
book	candlestick	air conditioner	sidewalk	road
mountain	traffic light	flag	armchair	television receiver
flag	jar	monitor	poster	grass
monitor	stairs	minibike	stool	shrub
basket	magazine	can	food	table
shrub	monitor	book	signboard	streetlight
stairs	pot	van	truck	sidewalk
plaything	field	basket	sink	earth
van	car	shoe	grass	bucket
jar	glass	stairs	desk	poster
umbrella	streetlight	plaything	painting	figurine
candlestick	mountain	floor	chest of drawers	sky
shoe	path	umbrella	field	magazine
minibike	wall socket	jar	chandelier	air conditioner
streetlight	van	bicycle	railing	monitor
magazine	umbrella	rock	stove	person
bicycle	figurine	candlestick	sky	awning
rug	switch	magazine	swivel chair	stairs
pot	fluorescent	streetlight	column	fluorescent
fluorescent	light	pot	blind	van
figurine	telephone	figurine	palm	traffic light
glass	spotlight	trade name	sea	shoe
wall socket	grass	rug	curtain	ashcan
rock	shoe	glass	base	field
spotlight	bicycle	fluorescent	road	flag
light	rock	animal	computer	mountain
telephone	shrub	spotlight	fireplace	path
switch	minibike	wall socket	windowpane	animal
person	trade name	light	building	rock
trade name	person	telephone	bookcase	boat
animal	ball	switch	house	trade name
ball	animal	ball	skyscraper	sea

BIBILIOGRAPHY

Anderson, B. L. (2020). Mid-level vision. *Current Biology*, *30*(3), R105–R109. https://doi.org/10.1016/j.cub.2019.11.088

Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*, *364*(6439). https://doi.org/10.1126/science.aav9436

Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. (2017). *Network Dissection: Quantifying Interpretability of Deep Visual Representations*. 6541–6549. https://openaccess.thecvf.com/content_cvpr_2017/html/Bau_Network_Dissection_Quantifying_CVPR_2017_paper.html

Bau, D., Zhu, J.-Y., Strobelt, H., Lapedriza, A., Zhou, B., & Torralba, A. (2020). Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, *117*(48), 30071–30078. https://doi.org/10.1073/pnas.1907375117

Bonner, M. F., & Epstein, R. A. (2017). Coding of navigational affordances in the human visual system. *Proceedings of the National Academy of Sciences*, *114*(18), 4793–4798. https://doi.org/10.1073/pnas.1618228114

Bonner, M. F., & Epstein, R. A. (2018). Computational mechanisms underlying cortical responses to the affordance properties of visual scenes. *PLOS Computational Biology*, *14*(4), e1006111. https://doi.org/10.1371/journal.pcbi.1006111

Bonner, M. F., & Epstein, R. A. (2020). Object representations in the human brain reflect the co-occurrence statistics of vision and language. *BioRxiv*, 2020.03.09.984625. https://doi.org/10.1101/2020.03.09.984625

Cai, W., & Wei, Z. (2020). PiiGAN: Generative Adversarial Networks for Pluralistic Image Inpainting. *IEEE Access*, *8*, 48451–48463. https://doi.org/10.1109/ACCESS.2020.2979348

Chang, N., Pyles, J. A., Marcus, A., Gupta, A., Tarr, M. J., & Aminoff, E. M. (2019). BOLD5000, a public fMRI dataset while viewing 5000 visual images. *Scientific Data*, *6*(1), 49. https://doi.org/10.1038/s41597-019-0052-3

Chao, L. L., Haxby, J. V., & Martin, A. (1999). Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nature Neuroscience*, *2*(10), 913–919. https://doi.org/10.1038/13217

Cichy, Radoslaw M., & Kaiser, D. (2019). Deep Neural Networks as Scientific Models. *Trends in Cognitive Sciences*, *23*(4), 305–317. https://doi.org/10.1016/j.tics.2019.01.009

