

---

# Assessing the representation of seen and unseen contents in human brains and deep artificial networks

---

*Author:*

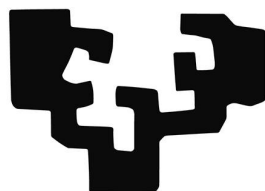
Ning Mei

*Supervisor:*

David Soto

Roberto Santana

YEAR 2022



UPV EHU

*A thesis submitted in fulfillment of the requirements  
for the degree of Doctor of Philosophy*

*in the*

*Basque Center on Cognition, Brain, and Language*

September 14, 2022



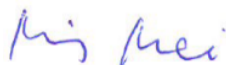
BASQUE CENTER  
ON COGNITION, BRAIN  
AND LANGUAGE

# Declaration of Authorship

I, Ning Mei, declare that this thesis titled, “Assessing the representation of seen and unseen contents in human brains and deep artificial networks” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:



---

Date: September 14, 2022

---

## *Acknowledgements*

This project would not have been possible without the support of many people. Many thanks to my advisors, David Soto and Roberto Santana, for their guidance and for the numerous discussions, revisions, and for helping me to make sense of the data. Also, I would like to express my sincere gratitude to David Soto who help me to enhance my basic knowledge and go deeper into the multidisciplinary research area of computational cognitive neuroscience. His patience and guidance led me to achieve several milestones in my predoctoral career. Also, special thanks to Pedro Margollés, Patxi Elozegi, and Usman Ayub Sheikh for offering generous help and suggestions during my experiments. Additionally, I want to give my best wishes to Sanjeev Nara and his families. Finally, I thank the Spanish Ministry of Science and Innovation for the financial support that allowed me to start this journey in the first place.

Part of this thesis has been already published with the following reference.

Mei, N., Santana, R., & Soto, D. (2022). Informative neural representations of unseen contents during higher-order processing in human brains and deep artificial networks. *Nature Human Behaviour*, 6(5), 720-731.

BASQUE CENTER ON COGNITION, BRAIN, AND LANGUAGE

*Abstract*

Doctor of Philosophy

**Assessing the representation of seen and unseen contents in human brains  
and deep artificial networks**

by Ning Mei

The functional scope of unconscious visual information processing and its implementation in the human brain remains a highly contested issue in cognitive neuroscience. The influential global workspace and higher-order theories predict that unconscious visual processing is restricted to representations in the visual cortex, which are not read-out further by parietal and prefrontal cortices. However, the prior work may lack sensitivity, with studies using a low number of trials per participant to pinpoint unconscious processing on an individual basis, and further lack a sound computational framework to understand the representation of unconscious knowledge. The present thesis employs functional magnetic resonance imaging (fMRI) and computational approaches to develop a high-precision, within-subject framework in order to define the properties of the brain representations of unconscious content associated with null perceptual sensitivity. Machine learning models were used to read-out multivariate unconscious content from fMRI signals throughout the ventral visual pathway, and model-based representational similarity analysis examined the properties of both conscious and unconscious representations. Finally, a wide range of feedforward convolutional neural network (FCNN) models was used to simulate the fMRI results, namely, to probe the existence of informative representations of visual objects associated with null perceptual sensitivity in artificial networks. The results show that even when human observers display null perceptual sensitivity at a behavioral level, there are neural representations of unconscious content widely distributed throughout the cortex, and these are not only contained in visual regions but also extend to higher-order regions in the ventral visual pathway, parietal and even prefrontal areas. The computational simulations with different FCNN models trained to perform the same visual task with noisy images demonstrated that even when the FCNN models failed to classify the category of the noisy images, the hidden representation of the FCNN models contained an informative representation that allowed for decoding of the image class. The implications of the results for models of visual consciousness are discussed. We also anticipate that the current study will inspire future work integrating neuroimaging, machine learning, and computational modeling to investigate the representation of conscious and unconscious knowledge across different cognitive domains.





# Contents

<b>Declaration of Authorship</b>	<b>i</b>
<b>Acknowledgements</b>	<b>ii</b>
<b>Related publications</b>	<b>iii</b>
<b>Abstract</b>	<b>v</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Theories of consciousness . . . . .	1
1.1.1 Global workspace theory of consciousness . . . . .	4
1.1.2 Recurrent processing theory of consciousness . . . . .	6
1.1.3 Higher-order theory of consciousness . . . . .	7
1.2 Behavioral evidence: Masking and Stimulus degradation . . . . .	8
1.3 M/EEG evidence . . . . .	9
1.4 fMRI evidence . . . . .	11
1.5 Controversy in unconscious processing research . . . . .	13
1.6 Lack of information-based approaches . . . . .	14
1.7 Objectives of the present thesis . . . . .	15
<b>2 Decoding unconscious and conscious contents</b>	<b>17</b>
2.1 Introduction . . . . .	17
2.1.1 Machine learning in cognitive neuroscience . . . . .	17
2.1.2 Limitations in previous decoding studies of unconscious processing .	18
2.2 Methods . . . . .	20

2.2.1	Participants . . . . .	20
2.2.2	Task and procedures . . . . .	20
2.2.3	Analysis of behavioral performance . . . . .	24
2.2.4	fMRI acquisition and preprocessing . . . . .	25
2.2.5	fMRI decoding pipeline within awareness states . . . . .	27
2.2.6	Generalization pipeline across awareness states . . . . .	28
2.3	Results . . . . .	29
2.3.1	Behavioral performance . . . . .	29
2.3.2	Decoding within each awareness state and generalization across awareness states . . . . .	33
2.4	Discussion . . . . .	40
<b>3</b>	<b>Assessing the representation of unseen contents using DNNs</b>	<b>42</b>
3.1	Introduction: Artificial deep neural network as a model of the human visual system at the representational level . . . . .	42
3.2	Methods . . . . .	44
3.2.1	A feedforward convolutional neural network simulation of unconscious processing at the representational level . . . . .	44
3.3	Results . . . . .	50
3.3.1	General results of the simulation models . . . . .	50
3.3.2	Robustness of the simulation models under various conditions . . . . .	57
3.4	Discussion . . . . .	62
<b>4</b>	<b>Neural signatures of conscious and unconscious contents</b>	<b>65</b>
4.1	Introduction: using information-based methods to study conscious and unconscious processing . . . . .	65
4.2	Methods . . . . .	66
4.2.1	Representations of the images from the state-of-the-art computer vision models . . . . .	66
4.2.2	Standard whole-brain searchlight RSA . . . . .	73

4.2.3	Encoding-based whole-brain searchlight RSA . . . . .	73
4.3	Results . . . . .	76
4.3.1	Standard whole-brain searchlight RSA . . . . .	76
4.3.2	Encoding-based whole-brain searchlight RSA . . . . .	78
4.4	Discussion . . . . .	84
<b>5</b>	<b>General Discussion</b>	<b>86</b>
5.1	Summary of findings and general conclusions . . . . .	86
5.1.1	High-precision, high-sampling fMRI data associated with null sensitivity reveal representations of unconscious contents . . . . .	87
5.1.2	A simulation of the fMRI results using feedforward convolutional neural network models . . . . .	87
5.1.3	Neural signatures of conscious and unconscious contents from an information-based perspective . . . . .	88
5.2	Integration of different results . . . . .	89
<b>6</b>	<b>Resumen amplio en castellano</b>	<b>94</b>

# List of Figures

2.1	Example trial of the fMRI experiment . . . . .	21
2.2	Example probe images . . . . .	23
2.3	Regions of interest . . . . .	26
2.4	Behavioral performance . . . . .	31
2.5	Decoding performance . . . . .	35
2.6	Additional decoding performance . . . . .	36
2.7	Robustness of decoding performance with a different penalty term . . . . .	38
2.8	Robustness of decoding performance with a different cross-validation procedure . . . . .	39
3.1	The feedforward convolutional neural network model . . . . .	47
3.2	Performance of the FCNN discriminating noisy images . . . . .	53
3.3	Performance of SVM decoding FCNN hidden representations when FCNN was at chance level . . . . .	54
3.4	Feature attribution of the FCNN to providing informative hidden representations . . . . .	55
3.5	Performance of the FCNN discriminating noisy images . . . . .	58
3.6	Difference of performance between SVM and FCNN . . . . .	59
3.7	Performance of SVM decoding FCNN hidden representations when FCNN was at chance level . . . . .	60
3.8	Performance of FCNN discriminating noisy images and the SVM decoding the first layer representations . . . . .	62
4.1	RDM of the AlexNet model . . . . .	68

4.2	RDM of the MobileNetV2 model . . . . .	69
4.3	RDM of the VGG19 model . . . . .	70
4.4	RDM of the ResNet50 model . . . . .	71
4.5	RDM of the DenseNet169 model . . . . .	72
4.6	General diagram of the encoding-based RSA pipeline . . . . .	75
4.7	RSA map of the unconscious condition for the five FCNNs . . . . .	77
4.8	RSA map of the conscious condition for the five FCNNs . . . . .	77
4.9	Encoding-based RSA map of the AlexNet model . . . . .	79
4.10	Encoding-based RSA map of the VGG19 model . . . . .	80
4.11	Encoding-based RSA map of the MobileNetV2 model . . . . .	81
4.12	Encoding-based RSA map of the ResNet50 model . . . . .	82
4.13	Encoding-based RSA map of the DensNet169 model . . . . .	83

# List of Tables

2.1	Behavioral performance . . . . .	32
3.1	Post-hoc test results regarding the feature importances . . . . .	56

# List of Abbreviations

<b>BCBL</b>	<b>B</b> asque Center on <b>C</b> ognition, <b>B</b> rain, and <b>L</b> anguage
<b>GNW</b>	<b>G</b> lobal <b>N</b> eural <b>W</b> orkspace
<b>GWT</b>	<b>G</b> lobal <b>W</b> orkspace <b>T</b> heory
<b>HOT</b>	<b>H</b> igher- <b>O</b> der <b>T</b> heory
<b>FFT</b>	<b>F</b> ast <b>F</b> ourier <b>T</b> ransformation
<b>MVPA</b>	<b>M</b> ultivariate <b>P</b> attern <b>A</b> nalysis
<b>MEG</b>	<b>M</b> agnetoencephalography
<b>EEG</b>	<b>E</b> lectroencephalography
<b>fMRI</b>	<b>F</b> unctional <b>M</b> agnetic <b>R</b> esonance <b>I</b> maging
<b>FCNN</b>	<b>F</b> eedforward <b>C</b> onvolutional <b>N</b> eural <b>N</b> etwork
<b>ReLU</b>	<b>R</b> etified <b>L</b> inear <b>U</b> nit
<b>ELU</b>	<b>E</b> xponential <b>L</b> inear <b>U</b> nit
<b>SELU</b>	<b>S</b> caled <b>E</b> xponential <b>L</b> inear <b>U</b> nit function
<b>SVM</b>	<b>S</b> upport <b>V</b> ector <b>M</b> achine
<b>PCA</b>	<b>P</b> rinciple <b>C</b> omponent <b>A</b> nalysis
<b>PFC</b>	<b>P</b> re- <b>F</b> rontal <b>C</b> ortex
<b>IT</b>	<b>I</b> nferior <b>T</b> emporal cortex
<b>RSA</b>	<b>R</b> epresentational <b>S</b> imilarity <b>A</b> nalysis
<b>RDM</b>	<b>R</b> epresentational <b>D</b> issimilarity <b>M</b> atrix
<b>ROC AUC</b>	<b>R</b> eceiver <b>O</b> perating <b>C</b> urve, <b>A</b> rea <b>U</b> nder the <b>C</b> urve
<b>TP</b>	<b>T</b> rue <b>P</b> ositive
<b>FP</b>	<b>F</b> alse <b>P</b> ositive
<b>TN</b>	<b>T</b> rue <b>N</b> egative

**FN**          **F**alse **N**egative



# List of Symbols

$d'$	d-prime
$meta - d'$	meta-d-prime
$\sigma$	variance
$\eta^2$	eta squared, measure of effect size
$\Delta$	difference between two variables
$\phi$	activation function
$\Sigma$	sum of items

# 1 Introduction

## 1.1 Theories of consciousness

It is very difficult to write down a general definition of consciousness. Despite recent methodological and conceptual development, the neuroscience of consciousness is still in its infancy. Consequently, different authors hold very different opinions on what it means to be conscious. So, as a first approach, we can simply bypass all these theoretical considerations and focus on everyday examples of what it is like to be conscious. For instance, if you ever used glasses, you know what it is like to go from not recognizing the face in front of you to immediately recognizing your friends' faces after wearing your lenses. Philosophers would say that there is something like seeing this person's face which is different from other possible experiences (Nagel, 1974). This subjective dimension of cognition is considered the basis of our conscious experience. Thus, the neuroscience of consciousness aims to explain the neurobiological basis of subjective experience - the personal stream of perceptions, thoughts, and beliefs that constitute our inner world.

In this regard, the conscious experience should not be confused with the state of alertness of being consciously awake. The first consciousness is used as a transitive verb, like in the case of "I am conscious of my back pain" or "I am aware of the smell of spring flowers". While, the second consciousness is used as an intransitive verb that does not require an object (i.e., the patient gains consciousness after anesthesia). In the present manuscript, the term "consciousness" therefore refers to the subjective experience of the external world. The conscious experience is always subjective and sometimes mixed with attended experience. The focus of the subjective experience aims to investigate whether

such experience can be measured quantitatively. When people fail to match their experience and the external world, they are “unconscious”. Introducing the notion of conscious perception implicitly assumes the existence of its contrary, namely, unconscious perception. Even though most researchers in Cognitive Neuroscience agree upon the existence of conscious perception based on their personal experience, there is less consensus regarding the existence of unconscious perception. In fact, it has been argued that understanding the distinction between conscious and unconscious information processing in the human brain is one of the main problems in Psychology and Neuroscience. For instance, the scope of the functional operations that can take place during unconscious information processing remains unclear (Dehaene, 2014).

A spectrum of paradigms (see review by Kim & Blake, 2005) allow us to test conscious and unconscious stimuli extensively in research labs (Baars, 1993). There are two distinctive kinds of unconscious stimuli, one is subliminal and the other is preconscious (Dehaene et al., 2006; Kanai et al., 2010). Subliminal stimuli are processed bottom-up with no top-down constraints, and researchers often flash them rapidly or mask them so that they are undetectable even with focused attention (Dehaene & Changeux, 2011). On the other hand, preconscious stimuli are visible at a certain level, but people are not consciously aware of them due to distraction or inattention (Dehaene & Changeux, 2011). The current thesis aims to test the scope of unconscious processing triggered by subliminal stimuli. However, due to the nature of introspective self-report of one’s conscious state, it is impossible to reliably confirm that every unconscious experience is only triggered by subliminal stimuli. Therefore, it is important to introduce objective behavioral measurements, such as those of signal detection theory (Macmillan & Creelman, 2004), which will be discussed later.

Recent research has suggested that people can maintain information in working memory even though they lack conscious access to it (i.e., King et al., 2016; Soto et al., 2011) and that “unconscious” working memory may engage prefrontal substrates to support the encoding and maintenance of non-conscious information (Dutta et al., 2014). Likewise, blindsight patients with lesions in the primary visual cortex appear able to respond to

and discriminate information in their impaired visual field for which they lack conscious experience (Weiskrantz, 2009) and a case-study using electroencephalography (EEG) documented brain responses to stimuli in the "blind" hemifield (Rossion et al., 2000). "Blindsight" and "unconscious" working memory processes are interesting phenomena because they indicate that subjective perceptual reports and objective discrimination performance can dissociate. However, the blindsight phenomena can only provide limited insights into the properties of unconscious information processing, because there remain fundamental issues to reliably differentiate unconscious from conscious processing. For instance, many studies, including those on "unconscious" working memory and blindsight cited above, employed subjective measures of perceptual awareness to separate non-conscious and conscious trials (Overgaard et al., 2010). However, subjective measures are unreliable because they are affected by decision biases. For instance, people may report a lack of awareness for otherwise conscious stimuli if they had very low confidence about what they were perceiving (Lau, 2008).

Understanding the distinction between conscious and unconscious information processing remains a key unresolved issue, but is paramount for developing a comprehensive neuroscientific account of consciousness and its role in cognition and behavior. Influential neurocognitive models of visual consciousness such as the global neuronal workspace model (Dehaene, 2014) propose that conscious awareness is associated with sustained activity in large-scale association networks involving the frontoparietal cortex, making information globally accessible to systems involved in working memory, report one's experience verbally, and behavioral control. Unconscious visual processing, on the other hand, is thought to be transient and operates locally in domain-specific systems, supporting low-level perceptual analysis (Lamme, 2020).

Recent studies have, however, confronted this view with intriguing data suggesting that unconscious information processing is implicated in higher-order operations associated with cognitive control (Van Gaal & Lamme, 2012), memory-guided behavior across both short- and long-term delays (Chong et al., 2014; Rosenthal et al., 2016; Soto et al., 2011; Trübtschek et al., 2017; Wuethrich et al., 2018), and also language computations

(Hassin, 2013). However, subsequent work has failed to support these findings (Rabagliati et al., 2018), and even the evidence for unconscious semantic priming has been recently called into question (Kouider & Dehaene, 2007; Stein et al., 2020). The limits and scope of unconscious information processing remain to be determined. This is critical for producing a contour map of information processing from unconscious to conscious activity (Dehaene, 2014).

In the next few sections (sections 1.1.1, 1.1.2, and 1.1.3), we will introduce some of the mainstream theories of consciousness.

### 1.1.1 Global workspace theory of consciousness

Baars (1993) first coined the psychological term "global workspace", which was inspired by its use in artificial intelligence, referring to a memory space that allows programs to solve problems in a collective fashion (Baars et al., 2021). Baars et al. (2021) pointed out that global workspace theory centers on answering this question of his influential paper: "How does a serial, integrated and very limited stream of consciousness emerge from a nervous system that is mostly unconscious, distributed, parallel, and of enormous capacity"?

To answer this question, Stanislas Dehaene and Jean-Pierre Changeux have developed many experimental paradigms to test different aspects of the global workspace theory of consciousness in the human brain (Dehaene & Changeux, 1989, 2000; Dehaene et al., 2001; Dehaene et al., 1998b; Naccache & Dehaene, 2001; Sergent et al., 2005; Sergent & Dehaene, 2004a, 2004b). In these studies, stimuli, such as letters (Naccache & Dehaene, 2001), numbers (Naccache & Dehaene, 2001), and words (Dehaene et al., 2001), were flashed briefly, masked, and primed, so that they were difficult to perceive.

Dehaene and Changeux (2005) presented a detailed biologically plausible model, global neuronal workspace (GNW), that illustrated that the "global workspace" was grounded

by long-distance cortical connections, “particularly those linking prefrontal cortex to associative areas of the parietal and cingulate cortices” (Dehaene & Changeux, 2005, section "Comparison with Other Theories of Consciousness", p. 0923). In other words, a conscious experience occurs when the long-range connected neurons ignite widespread whole-brain feedforward and feedback activation (Dehaene et al., 1998a; Sergent & Dehaene, 2004b). Unconscious contents/representations oscillate within the sensory cortices. These activities can be attentionally selected and broadcasted to the whole brain and then form conscious contents/representations (Dehaene, 2014). Many parts of the brain, such as the inferior temporal cortex, the middle frontal cortex, and the prefrontal cortex, have reciprocal neuronal projections with the sensory cortex. Therefore, conscious representations emerge from such feedforward and feedbackward networks. The ignition of the "global workspace" is a psychological marker of conscious access to the sensory contents, thus, when the ignition is absent, the sensory contents are subjectively unconscious, even though the contents form representations in the sensory cortices at some level.

Broadly speaking, the GNW proposes that a brain mechanism integrates input from the sensory cortices, the long-term memory system, the temporal cortex, and the prefrontal cortex to form a global workspace that shares all the information (Dehaene & Changeux, 2011, 2005; Dehaene et al., 1998a; Dehaene & Naccache, 2001; Dehaene et al., 2003). GNW assumes contents are encoded in the shared neural representational space, namely the "working space", and the long-distance connections between different areas of the brain allow the whole brain to be efficiently integrated, leading to the emergence of conscious contents. According to GNW, unconscious contents are bottom-up driven in the sensory cortices, and unconscious visual stimuli elicit small magnitudes of neural activity that are not projected to the shared neural representational space. Thus, unconscious neural representations are rapidly triggered and processed by the low-level sensory cortex, but quickly vanish in a few milliseconds (Dehaene & Changeux, 2011). GWT predicts that unseen and unconscious processing stop at the early perceptual level (Dehaene et al., 2014). Thus, no brain activity related to the unconscious representations should be observed beyond the primary visual cortex (i.e., prefrontal -PFC-) if this theory is correct.

Also, the GWT does not predict that the representation of seen contents is similar to that of unseen contents. This will be tested in the present thesis.