Cichy, Radoslaw Martin, Chen, Y., & Haynes, J.-D. (2011). Encoding the identity and location of objects in human LOC. *NeuroImage*, *54*(3), 2297–2307. https://doi.org/10.1016/j.neuroimage.2010.09.044

Cichy, Radoslaw Martin, Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*(1), 27755. https://doi.org/10.1038/srep27755

Coutanche, M. N., & Koch, G. E. (2018). Creatures great and small: Real-world size of animals predicts visual cortex representations beyond taxonomic category. *NeuroImage*, *183*, 627–634. https://doi.org/10.1016/j.neuroimage.2018.08.066

Cox, D. D. (2014). Do we understand high-level vision? *Current Opinion in Neurobiology*, *25*, 187–193. https://doi.org/10.1016/j.conb.2014.01.016

Cox, R. W. (1996). AFNI: Software for Analysis and Visualization of Functional Magnetic Resonance Neuroimages. *Computers and Biomedical Research*, *3*(29), 162–173. https://doi.org/10.1006/cbmr.1996.0014

Dolhansky, B., & Ferrer, C. C. (2018). *Eye In-Painting With Exemplar Generative Adversarial Networks*. 7902–7911. https://openaccess.thecvf.com/content_cvpr_2018/html/Dolhansky_Eye_In-Painting_With_CVPR_2018_paper.html

Eger, E., Ashburner, J., Haynes, J.-D., Dolan, R. J., & Rees, G. (2008). FMRI Activity Patterns in Human LOC Carry Information about Object Exemplars within Category. *Journal of Cognitive Neuroscience*, *20*(2), 356–370. https://doi.org/10.1162/jocn.2008.20019

El-Shamayleh, Y., & Pasupathy, A. (2016). Contour Curvature As an Invariant Code for Objects in Visual Area V4. *Journal of Neuroscience*, *36*(20), 5532–5543. https://doi.org/10.1523/JNEUROSCI.4139-15.2016

Epstein, R. A., & Baker, C. I. (2019). Scene Perception in the Human Brain. *Annual Review of Vision Science*, *5*(1), 373–397. https://doi.org/10.1146/annurev-vision-091718-014809

Epstein, R. A., & Morgan, L. K. (2012). Neural responses to visual scenes reveals inconsistencies between fMRI adaptation and multivoxel pattern analysis. *Neuropsychologia*, *50*(4), 530–543. https://doi.org/10.1016/j.neuropsychologia.2011.09.042

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, *392*(6676), 598–601. https://doi.org/10.1038/33402

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S.,

Wright, J., Durnez, J., Poldrack, R. A., & Gorgolewski, K. J. (2019). fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), 111–116. https://doi.org/10.1038/s41592-018-0235-4

Freeman, J., Brouwer, G. J., Heeger, D. J., & Merriam, E. P. (2011). Orientation Decoding Depends on Maps, Not Columns. *Journal of Neuroscience*, *31*(13), 4792–4804. https://doi.org/10.1523/JNEUROSCI.5160-10.2011

Grill-Spector, K. (2003). The neural basis of object perception. *Current Opinion in Neurobiology*, *13*(2), 159–166. https://doi.org/10.1016/S0959-4388(03)00040-0

Grill-Spector, K., Kourtzi, Z., & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Research*, *41*(10), 1409–1422. https://doi.org/10.1016/S0042-6989(01)00073-6

Groen, I. I. A., Silson, E. H., & Baker, C. I. (2017). Contributions of low- and high-level properties to neural processing of visual scenes in the human brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1714), 20160102. https://doi.org/10.1098/rstb.2016.0102

Groen, I. I., Greene, M. R., Baldassano, C., Fei-Fei, L., Beck, D. M., & Baker, C. I. (2018). Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *ELife*, *7*, e32962. https://doi.org/10.7554/eLife.32962

Habak, C., Wilkinson, F., Zakher, B., & Wilson, H. R. (2004). Curvature population coding for complex shapes in human vision. *Vision Research*, *44*(24), 2815–2823. https://doi.org/10.1016/j.visres.2004.06.019