### 1.1.2 Recurrent processing theory of consciousness

The recurrent processing theory (Beck et al., 2001; Goebel et al., 2001; Lamme, 2010; Lamme et al., 2002) suggests consciousness emerges when recurrent top-down and bottom-up activities occur in sensory areas of the brain such as early visual cortex.

The recurrent processing theory proposes that the brain first processes the input signals with a "feedforward sweep" (Lamme, 2006). The "feedforward sweep" rapidly passes information through the visual areas and is further passed to the higher areas, such as the motor cortex and the frontoparietal areas (i.e., for a potential response). The "feedforward sweep" information might not lead to conscious contents if the information stays in the early sensory areas, such as the primary visual cortex (Dehaene et al., 2006; Lamme, 2010). A deeper "feedforward sweep" that involves the motor areas and the frontoparietal areas, where the stimulus is masked and invisible, may influence the responses (i.e., unconscious priming, see the review of Dehaene et al., 1998b; Lamme, 2010) but not lead to conscious awareness. Without recurrent feedback within the sensory areas, information remains unconscious. Therefore, the stage of information processing known as "localized recurrent processing" (Lamme, 2006) is critical for the processed information to become conscious.

When the input signals in the primary sensory areas are strong enough to evoke recurrent processing within the sensory areas, the input image signals emerge into consciousness and can be reported and acted upon (i.e., through verbal or motor responses). Thus, recurrent processing, even only within the sensory cortex, allows the second or third waves of "feedforward sweep" information to reach higher areas (Lamme, 2010), and also allows dynamic interactions between areas so that object recognition reaches semantic/concept level (Lamme et al., 2000; Lamme et al., 1998; Lamme et al., 1993).

### 1.1.3 Higher-order theory of consciousness

Rosenthal (2004, P. 24) pointed out that “we are introspectively conscious of a state when we are not simply aware of that state, but aware of it in a deliberate, attentively focused way”. The higher-order theory of consciousness (Rosenthal, 1993) describes the conscious experience as a subjective mental state that is supported by not only first-order representations from the primary sensory cortex, but also some higher-order mechanisms from higher-order regions, such as the prefrontal areas (Brown et al., 2019). First-order representations can arise when the sensory cortices perceive external stimuli, and their processing is automatic (Brown et al., 2019). However, neuroimaging studies in blindsight patients (Persaud et al., 2011) and relative blindsight in normal observers (Lau & Passingham, 2006) show that “perceptual awareness is better correlated with activity in brain areas responsible for high-level cognition rather than in early sensory regions” (Brown et al., 2019, section "Current Status of HOT"). Both global workspace theory and higher-order theory argue that consciousness emerges from additional processes on top of early sensory perceptions. However, in global workspace theory, the early sensory representations are boosted and stabilized to enter the "global workspace". Provided the sensory representations are globally broadcast, the representations are conscious contents. Whereas in higher-order theory, inner awareness is represented by higher-order thoughts (i.e., meta-representations of sensory perceptions, Brown et al., 2019), thus, such higher-order representations, particularly in the pre-frontal cortex (PFC), mark the distinction between consciousness and unconscious processing. Similar to the global workspace theory, the higher-order theory predicts no brain activity related to the unconscious contents should be observed beyond the primary visual cortex in the current thesis.



## 1.2 Behavioral evidence: Masking and Stimulus degradation

The history of visual masking can be traced back to the 1960s (Breitmeyer et al., 2006; Kouider & Dehaene, 2007). This paradigm has been used for investigating unconscious stimuli, visual limits, and perceptual impairments (Bachmann & Francis, 2013). One frequently used masking paradigm is simply perceptual and involves showing the probe stimulus (i.e., an object or a word) followed by a mask (Breitmeyer, 2007). and participants are required to make a perceptual choice on the stimulus. Another type of masking paradigm is masked priming, which involves masking the probe stimulus and then presenting a stimulus that is related or unrelated to the probe, and asking participants to report if the presented stimulus was seen or unseen (Wickens, 1973). Marcel et al. (1974) showed that after briefly showing a word followed by a pattern mask, participants could correctly report perceived words. In follow-up studies, Marcel and colleagues (Marcel, 1980, 1983) demonstrated that the masking paradigm elicited automatic unconscious visual processing. In these studies, a prime word was masked and then two probe words were presented. One probe word was related to the prime word and the other word was unrelated. Participants were asked to choose one of the probe words and they chose the related one on 60% of the trials. These studies showed that unconscious processing reached the semantic level (Kouider & Dehaene, 2007).

Continuous flash suppression (or CFS) is also another form of masking that is used for studying conscious and unconscious processing (Pournaghdali & Schwartz, 2020). It involves using a stereoscope to present images to each eye separately (Tsuchiya & Koch, 2004). While presenting the probe image to one eye, colored mondrian-like pattern images are presented to the other eye, hence masking the probe image presented in the other eye. Many subjects reported they did not see the probe images even though the probe images were presented for a long time, but still observed that the masked content could be unconsciously processed and influence behavioral performance (Tsuchiya & Koch, 2005).

Many studies have investigated the boundary between unconscious and conscious processing using masking, CFS, and priming (Breitmeyer et al., 2005; Breitmeyer et al., 2004; Dehaene et al., 2004; Dehaene et al., 1998b; Evett & Humphreys, 1981; Fahrenfort et al., 2007; Greenwald et al., 1996; Humphreys et al., 1988; Humphreys et al., 1982; Koechlin et al., 1999; Lamme et al., 2002; Naccache et al., 2002; Suzuki & Fukuda, 2013; Tsuchiya & Koch, 2004, 2005). However, these studies have been extensively debated and questioned regarding the validity of the measures of awareness (Cheesman & Merikle, 1984). The caveats regarding the use of the subjective and objective measure of conscious awareness will be discussed in section 1.5, but in short, it is concerned that a weak but above chance level sensitivity of the stimuli may raise conscious awareness (Soto et al., 2019), and this raises a discrepancy between subjective and objective measurements of consciousness. The subjective measurement is self-report while objective measurement involves post-hoc signal detection theory tests (see section 1.5). Cheesman and Merikle (1986) argued that the ease of distinguishing between subjective consciousness and unconsciousness depends on the form of the processing being investigated. This will be further discussed in section 1.5.

### 1.3 M/EEG evidence

Magnetoencephalography (MEG) (Hämäläinen et al., 1993) and electroencephalography (EEG) (Henry, 2006) are non-invasive methods to study brain functions through the recording and analysis of electromagnetic field potentials in the scalp. Due to the precision of the temporal resolution, M/EEG are often used to study the distinction between unconscious and conscious processing. King et al. (2016) masked a briefly presented Gabor patch (tilted left or right) and then, following a delay period, they presented a probe Gabor patch. They asked the subjects to compare the orientation of masked and probe Gabor patches and also rate their visibility of the masked Gabor patch. The discrimination performance of the trials in which subjects reported having no awareness whatsoever of the masked Gabor (unseen trials) was significantly above the chance level.

Using multivariate pattern analyses of the MEG signals, they showed that the presence (or absence) of the Gabor, its rotation and contrast levels, the phase, and the spatial frequency of the masked unseen Gabor patches could be decoded by machine learning algorithms significantly above chance levels across the delay period before the onset of the visible probe. Multivariate pattern analysis (decoding) applied to neural data from the masking paradigms can be more sensitive to detecting unconscious processing signals. In an eight-location masking experiment, the target line segment was post-masked (Salti et al., 2015). The performance of identifying the reportedly "unseen" line segments was way higher than the chance level. The location of "unseen" segments can be decoded from MEG and EEG signals for up to 800 ms. Additionally, another MEG masking study showed that the increases in neural activities in higher-order visual areas correlated with conscious visual perception (Fisch et al., 2009).

Moving beyond the study of low-level perceptual features, Van Gaal et al. (2014) applied a spatial negation paradigm with EEG to investigate the cognitive processing of grammar rules that were traditionally thought to be slow and require conscious access (Deutsch et al., 2006). The key aspects of the stimuli are a modifier (i.e., "not" or "very"), an adjective word (i.e., "good"), and a target noun (i.e., "peace"). The modifier and the adjective word were presented together briefly, and they were pre- and post-masked. The discrimination performance of the masked words was at chance (26.4%) compared to unmasked words (75.1%). Van Gaal et al. (2014) found that the N400 effect of the masked words was remarkably similar to the unmasked words.

M/EEG techniques provide excellent temporal resolution of cognitive functions, however, they lack neuroanatomical precision to detect where in the brain the effect happens. An important question in the study of consciousness is to define whether unconscious visual contents are merely represented in the visual and/or whether they also involve higher-order areas that are relevant for decision-making and working memory functions, namely, the frontoparietal areas. Below, I review the neuroimaging evidence regarding this issue.

## 1.4 fMRI evidence

Functional magnetic resonance imaging (fMRI) is used for studying blood oxygen-level dependent brain responses triggered by stimuli and tasks (Ekstrom, 2010). fMRI has been a dominant neuroimaging method to study maps of brain activity and the spatial correlation between different brain areas in different cognitive functions. Using the word masking paradigm, Dehaene and colleagues showed that when the words were masked and invisible, participants recognized the masked words at chance level (Dehaene et al., 2001), and weak brain activations were only observed in the visual areas, such as the fusiform gyrus. When the words were visible, participants recognized the words successfully, and there were activations across the visual cortex as well as in the parietal areas. The activations were much stronger than those in the unconscious condition. Diaz and McCarthy (2007) conducted two masking experiments similar to Dehaene et al. (2001). In one of the experiments, subjects first decided whether a visible stimulus was a real-word or a non-word. And then a masked stimulus that could be a real-word or a non-word. The masked real-word could be semantically related to the visible real-word or not related. The responses to the masked, semantically related words were significantly faster. In the other experiment, which was an fMRI experiment, subjects were presented with masked real-words and non-words. Subjective and objective measurements of consciousness were used to ensure subjects were not aware of the masked stimuli. They found that the masked real-words elicited significantly greater responses in a distributed network of the left hemisphere, “including the inferior frontal gyrus, the angular gyrus, and posterior regions of the lateral temporal cortex”, compared to the masked non-words. In a recent fMRI study using masked words, multivariate patterns of the unseen words could be decoded by machine learning algorithms from the fMRI data in areas including the visual and inferior frontal areas (Sheikh et al., 2019). In another study, Stein et al. (2021) presented masked images of faces and houses and sampled trials as subjectively invisible (based on participants’ awareness reports) and objective invisible conditions (based on signal detection measures). The perceptual sensitivity in the subjectively invisible trials was greater than

zero. The objective invisible trials had a fixed mask contrast and  $d'$  of these trials were not different from zero. A localizer scan in which the different clear images were presented was collected to assess fMRI patterns and correlate the subjective and objective invisible fMRI patterns. The correlation between the localizer and the subjective invisible trials showed the role of the fusiform gyrus in processing the invisible contents, but as expected the correlation between the localizer and the objective invisible trials was weaker, likely due to stronger masking. Further, the correlation between the localizer and the subjective invisible trials showed patterns of unseen contents that extended to the inferior temporal gyrus.