Han, X., Zhong, Y., Cao, L., & Zhang, L. (2017). Pre-Trained AlexNet Architecture with Pyramid Pooling and Supervision for High Spatial Resolution Remote

Sensing Image Scene Classification. *Remote Sensing*, *9*(8), 848. https://doi.org/10.3390/rs9080848

Haxby, J. V. (2012). Multivariate pattern analysis of fMRI: The early beginnings. *NeuroImage*, *62*(2), 852–855. https://doi.org/10.1016/j.neuroimage.2012.03.016

Hayworth, K. J., & Biederman, I. (2006). Neural evidence for intermediate representations in object recognition. *Vision Research*, *46*(23), 4024–4031. https://doi.org/10.1016/j.visres.2006.07.015

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. *ArXiv:1512.03385 [Cs]*. http://arxiv.org/abs/1512.03385

Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Wicklin, C. V., & Baker, C. I. (2019). THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLOS ONE*, *14*(10), e0223792. https://doi.org/10.1371/journal.pone.0223792

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, *4*(11), 1173–1185. https://doi.org/10.1038/s41562-020-00951-3

Henriksson, L., Mur, M., & Kriegeskorte, N. (2019). Rapid Invariant Encoding of Scene Layout in Human OPA. *Neuron*, *103*(1), 161-171.e3. https://doi.org/10.1016/j.neuron.2019.04.014

Ibrahim, M. I. S. (n.d.). *Wavelet Based Approaches for Detection and Recognition in Ear Biometrics*. 153.

Jozwik, K. M., Kriegeskorte, N., Storrs, K. R., & Mur, M. (2017). Deep Convolutional Neural Networks Outperform Feature-Based But Not

Categorical Models in Explaining Object Similarity Judgments. *Frontiers in Psychology*, *8*, 1726. https://doi.org/10.3389/fpsyg.2017.01726

Julian, J. B., Ryan, J., & Epstein, R. A. (2017). Coding of Object Size and Object Category in Human Visual Cortex. *Cerebral Cortex*, *27*(6), 3095–3109. https://doi.org/10.1093/cercor/bhw150

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLOS Computational Biology*, *10*(11), e1003915. https://doi.org/10.1371/journal.pcbi.1003915

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K. A., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, *116*(43), 21854–21863. https://doi.org/10.1073/pnas.1905544116

Kim, J. G., Biederman, I., Lescroart, M. D., & Hayworth, K. J. (2009). Adaptation to objects in the lateral occipital complex (LOC): Shape or semantics? *Vision Research*, *49*(18), 2297–2305. https://doi.org/10.1016/j.visres.2009.06.020

Konkle, T., & Caramazza, A. (2013). Tripartite Organization of the Ventral Stream by Animacy and Object Size. *Journal of Neuroscience*, *33*(25), 10235–10242. https://doi.org/10.1523/JNEUROSCI.0983-13.2013

Konkle, T., & Oliva, A. (2012). A Real-World Size Organization of Object Responses in Occipitotemporal Cortex. *Neuron*, *74*(6), 1114–1124. https://doi.org/10.1016/j.neuron.2012.04.036

Kriegeskorte, N., & Douglas, P. K. (2018). Cognitive computational neuroscience. *Nature Neuroscience*, *21*(9), 1148–1160. https://doi.org/10.1038/s41593-018-0210-5

Lage-Castellanos, A., Valente, G., Formisano, E., & Martino, F. D. (2019). Methods for computing the maximum performance of computational models of fMRI responses. *PLOS Computational Biology*, *15*(3), e1006397. https://doi.org/10.1371/journal.pcbi.1006397

Large, M.-E., Aldcroft, A., & Vilis, T. (2007). Task-related laterality effects in the lateral occipital complex. *Brain Research*, *1128*, 130–138. https://doi.org/10.1016/j.brainres.2006.10.023