In line with the above study, Sterzer et al. (2008) used a CFS fMRI paradigm in which faces and houses were presented only to one eye and masks were presented to the other eye to facilitate suppression of the faces and houses ( $N = 4$ ). Subjects were at chance of discriminating between faces and houses ( $d' = -0.05 \pm 0.14$ ), but these unseen contents were decoded from the fMRI data in areas such as the fusiform face area and parahippocampal place area better than chance.

Jiang et al. (2007) recruited five subjects and presented flickers at different frequencies (5 Hz at full contrast, 5 Hz at subthreshold contrast, and 30 Hz). Behaviorally, subjects clearly viewed the 5 Hz flickers at full contrast, but their responses to the 5 Hz flickers at sub-threshold contrast as well as the 30 Hz flickers were at chance by an off-line post-hoc test. Interestingly, the 30 Hz flickers elicited stronger brain activities in V1 through V4 compared to the 5 Hz flickers at subthreshold. They argued that stronger brain responses in the visual areas did not suggest conscious experience.

The fMRI studies (Dehaene et al., 2001; Sheikh et al., 2019; Stein et al., 2021; Sterzer et al., 2008) provide a more precise spatial resolution of the brain areas associated with unconscious and conscious information processing. However, the behavioral measurements of awareness and lack of awareness in these fMRI studies do not follow a consistent criterion. This will be discussed further in section 1.5.

## 1.5 Controversy in unconscious processing research

As mentioned in the previous sections, there remain significant hurdles in studying unconscious processing. For instance, in "subliminal priming" studies (i.e., Kouider and Dehaene, 2007), a briefly presented stimulus (the prime) is "masked" and followed by a visible target stimulus, and the subjects must respond to the category of the target. The masked prime can be congruent or incongruent with the response to the target thereby affecting the response latencies to the targets. Subliminal priming is inferred if, in a subsequent post-hoc offline perceptual test, the subjects can not distinguish the identity of the masked prime presented alone in a forced-choice test. However, in masked priming studies, no awareness measures are collected during the priming task. Considering that perceptual thresholds may vary across time, this leaves unclear the extent to which subjects have some degree of awareness in some trials hence leading to the behavioral priming effect. Studies of unconscious processing, in particular priming studies, do not include both subjective, trial-by-trial reports of awareness, and also objective criteria to pinpoint unconscious and consciousness states. Hence, it is difficult to know (i) whether or not subjects were aware of the prime in some trials and (ii) whether the perceptual threshold associated with null prime identification in the post-hoc test was similar to the perceptual threshold during the priming task (e.g. due to practice or fatigue effects).

Some studies report either subjective awareness or objective signal detection measures (see Dehaene et al., 2001; Salti et al., 2015) but it is important to integrate both subjective and objective behavioral measurements to provide robust evidence of null awareness and sensitivity for masked stimuli at the behavioral level (Soto et al., 2019). While subjective reports are essential for experimenters to acknowledge subjects' conscious states on a moment-to-moment basis, signal detection measures (Macmillan & Creelman, 2004) are critical to demonstrating null perceptual sensitivity at the behavioral level (Soto et al., 2019). As pointed out by Newell and Shanks (2014), both subjective reports and offline measures of perceptual sensitivity considered separately are of little use for understanding unconscious information processing. Thus, it seems clear that signal detection theory

(SDT) measures of objective perceptual performance (i.e.,  $d'$ ) collected online during the critical task (Macmillan & Creelman, 2004), in combination with trial-by-trial subjective reports, are best suited to demonstrate whether or not observers lack awareness of the target stimulus (i.e., the null sensitivity criterion, Newell and Shanks, 2014).

## 1.6 Lack of information-based approaches

Combining both subjective and objective measures (i.e., the null sensitivity at the behavioral level on trials reported as unaware, Soto et al., 2019) to study researchers were able to advance our understanding of unconscious processing (Lamme et al., 2002; Soto et al., 2011). The representation of unconscious processing may be delineated based on the information patterns extracted with visible contents through fine-grained analyses of brain data (i.e., M/EEG and fMRI), which are then extrapolated to the context of null perceptual sensitivity. Finding robust brain signals of information processing in the presence of null behavioral sensitivity could be the neural marker of unconscious processing. Studies mentioned in sections 1.3 and 1.4 have provided insights into the distinction between conscious and unconscious processing (Dehaene et al., 2001; King et al., 2016; Kouider & Dehaene, 2007; Salti et al., 2015; Stein et al., 2021). However, many of the studies did not demonstrate the null sensitivity at a behavioral level (see King et al., 2016), and could not draw robust inferences about unconscious processing (see Stein et al., 2021) due to a lack of statistical power. To address the latter matter, one should turn to information-based approaches (Kriegeskorte et al., 2006) to study multivariate patterns in the brain signals while collecting a high number of trials (i.e. over 1000 trials) per participant. This will allow determining the presence of null sensitivity while exploiting machine learning (Haxby et al., 2001), pattern analyses (Kriegeskorte, 2011; Kriegeskorte et al., 2008a), and computational models (Kriegeskorte, 2015) to significantly improve the reliability and reproducibility of the results in studying unconscious processing. In addition, by exploiting machine learning algorithms, representational dissimilarity, and computational models as information-based approaches, one could measure the level of representational invariance

between different levels of (un)conscious processing by investigating the generalization of the brain patterns learned by the models across the different states of (un)awareness (Soto et al., 2019). This approach allows the re-use of learned representations associated with the conscious state to inform the unconscious representation.

## 1.7 Objectives of the present thesis

Considering the neuroimaging evidence to date, the extent to which unconscious information associated with null perceptual sensitivity can permeate different stages of brain processing is unclear. The standard approach to neuroimaging studies of unconscious processing is limited because it lacks high within-subject precision. Previous studies used group-based statistical inference with few participants, and crucially also, very low numbers of trials per participant, which makes it underpowered to test the null sensitivity assumption and pinpoint the brain representation of unconscious knowledge reliably within each subject. Here we present a computational framework that circumvents these limitations.

A key objective of this thesis is to develop sensitive, within-subject neural markers of the scope of unconscious and conscious information processing. Here we use a high-precision, highly-sampled, within-subject computational approach to define the properties of the brain representations associated with unconscious perceptual content, defined by the absence of null perceptual sensitivity. In order to ensure null sensitivity, researchers must make a trade-off with the risk of being unable to detect brain signals of unconscious information processing (Soto et al., 2019). Therefore, machine learning algorithms and computational models are alternative tools to provide an information-based approach (Kriegeskorte et al., 2006) to test the existence of unconscious representations in brain activity patterns with higher sensitivity. This is something that traditional mass-univariate brain mapping approaches (Friston et al., 1995; Friston et al., 1994; Maris & Oostenveld, 2007) cannot achieve.



---

In addition, a further objective of this thesis is to provide information-based approaches, including a dense sampling of the brain signals, machine learning algorithms for pattern analysis, and computational models based on deep neural networks serving as encoding models to explain brain activity. These approaches provide us with very fine-grain neural representations of conscious and unconscious contents.

The final objective of this thesis is to provide a computational simulation of the representation of unseen contents using state-of-the-art deep neural network models (Kriegeskorte, 2015; LeCun & Bengio, 1995; LeCun et al., 2015; Lindsay, 2021). The simulation results allow us to understand better whether and how the representations of unseen contents differ from seen contents in both the human brain and artificial intelligence models of the human brain.

## 2 Decoding unconscious and conscious contents from fMRI data

### 2.1 Introduction

#### 2.1.1 Machine learning in cognitive neuroscience

Machine learning is an important component of data science. It uses statistical algorithms (i.e., gradient descent Ruder, 2016) to learn key aspects of the data, such as classifying categories or revealing structures in the data. Machine learning models often use different algorithms that have different data assumptions to learn the key aspects from the data based on predefined questions. For instance, classification models are often used in studying the problems from the level of neuronal (Li et al., 2015) to neuroimaging (Haxby et al., 2001) to behaviors (Mei et al., 2020), and this method is often used to illustrate whether there exists information in regions of the brain according to the experimental conditions. When a model maps the brain signals to the experimental conditions, it is a "decoding model"; while a model that maps the experimental information to the brain signals, it is an "encoding model" (Kriegeskorte & Douglas, 2019).

Decoding (Haxby et al., 2001) and encoding (Naselaris et al., 2011) are modeling approaches that have been used in making sense of neuroimaging data such as fMRI (Kriegeskorte & Douglas, 2019). Machine learning algorithms (Jordan & Mitchell, 2015) are the core of the decoding and encoding approaches. These algorithms often utilize linear models, such as logistic regression (King et al., 2016; Wright, 1995) and linear support vector machine (Bhavsar & Panchal, 2012; Sheikh et al., 2019) to decode the experimental conditions from brain signals or encode experimental information to brain

signals (Naselaris et al., 2011). The performance of the decoding models is computed using cross-validation and it is often used as the measure of experimental information that is contained in the multivariate pattern of brain signals.

Computational models, a subset of machine learning algorithms, are often used as alternative models to explain how information is represented in the brain. Particularly, in the field of computer vision, models like HMAX (Riesenhuber & Poggio, 1999) and convolutional neural network (CNN, Fukushima, 1980; Hinton et al., 2015; LeCun and Bengio, 1995; McFee et al., 2018) are used to explain how visual information is processed in the primary visual cortex. CNNs are now better models for predicting brain signals in several benchmarking visual tasks (Schrimpf et al., 2020), and this makes them the best alternative encoding models of visual processing and object recognition (Kriegeskorte, 2015).

Distance-based measurements such as Euclidean distance Kriegeskorte et al., 2006 are also used to compare the distance between pairs of activity-patterns. The distances indicate the relative similarities among pairs of activity-patterns from different levels of analyses (i.e. brain, computational model, across species) and different principles of information processing (see Chapter 4).

Additionally, computational models, such as CNNs, can be used to model how an ideal observer performs a visual processing task (i.e., object recognition), approximating the visual information being processed in the human brain, which is often referred to as simulation-based approaches (Kriegeskorte, 2015; Kriegeskorte and Douglas, 2018, see Chapter 3).