Long, B., Yu, C.-P., & Konkle, T. (2018). Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences*, *115*(38), E9015–E9024. https://doi.org/10.1073/pnas.1719616115

Long, J., Shelhamer, E., & Darrell, T. (2015). *Fully Convolutional Networks for Semantic Segmentation*. 3431–3440. https://openaccess.thecvf.com/content_cvpr_2015/html/Long_Fully_Convolutional_Networks_2015_CVPR_paper.html

Long Sha, Haxby, J. V., Abdi, H., Swaroop Guntupalli, J., Oosterhof, N. N., Halchenko, Y. O., & Connolly, A. C. (2015). The Animacy Continuum in the Human Ventral Vision Pathway. *Journal of Cognitive Neuroscience*, *27*(4), 665–678. https://doi.org/10.1162/jocn_a_00733

Malach, R., Reppas, J. B., Benson, R. R., Kwong, K. K., Jiang, H., Kennedy, W. A., Ledden, P. J., Brady, T. J., Rosen, B. R., & Tootell, R. B. (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proceedings of the National Academy of Sciences*, *92*(18), 8135–8139. https://doi.org/10.1073/pnas.92.18.8135

Marchette, S. A., Vass, L. K., Ryan, J., & Epstein, R. A. (2015). Outside Looking In: Landmark Generalization in the Human Navigational System. *Journal of Neuroscience*, *35*(44), 14896–14908. https://doi.org/10.1523/JNEUROSCI.2270-15.2015

Mehrer, J., Spoerer, C. J., Kriegeskorte, N., & Kietzmann, T. C. (2020). *Individual differences among deep neural network models* [Preprint]. Neuroscience. https://doi.org/10.1101/2020.01.08.898288

Mehrotra, R., Namuduri, K. R., & Ranganathan, N. (1992). Gabor filter-based edge detection. *Pattern Recognition*, *25*(12), 1479–1494. https://doi.org/10.1016/0031-3203(92)90121-X

Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, *73*, 1–15. https://doi.org/10.1016/j.dsp.2017.10.011

Morgan, L. K., MacEvoy, S. P., Aguirre, G. K., & Epstein, R. A. (2011). Distances between Real-World Locations Are Represented in the Human Hippocampus. *Journal of Neuroscience*, *31*(4), 1238–1245. https://doi.org/10.1523/JNEUROSCI.4667-10.2011

Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage*, *59*(3), 2636–2643. https://doi.org/10.1016/j.neuroimage.2011.08.076

Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., & Gallant, J. L. (2009). Bayesian Reconstruction of Natural Images from Human Brain Activity. *Neuron*, *63*(6), 902–915. https://doi.org/10.1016/j.neuron.2009.09.006

Nasr, S., Echavarria, C. E., & Tootell, R. B. H. (2014). Thinking Outside the Box: Rectilinear Shapes Selectively Activate Scene-Selective Cortex. *Journal of Neuroscience*, *34*(20), 6721–6735. https://doi.org/10.1523/JNEUROSCI.4802-13.2014

Nasr, S., & Tootell, R. B. H. (2012). A Cardinal Orientation Bias in Scene-Selective Visual Cortex. *Journal of Neuroscience*, *32*(43), 14921–14926. https://doi.org/10.1523/JNEUROSCI.2036-12.2012

Nayebi, A., Sagastuy-Brena, J., Bear, D. M., Kar, K., Kubilius, J., Ganguli, S., Sussillo, D., DiCarlo, J. J., & Yamins, D. L. K. (2021). Goal-Driven Recurrent Neural Network Models of the Ventral Visual Stream. *BioRxiv*, 2021.02.17.431717. https://doi.org/10.1101/2021.02.17.431717

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430. https://doi.org/10.1016/j.tics.2006.07.005

Nunez-Elizalde, A. O., Huth, A. G., & Gallant, J. L. (2019). Voxelwise encoding models with non-spherical multivariate normal priors. *NeuroImage*, *197*, 482–492. https://doi.org/10.1016/j.neuroimage.2019.04.012

Olshausen, B. A., & Field, D. J. (2000). Vision and the Coding of Natural Images: The human brain may hold the secrets to the best image-compression algorithms. *American Scientist*, *88*(3), 238–245.

Oord, Aäron van den, & Kalchbrenner, N. (2016). Pixel RNN. *ICML*.

Oord, Aaron van den, Kalchbrenner, N., & Kavukcuoglu, K. (2016). Pixel Recurrent Neural Networks. *ArXiv:1601.06759 [Cs]*. http://arxiv.org/abs/1601.06759

Park, J., & Park, S. (2017). Conjoint representation of texture ensemble and location in the parahippocampal place area. *Journal of Neurophysiology*, *117*(4), 1595–1607. https://doi.org/10.1152/jn.00338.2016

Park, S., Li, D. S. P., Shao, J., Lu, Z., & McCloskey, M. (2020). A scene with an invisible wall—The role of navigational experience in visual scene perception. *Journal of Vision*, *20*(11), 990–990. https://doi.org/10.1167/jov.20.11.990

Ponce, C. R., Hartmann, T. S., & Livingstone, M. S. (2017). End-Stopping Predicts Curvature Tuning along the Ventral Stream. *Journal of Neuroscience*, *37*(3), 648–659. https://doi.org/10.1523/JNEUROSCI.2507-16.2016

Rajimehr, R., Devaney, K. J., Bilenko, N. Y., Young, J. C., & Tootell, R. B. H. (2011). The "Parahippocampal Place Area" Responds Preferentially to High Spatial Frequencies in Humans and Monkeys. *PLOS Biology*, *9*(4), e1000608. https://doi.org/10.1371/journal.pbio.1000608

Rice, G. E., Watson, D. M., Hartley, T., & Andrews, T. J. (2014). Low-Level Image Properties of Visual Objects Predict Patterns of Neural Response across Category-Selective Regions of the Ventral Visual Pathway. *Journal of Neuroscience*, *34*(26), 8837–8844. https://doi.org/10.1523/JNEUROSCI.5265-13.2014

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Schmidt, K., Yamins, D. L. K., & DiCarlo, J. J. (2018). Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? *BioRxiv*, 407007. https://doi.org/10.1101/407007

Sha, L., Haxby, J. V., Abdi, H., Guntupalli, J. S., Oosterhof, N. N., Halchenko, Y. O., & Connolly, A. C. (2015). The Animacy Continuum in the Human Ventral

Vision Pathway. *Journal of Cognitive Neuroscience*, *27*(4), 665–678. https://doi.org/10.1162/jocn_a_00733

Shpaner, M., Molholm, S., Forde, E., & Foxe, J. J. (2013). Disambiguating the roles of area V1 and the lateral occipital complex (LOC) in contour integration. *NeuroImage*, *69*, 146–156. https://doi.org/10.1016/j.neuroimage.2012.11.023

Silson, Edward H., Groen, I. I. A., Kravitz, D. J., & Baker, C. I. (2016). Evaluating the correspondence between face-, scene-, and object-selectivity and retinotopic organization within lateral occipitotemporal cortex. *Journal of Vision*, *16*(6), 14–14. https://doi.org/10.1167/16.6.14

Silson, Edward Harry, Chan, A. W.-Y., Reynolds, R. C., Kravitz, D. J., & Baker, C. I. (2015). A Retinotopic Basis for the Division of High-Level Scene Processing between Lateral and Ventral Human Occipitotemporal Cortex. *Journal of Neuroscience*, *35*(34), 11921–11935. https://doi.org/10.1523/JNEUROSCI.0137-15.2015

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural Image Statistics and Neural Representation. *Annual Review of Neuroscience*, *24*(1), 1193–1216. https://doi.org/10.1146/annurev.neuro.24.1.1193