### **2.1.2 Limitations in previous decoding studies of unconscious processing**

There is extensive literature on using machine learning to learn about how the human brain processes information (Haxby et al., 2001). Particularly, multivoxel pattern analysis (MVPA), which is also called "decoding", maps the multivariate brain patterns to the

experimental conditions. It is popularized in studying the neural correlates of unconscious processing of simple visual patterns in the visual cortex (Haynes, 2009), such as grating patterns (King & Dehaene, 2014; Salti et al., 2015), letters (Trübtschek et al., 2017), and words (Sheikh et al., 2019).

Considering the available neuroimaging evidence to date, the extent to which unconscious information associated with null perceptual sensitivity can permeate different stages of brain processing is unclear. The standard approach to neuroimaging studies of unconscious processing is limited because it lacks high within-subject precision. Previous studies used group-based statistical inference but low sample sizes in terms of subjects, and crucially also, very low numbers of trials per participant, making it hard to test the null sensitivity assumption and pinpoint the brain representation of unconscious knowledge reliably within each subject (also see section 1.6).

In this chapter, we introduced a within-subject design fMRI experiment ( $N = 7$ ; over 1700 trials per participant) (i) to shed new light on the scope of unconscious information processing in the human brain; (ii) to reveal the properties of the brain representations during unconscious information processing and the extent to which they are similar to the conscious counterparts. We used a standard linear classifier applied to the fMRI signals to decode the categories of living and non-living images presented at different levels of visibility and assessed generalization for "out-of-sample" images (i.e., cat, bicycle) not used during classifier training, and, critically, we assessed generalization across different states of awareness within-subjects utilizing a large number of instances. It is unclear whether conscious and non-conscious items are represented similarly in the human brain. We examined how much information we could read-out and generalize from areas in the ventral visual pathway and frontoparietal cortex. Using a highly-sampled within-subject design, we tested the main hypothesis that visual and semantic representations of unconscious items in the ventral visual pathway, and perhaps also in other higher-order nodes of the frontoparietal network, can share similar properties to their conscious counterparts.

## 2.2 Methods

### 2.2.1 Participants

We recruited seven subjects from the Basque Center on Cognition, Brain, and Language, one of the subjects was female. Subjects agreed to participate in six fMRI scanning sessions by signing the BCBL fMRI consent agreement. Subjects had normal visions or corrected visions.

### 2.2.2 Task and procedures

In the fMRI experiment, subjects ( $N = 7$ , one female) were presented with images of animate and inanimate objects (Moreno-Martínez & Montoro, 2012). We selected 96 unique items (48 animate and 48 inanimate, i.e., cat, boat) for the experiment. These images could also be grouped into 10 subcategories (i.e., animal, vehicle) and 2 categories (i.e., animate/living v.s. inanimate/nonliving). The experiment was an event-related design. On each day, subjects carried out nine blocks of 32 trials (16 animate and 16 inanimate) and every three blocks contained a full set of all the unique items (96 unique images). Subjects came back six times within two weeks. There were hence 288 trials per day and 1728 trials in total per subject. The probe images were gray-scaled and presented in different orientations. In other words, the original images were augmented before the experiments using Tensorflow-Keras (Abadi et al., 2015; Chollet et al., 2018). A random-phase noise background generated from the images was added to the target image before the experiment to facilitate masking. The random-phase noise background was generated by converting the image to the Fourier domain using fast-Fourier-Transformation (FFT), and then we added Gaussian noises to the Fourier domain values. In the end, we inverted the FFT to get the random-phase noise background.

The experiment was programmed using Psychopy (Peirce, 2007). The experiment was carried out on a monitor with a refresh rate of 100 Hz. Figure 2.1 illustrates the sequence of events on each trial. A fixation point appeared for 500 *ms* and was followed by a

blank screen for 500 *ms*. Twenty frames of Gaussian noise masks were presented and then followed by the probe image, which was followed by another twenty frames of masks. Then, there was a jittered blank period (1500 - 3500 *ms*) with a pseudo-exponential distribution in 500 *ms* steps (sixteen 1500 *ms*, eight 2000 *ms*, four 2500 *ms*, two 3000 *ms*, and two 3500 *ms*) selected randomly and without replacement on each block of 32 trials. Following the jittered blank period, subjects were required (i) to identify and respond to the category of the image and (ii) to rate the state of visual awareness associated with the image. There was a 1500 *ms* deadline for each response and subjects were allowed to change their responses during this time window. Only the last response was recorded. For the categorization decision task, two choices were presented on the screen - living (V) and nonliving (nV) - i.e., "V nV" or "nV V", and the left-right order of the choices was randomly selected for each trial. Subjects pressed "1" (left) or "2" (right) to indicate the probe condition. For the awareness decision task, there were 3 choices: (i) "I did not see anything that allowed me to categorize the item, I was completely guessing"; henceforth, the unconscious trials, (ii) partially unconscious ("I saw a brief glimpse but I am not confident of the response"), and (iii) conscious ("I saw the object clearly or almost clearly and I am confident of the categorization decision"). The inter-trial interval was then followed with a jittered blank period of 6000 - 8000 *ms* with a pseudo-exponential distribution in 500 *ms* steps.

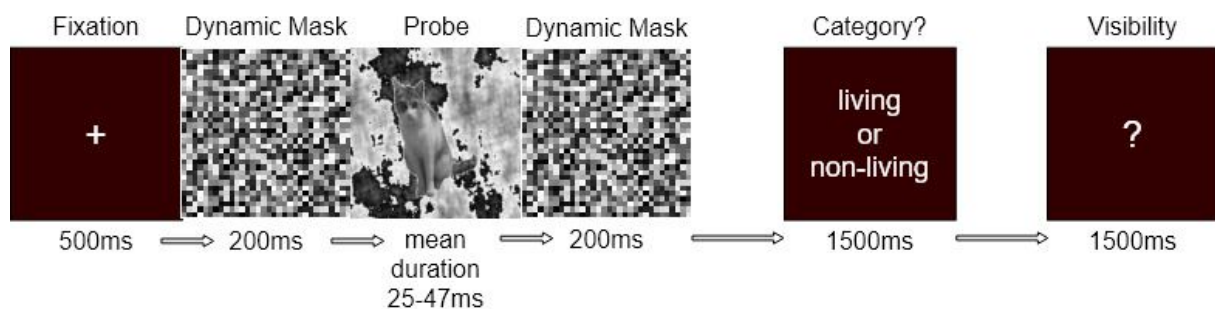


FIGURE 2.1: Example of the sequence of events within an experimental trial. Subjects were asked to discriminate the category of the masked image (living vs. non-living, categorization decision task) and then rate their visual awareness on a trial-by-trial basis. Example of a trial with a masked image of a cat. The stimuli were selected and augmented from an open source image set (Moreno-Martínez & Montoro, 2012).

The duration of the probe image was based on an adaptive staircase that was running

---

throughout the trials. Specifically, based on pilot tests, we elected to use the staircase to get a high proportion of unconscious trials while ensuring that perceptual sensitivity was not different from chance level. If the subject reported "glimpse", the number of frames of stimulus presentation was reduced by one frame for the next trial, unless it was already only one frame of presentation; if the subject reported "conscious", the number of frames of presentation would be reduced by two or three frames for the next trial, unless it was less than two to three frames, in which case it would be reduced by one frame; if the subject reported "unconscious", the number of frames of presentation increased by one or two frames for the next trial randomly. Examples of probe images were shown in Figure 2.2.

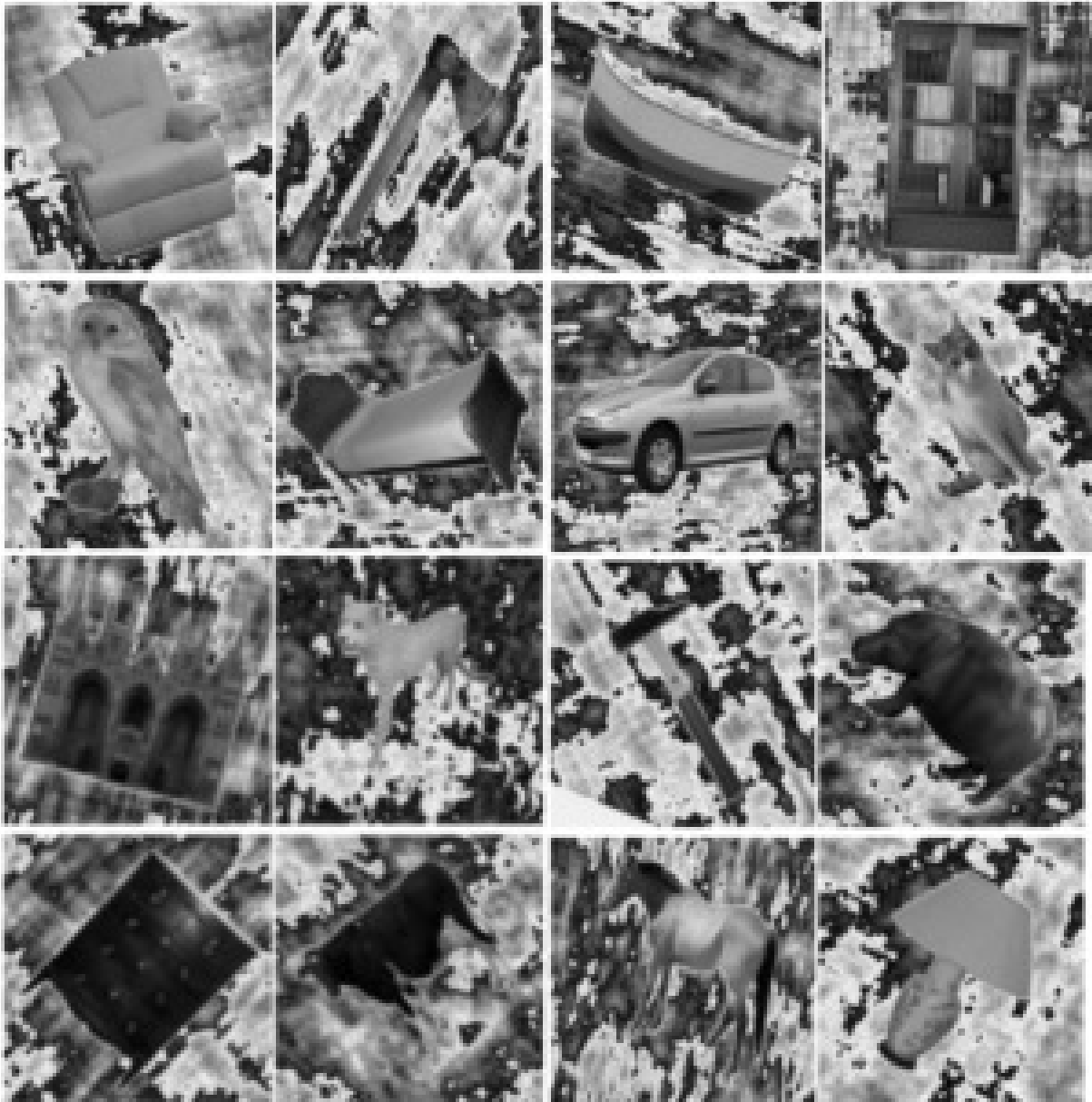


FIGURE 2.2: Example probe images used in the fMRI experiment. Subjects were asked to perform a perceptual categorization task and to rate their awareness (Moreno-Martínez & Montoro, 2012).