Srinath, R., Emonds, A., Wang, Q., Lempel, A. A., Dunn-Weiss, E., Connor, C. E., & Nielsen, K. J. (2021). Early Emergence of Solid Shape Coding in Natural and Deep Network Vision. *Current Biology*, *31*(1), 51-65.e5. https://doi.org/10.1016/j.cub.2020.09.076

Swisher, J. D., Gatenby, J. C., Gore, J. C., Wolfe, B. A., Moon, C.-H., Kim, S.-G., & Tong, F. (2010). Multiscale Pattern Analysis of Orientation-Selective Activity in the Primary Visual Cortex. *Journal of Neuroscience*, *30*(1), 325–330. https://doi.org/10.1523/JNEUROSCI.4811-09.2010

Szegedy, C., Toshev, A., & Erhan, D. (n.d.). *Deep Neural Networks for Object Detection*. 9.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network: Computation in Neural Systems*, *14*(3), 391–412. https://doi.org/10.1088/0954-898X_14_3_302

Troiani, V., Stigliani, A., Smith, M. E., & Epstein, R. A. (2014). Multiple Object Properties Drive Scene-Selective Regions. *Cerebral Cortex*, *24*(4), 883–897. https://doi.org/10.1093/cercor/bhs364

Walker, E. Y., Sinz, F. H., Cobos, E., Muhammad, T., Froudarakis, E., Fahey, P. G., Ecker, A. S., Reimer, J., Pitkow, X., & Tolias, A. S. (2019). Inception loops discover what excites neurons most using deep predictive models. *Nature Neuroscience*, *22*(12), 2060–2065. https://doi.org/10.1038/s41593-019-0517-x

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111

Yau, J. M., Pasupathy, A., Brincat, S. L., & Connor, C. E. (2013). Curvature Processing Dynamics in Macaque Area V4. *Cerebral Cortex*, *23*(1), 198–209. https://doi.org/10.1093/cercor/bhs004

Yeh, R. A., Chen, C., Yian Lim, T., Schwing, A. G., Hasegawa-Johnson, M., & Do, M. N. (2017). *Semantic Image Inpainting With Deep Generative Models*. 5485–5493.

https://openaccess.thecvf.com/content_cvpr_2017/html/Yeh_Semantic_Image_Inpainting_CVPR_2017_paper.html

Yue, X., Pourladian, I. S., Tootell, R. B. H., & Ungerleider, L. G. (2014). Curvature-processing network in macaque visual cortex. *Proceedings of the National Academy of Sciences*, *111*(33), E3467–E3475. https://doi.org/10.1073/pnas.1412616111

Yue, X., Robert, S., & Ungerleider, L. G. (2020). Curvature processing in human visual cortical areas. *NeuroImage*, *222*, 117295. https://doi.org/10.1016/j.neuroimage.2020.117295

Zhang, H., Hu, Z., Luo, C., Zuo, W., & Wang, M. (2018). Semantic Image Inpainting with Progressive Generative Networks. *Proceedings of the 26th ACM International Conference on Multimedia*, 1939–1947. https://doi.org/10.1145/3240508.3240625

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2018). Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(6), 1452–1464. https://doi.org/10.1109/TPAMI.2017.2723009

Zhou, Bolei, Zhao, H., Puig, X., Fidler, S., Barriuso, A., & Torralba, A. (2017). *Scene Parsing Through ADE20K Dataset*. 633–641. https://openaccess.thecvf.com/content_cvpr_2017/html/Zhou_Scene_Parsing_Through_CVPR_2017_paper.html

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. L. K. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, *118*(3). https://doi.org/10.1073/pnas.2014196118

CURRICULUM VITAE

Donald Li was born on December 20, 1993, in Hong Kong. He received his B.Eng. in Aerospace Engineering and Engineering Mechanics from Tsinghua University in July 2015. In August 2015, Donald joined the Cognitive Science Department at Johns Hopkins University under the mentorships of Drs. Michael Bonner, Brenda Rapp and Soojin Park.