### 2.2.3 Analysis of behavioral performance

We assessed whether the level of discrimination accuracy of the image departed from chance in each of the awareness conditions. The metric to measure accuracy was the area under the receiver operating curve (ROC AUC). A response was defined as a "true positive" (TP) when "living" was both responded to and presented. A response was defined as a "false positive" (FP) when "living" was responded to while "nonliving" was presented. A response was defined as a "false negative" (FN) when "nonliving" was responded to while "living" was presented. A response was defined as a "true negative" (TN) when "nonliving" was both responded to and presented. Thus, a hit rate (H) was the ratio between TP and the sum of TP and FN, and a false alarm rate (F) was the ratio between FP and the sum of FP and TN. We first calculated ROC AUC associated with the individual behavioral performance within each of the different states of awareness (henceforth called the experimental ROC AUC). Then, we applied permutation tests to estimate the empirical chance level. We bootstrapped trials for a given awareness state with replacement (Horowitz, 2001); the order of the responses was shuffled while the order of the correct answers remained the same to estimate the empirical chance level. We calculated the ROC AUC based on the shuffled responses and correct answers to estimate the chance level of the behavioral performance, and we called this the chance level ROC AUC.

This procedure was repeated 10,000 times to estimate the distribution of the empirical chance level  $A'$  for each awareness state and each subject. The probability of empirical chance level ROC AUC being greater or equal to the experimental ROC AUC was the statistical significance level (one-tailed p-value, Ojala and Garriga, 2010). Hence, we determined whether the ROC AUC of each individual was above chance across the different awareness states (Bonferroni corrected for multiple tests).

## 2.2.4 fMRI acquisition and preprocessing

A 3-Tesla SIEMENS's Magnetom Prisma-fit scanner and a 64-channel head coil was used. In each fMRI session, a multiband gradient-echo echo-planar imaging sequence with an acceleration factor of 6, resolution of  $2.4 \times 2.4 \times 2.4 \text{ mm}^3$ , TR of 850 *ms*, TE of 35 *ms*, and bandwidth of 2582 Hz/Px was used to obtain 585 3D volumes of the whole brain (66 slices; FOV = 210 *mm*). For each subject, one high-resolution T1-weighted structural image was also collected.

The visual stimuli were projected on an MRI-compatible out-of-bore screen using a projector placed in the room adjacent to the MRI room. A small mirror, mounted on the head coil, reflected the screen for presentation to the subjects. The head coil was also equipped with a microphone that enabled the subjects to communicate with the experimenters in between the scanning blocks.

The first 10 volumes of each block were discarded to ensure steady-state magnetization; to remove non-brain tissue, a brain extraction tool (BET; Smith, 2002) was used; volume realignment was performed using MCFLIRT (Jenkinson et al., 2002); minimal spatial smoothing was performed using a Gaussian kernel with full width at half maximum (FWHM) of 3 *mm*. Next, Independent component analysis based on automatic removal of motion artifacts (ICA-AROMA) was used to remove motion-induced signal variations (Pruim et al., 2015) and this was followed by a high-pass filter with a cutoff of 60 seconds. The sessions are aligned to a reference volume of the first session<sup>1</sup>. All the processing of the fMRI scans were performed within the FSL (FMRIB Software Library, v6.0.0; Jenkinson et al., 2012) framework and were executed using the NiPype Python library (Gorgolewski et al., 2011).

For each subject, the relevant time points or scans of the preprocessed fMRI data of each run were labeled with attributes such as (i.e., cat, boat), category (i.e., animal, vehicle), and condition (i.e., living vs. nonliving) using the behavioral data files generated by Psychopy (v1.84, Peirce, 2007). Next, data from all sessions were stacked and each voxel's time series was block-wise z-scored (normalized) and linear detrended. Finally, to

---

<sup>1</sup>Detailed documentation can be addressed in <https://tinyurl.com/up2txma>

account for the hemodynamic lag, examples were created for each trial by averaging the 3 or 4 volumes between the interval of 4 and 7 seconds after image onset.

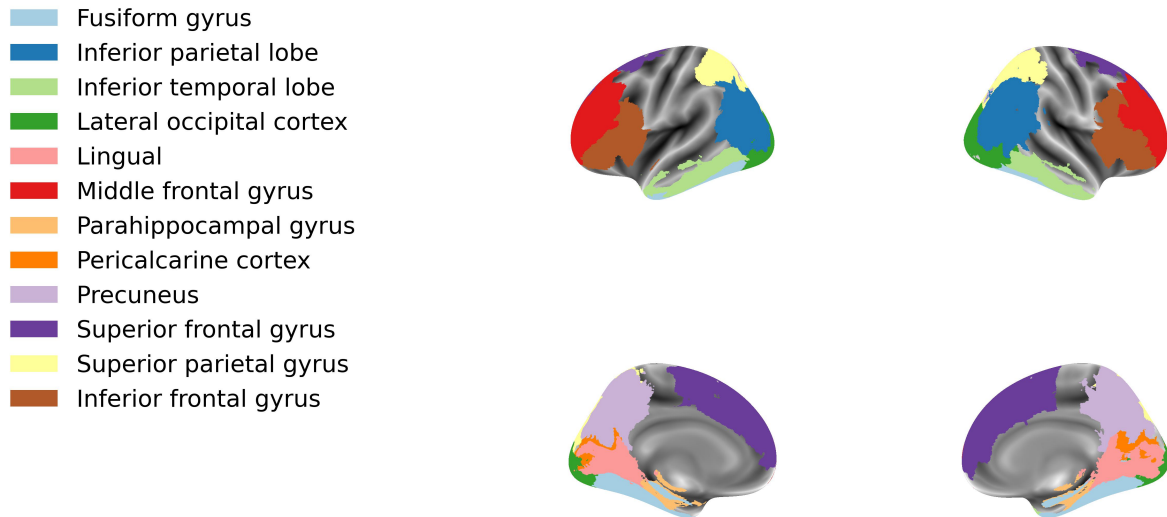


FIGURE 2.3: Selected regions of interest projected on an MNI standard template image. Initially, 12 ROIs were extracted for each hemisphere fusiform gyrus, inferior parietal lobe, inferior temporal lobe, lateral occipital cortex, lingual, middle frontal gyrus, parahippocampal gyrus, pericalcarine cortex, precuneus, superior frontal gyrus, superior parietal gyrus, pars opercularis gyrus, pars triangularis gyrus, and pars orbitalis gyrus. The last three ROIs were combined into a larger, bilateral ROI, namely the inferior frontal gyrus.

For a given awareness state, examples of BOLD activity patterns were collected for each of the 12 regions of interest (ROIs). There were 12 ROIs for each hemisphere (see Figure 2.3). The ROIs included the lingual gyrus, pericalcarine cortex, lateral occipital cortex, fusiform gyrus, parahippocampal gyrus, inferior temporal lobe, inferior parietal lobe, precuneus, superior parietal gyrus, superior frontal gyrus, middle frontal gyrus, and inferior frontal gyrus (comprising pars opercularis gyrus, pars triangularis gyrus, and pars orbitalis gyrus). Automatic segmentation of the high-resolution structural scan was done with FreeSurfer’s automated algorithm `recon-all` (v6.0.0). The resulting masks were transformed into functional space using 7 degrees of freedom linear registrations implemented in FSL FLIRT (Jenkinson et al., 2002) and binarized. All further analyses were performed in the native BOLD space within each subject.

### 2.2.5 fMRI decoding pipeline within awareness states

MVPA in the current chapter was conducted using scikit-learn (Pedregosa et al., 2011) and Nilearn (Abraham et al., 2014), using a linear support vector machine (SVM) classifier (Cortes & Vapnik, 1995). SVM has limited complexity, hence reducing the probability of over-fitting (model performs well in training data but has a poor performance in testing data) and it has been shown to perform well with fMRI data (Lewis-Peacock & Norman, 2014; Pereira & Botvinick, 2011). We used an SVM with L1 regularization, nested with invariant voxels removal<sup>2</sup> and feature scaling between 0 and 1<sup>3</sup> as preprocessing steps<sup>4</sup>. The key parameter of the invariant voxel removal was to determine which features had a variance of zero, and the parameter of the scaling is to learning the minimum and the maximum of the features. The key parameters of the invariant voxel removal and scaling were fitted in the training set and applied to the testing set. Note that these preprocessing steps are different from the detrending and z-scoring of the BOLD signals and represent conventional machine learning practices (Bruha, 2000; King et al., 2016).

During cross-validation, trials corresponding to one living (i.e., cat) and one non-living (i.e., boat) item for a given awareness state (i.e., unconscious) were left out as the test set and the rest was used to fit the machine learning pipeline. With 96 unique items, 2256 cross-validation folds could be performed in principle. However, because the awareness states were randomly sampled for each unique item (i.e., cat), the proportion of examples for training and testing were not equal among different folds. Some subjects had less than 96 unique items for one or more than one of the awareness states. Thus, less than 2256 folds of cross-validations were performed in these cases. The number of total available instances ranged from 700 to 800 and the number of test sets ranged from 2 (one trial for living and one trial for nonliving) to 20 for either unconscious or conscious states. The performance of the fitted pipeline was estimated by comparing the predicted labels and the true labels using ROC AUC for the test set.

---

<sup>2</sup>sklearn.feature\_selection.VarianceThreshold

<sup>3</sup>sklearn.preprocessing.MinMaxScaler

<sup>4</sup>sklearn.pipeline.make\_pipeline

To get an empirical chance level of the decoding, the same cross-validation procedures were repeated by replacing the linear SVM classifier with a "dummy classifier"<sup>5</sup>, which makes predictions based on the distribution of the classes of the training set randomly without learning the relevant multivariate patterns. The same preprocessing steps were kept in the pipeline. It is important to note that the optimization of L1 regularization was SVM specific, thus, it did not apply to the "dummy classifier".

The mean differences between the true decoding scores and the chance-level decoding scores were computed as the experimental score. To estimate the null distribution of the performance differences, we performed permutation tests. First, we concatenate the true decoding scores and the chance level decoding scores and then shuffle the concatenated vector. Second, we split the concatenated vector into a new "decoding scores" vector and a new "chance level decoding scores" vector. The mean differences between these two vectors were computed. This procedure was repeated 10,000 times to estimate the null distribution of the performance differences. The probability that the experimental score was greater or equal to the null distribution was the statistically significant level (one-tailed p-value). Finally, the comparisons across awareness states and the number of ROIs were corrected using Bonferroni correction (Figure 2.5 columns 1 and 2).

### 2.2.6 Generalization pipeline across awareness states

Here the classifier was trained from data in a particular awareness state (the "source"; e.g., on conscious trials) and then tested on a different awareness state (the "target"; e.g. on unconscious trials) on top of the cross-validation procedure described above. Similar to the decoding analysis within each awareness state, instances corresponding to one living and one non-living item in both "source" and "target" were left out, but only the left-out instances in the "target" were used as the test set. The rest of the instances in "source" were used as the training set to fit the machine learning pipeline (preprocessing + SVM) as described above. Because the awareness states were randomly sampled for each unique item (i.e., cat), the proportion of examples for training and testing were not equal among

---

<sup>5</sup>`sklearn.dummy.DummyClassifier`

different folds. The number of total available instances ranged from 700 to 800 and the number of test sets ranged from 2 (one trial for living and one trial for nonliving) to 20. The performance of the fitted pipeline was estimated by comparing the predicted labels and the true labels using ROC AUC for the test set.

To get an empirical chance level of the decoding, a similar procedure to that described above with a "dummy classifier" was employed here. Similar permutation test procedures were used to estimate the empirical null distribution of the difference between the experimental and chance level ROC AUC and the estimation was repeated 10,000 times. The probability that the experimental score was greater or equal to the null distribution was the statistically significant level (one-tailed p-value). Finally, the comparisons across awareness states for each subject were corrected using Bonferroni correction (Figure 2.5 column 3).

## 2.3 Results

### 2.3.1 Behavioral performance

We assessed whether subjects' performance at discriminating the image category from chance in each of the awareness conditions by using a signal detection theoretic measure to index perceptual accuracy, namely, the non-parametric  $A'$  (Zhang & Mueller, 2005). Permutation tests were performed to estimate the empirical chance level within each subject (see Methods). All subjects displayed above chance perceptual sensitivity in both glimpse and visible trials ( $p < 0.001$ , permuted p-values). Four of the seven subjects showed null perceptual sensitivity ( $0.16 < p < 0.64$ ) in those trials in which subjects reported a lack of awareness of the images. Discrimination performance in two additional subjects deviated from chance (p values  $< 0.02$ ,  $0.03$ , uncorrected for the number of tests across the different awareness states) but only one subject clearly showed above chance performance in the unaware trials ( $p = 0.00036$ ). Figure 2.4 illustrates the distribution of  $A'$  values alongside the chance distribution for each participant.

---

While individual biases in reporting awareness are difficult to control experimentally, several features of the data indicate that subjects were using the subjective ratings appropriately. First, the meta- $d'$  scores that measure how well one's awareness ratings align with discrimination accuracy were high. Second, as noted above, the average target's duration set by the adaptive staircase was the lowest/highest in the trials rated as unaware/aware. Finally, an inspection of the data shows that the proportion of trials rated as aware was similar among the subjects showing null sensitivity on the unconscious trials and those shown above chance performance (see table 2.1). Table 2.1 shows that six out of seven subjects showed null perceptual sensitivity in the behavioral test in those trials in which subjects reported a lack of visual awareness of the images (Figure 2.4). In trials with partial awareness ("glimpse") and full awareness, behavioral discrimination performance was well above chance.

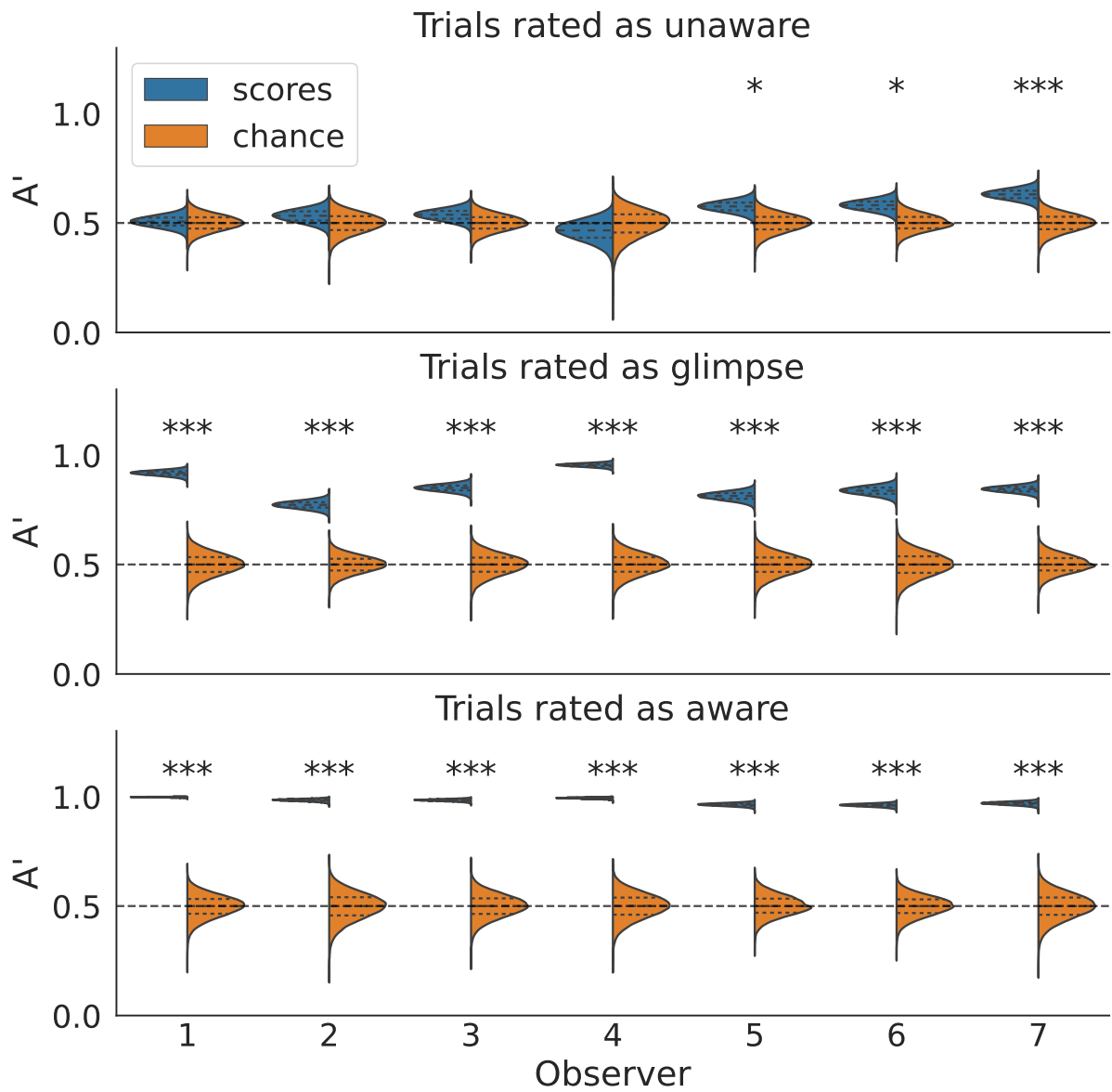


FIGURE 2.4: Behavioral performance. Distribution of within-subject  $A'$  scores with mean, first and third quartile, and the corresponding empirical chance distributions for each subject and awareness state. \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ .



Subject	Awareness	Proportion	Probe time (ms)	P(FA)	p(Hit)	d'	p	meta-d'
Sub-01	Unconscious	0.45	18.77 ± 10.56	0.51	0.51	0.02	0.4371	2.53
	Glimpse	0.26	29.96 ± 11.16	0.13	0.87	2.26	< 0.001***	
	Conscious	0.29	41.30 ± 13.54	0.01	0.99	9.30	< 0.001***	
Sub-02	Unconscious	0.35	16.48 ± 09.11	0.44	0.48	0.11	0.2341	1.64
	Glimpse	0.45	24.22 ± 13.42	0.29	0.71	1.12	< 0.001***	
	Conscious	0.20	36.74 ± 13.62	0.04	0.97	3.82	< 0.001***	
Sub-03	Unconscious	0.44	34.73 ± 18.91	0.49	0.54	0.12	0.1675	1.54
	Glimpse	0.31	50.79 ± 18.52	0.19	0.77	1.61	< 0.001***	
	Conscious	0.24	58.92 ± 20.79	0.04	0.98	3.81	< 0.001***	
Sub-04	Unconscious	0.44	34.92 ± 16.57	0.22	0.20	-0.06	0.6526	1.85
	Glimpse	0.32	51.55 ± 18.12	0.10	0.95	2.91	< 0.001***	
	Conscious	0.24	59.07 ± 18.27	0.01	0.99	6.37	< 0.001***	
Sub-05	Unconscious	0.42	19.80 ± 11.93	0.39	0.49	0.25	0.0276*	2.55
	Glimpse	0.28	31.43 ± 13.50	0.19	0.74	1.51	< 0.001***	
	Conscious	0.29	40.22 ± 17.03	0.09	0.96	3.14	< 0.001***	
Sub-06	Unconscious	0.44	22.70 ± 13.98	0.51	0.61	0.27	0.019*	1.72
	Glimpse	0.22	35.80 ± 15.42	0.21	0.77	1.53	< 0.001***	
	Conscious	0.34	43.18 ± 17.83	0.05	0.91	3.00	< 0.001***	
Sub-07	Unconscious	0.41	28.48 ± 16.39	0.39	0.57	0.47	< 0.001***	1.37
	Glimpse	0.39	41.19 ± 17.60	0.27	0.83	1.57	< 0.001***	
	Conscious	0.20	50.47 ± 20.39	0.06	0.95	3.21	< 0.001***	

TABLE 2.1: Individual proportion of awareness ratings, target duration, probability of hits and false alarms and corresponding d' scores, as well meta-d' scores for each subject. \*:p < 0.05, \*\*:p < 0.01, \*\*\*:p < 0.001; one-tailed p-values, Bonferroni corrected.

### 2.3.2 Decoding within each awareness state and generalization across awareness states

Out-of-sample generalization was applied to decode the categories of the items when they were consciously aware and when they were not. Figure 2.5 illustrates the decoding results in the unconscious and the conscious trials as well as the generalization performance from a decoder trained on the conscious trials and tested on the unconscious trials. Importantly, Figure 2.5 shows individual classification performance by using cross-validation and permutation tests run within each participant. ROC AUC was used as a measure of decoding accuracy (\*:  $p < 0.05$ , \*\*:  $p < 0.01$ , \*\*\*:  $p < 0.001$ , after multiple comparison corrections for the number of ROIs tested). In the unconscious trials, associated with null perceptual sensitivity in most subjects, significant decoding of the image class was achieved from multi-voxel activity patterns in ventral visual areas and even prefrontal regions. In the unconscious trials, multi-voxel patterns in the fusiform gyrus and middle frontal gyrus contained information that allowed for decoding of the item class in six of seven subjects. The unconscious item could also be decoded from activity in the inferior frontal gyrus in five subjects, and in four subjects from the inferior parietal lobe, inferior temporal lobe, lateral occipital cortex, parahippocampal gyrus, and superior parietal gyrus. Interestingly, the pattern of decoding results of the subjects that showed above chance perceptual sensitivity in those trials subjectively rated as unaware did not appear different in any brain area from the subjects in which sensitivity was at chance (Figure 2.6), although further research comparing a bigger number of subjects in the two cases is necessary to make further determinations.

We then investigated whether the multi-voxel patterns in the conscious trials were similar to the patterns in the unconscious trials. We applied cross-awareness-state generalization from the conscious to the unconscious condition. We found that the patterns in fusiform gyrus, lateral occipital cortex, and precuneus could be generalized from conscious to unconscious in all subjects. Additionally, the activity patterns in inferior parietal lobe, inferior temporal lobe, lingual, middle frontal gyrus, and superior parietal gyrus could

---

be generalized from conscious to unconscious in six out of seven subjects. The activity in inferior frontal lobe and pericalcarine cortex could be generalized from conscious to unconscious in five out of seven subjects. And lastly, the activity in parahippocampal gyrus and superior frontal gyrus could be generalized from conscious to unconscious in four out of seven subjects.

This pattern of results indicates the presence of invariant multivariate patterns in both the visual areas and the frontal regions for the same item categories in the conscious and unconscious conditions.

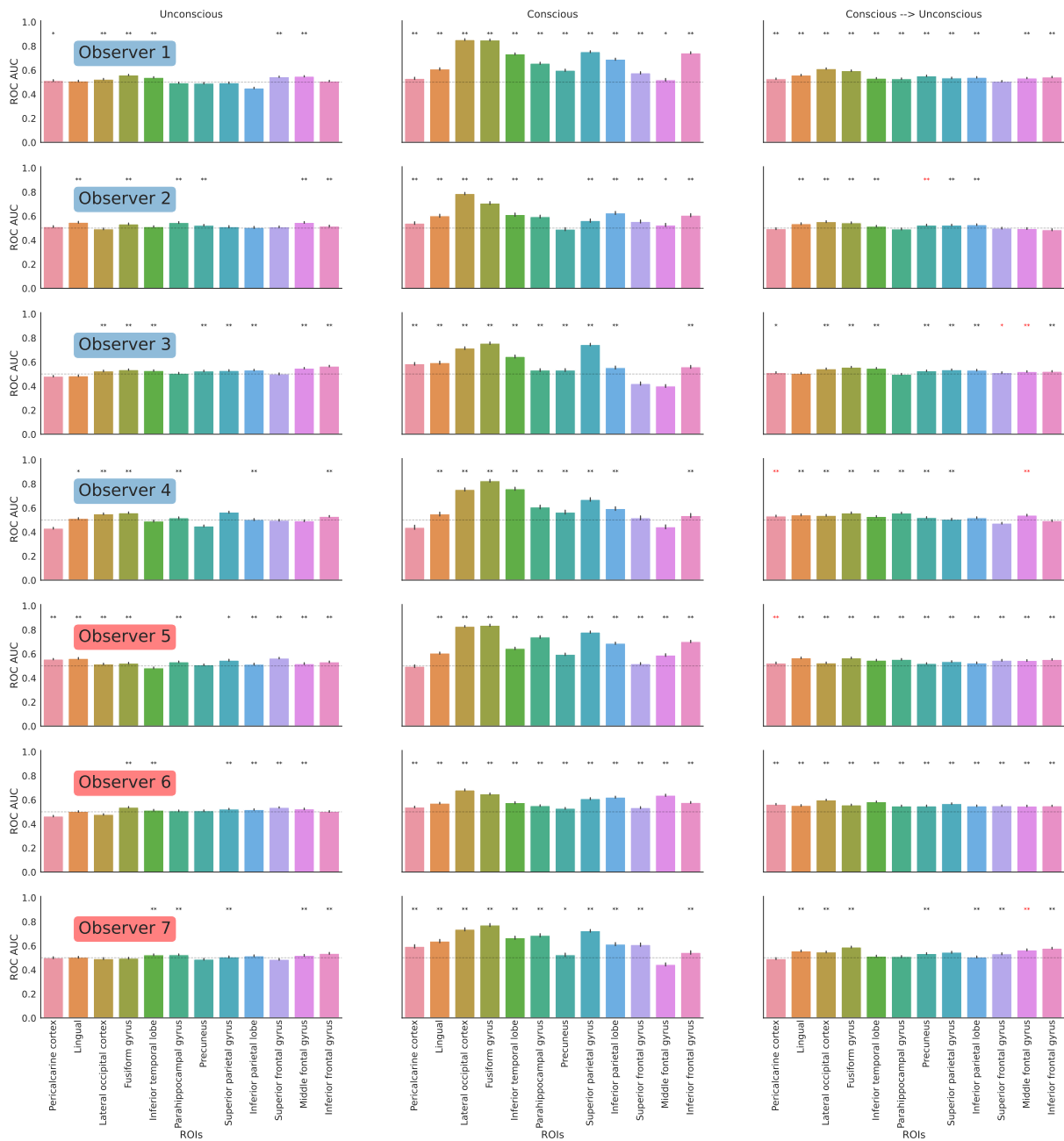


FIGURE 2.5: Decoding performance (ROC AUC) for out-of-sample images for each subject across the unconscious and conscious trials. \*:  $p < 0.05$ , \*\*:  $p < 0.01$ , \*\*\*:  $p < 0.001$ , one-tailed p-value, after multiple comparison correction for the number of ROIs tested for each subject. Error bars represent the standard error of the mean. Red asterisks indicate those ROIs in which the cross-awareness state generalization appeared above chance but the decoding in the conscious condition was at chance

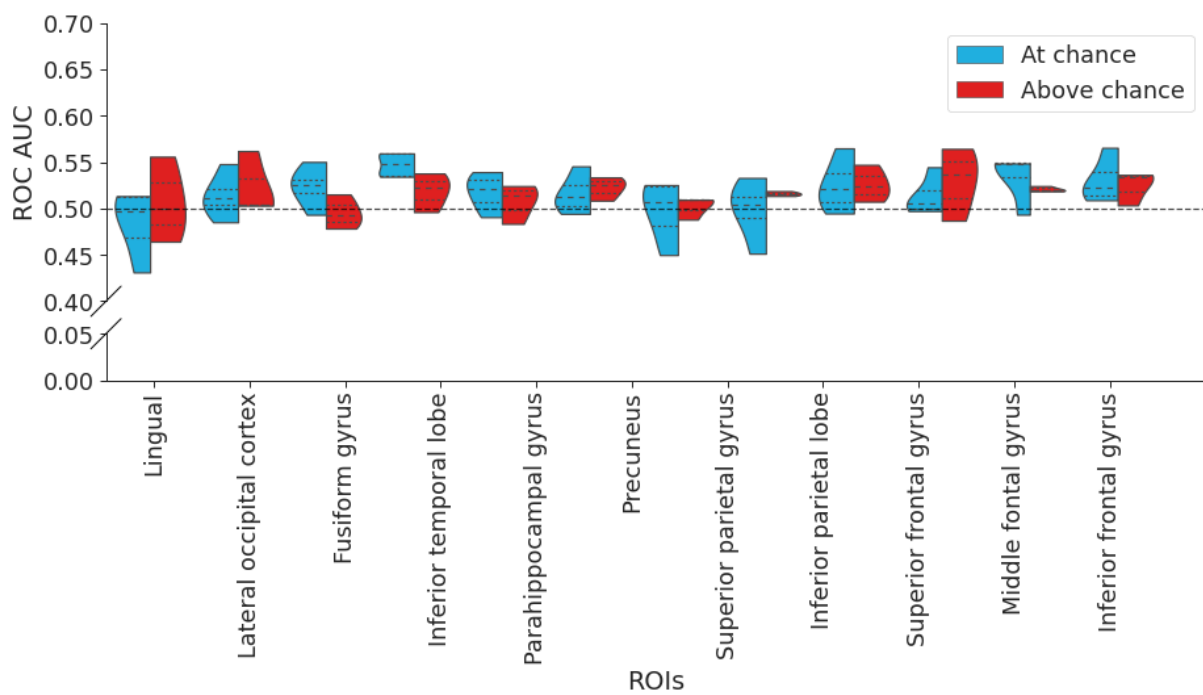


FIGURE 2.6: Additional decoding results. Distribution of decoding accuracy across the subjects whose perceptual sensitivity was at chance level and those who deviated from chance, including the mean, first, and third quartiles.

---

In addition to the standard MVPA we conducted on the fMRI data, we measure the robustness of the MVPA results by increasing the magnitude of the L1 regularization and changing the cross-validation paradigm (random stratified shuffling).

Shown as in Figure 2.7, when we increase the L1 regularization for the magnitude of 5, for the unconscious trials, we could decode the image categories from the lingual gyrus, fusiform gyrus, parahippocampal gyrus, middle frontal gyrus, and the inferior frontal gyrus in all subjects. Compared to the initial model, the pattern decoding was more robust using higher regularization. Additionally, when we change the cross-validation partitioning method (using random stratified shuffling, Figure 2.8), similar patterns were observed. Thus, we can confirm the robustness of the decoding pipeline used in the initial analysis.

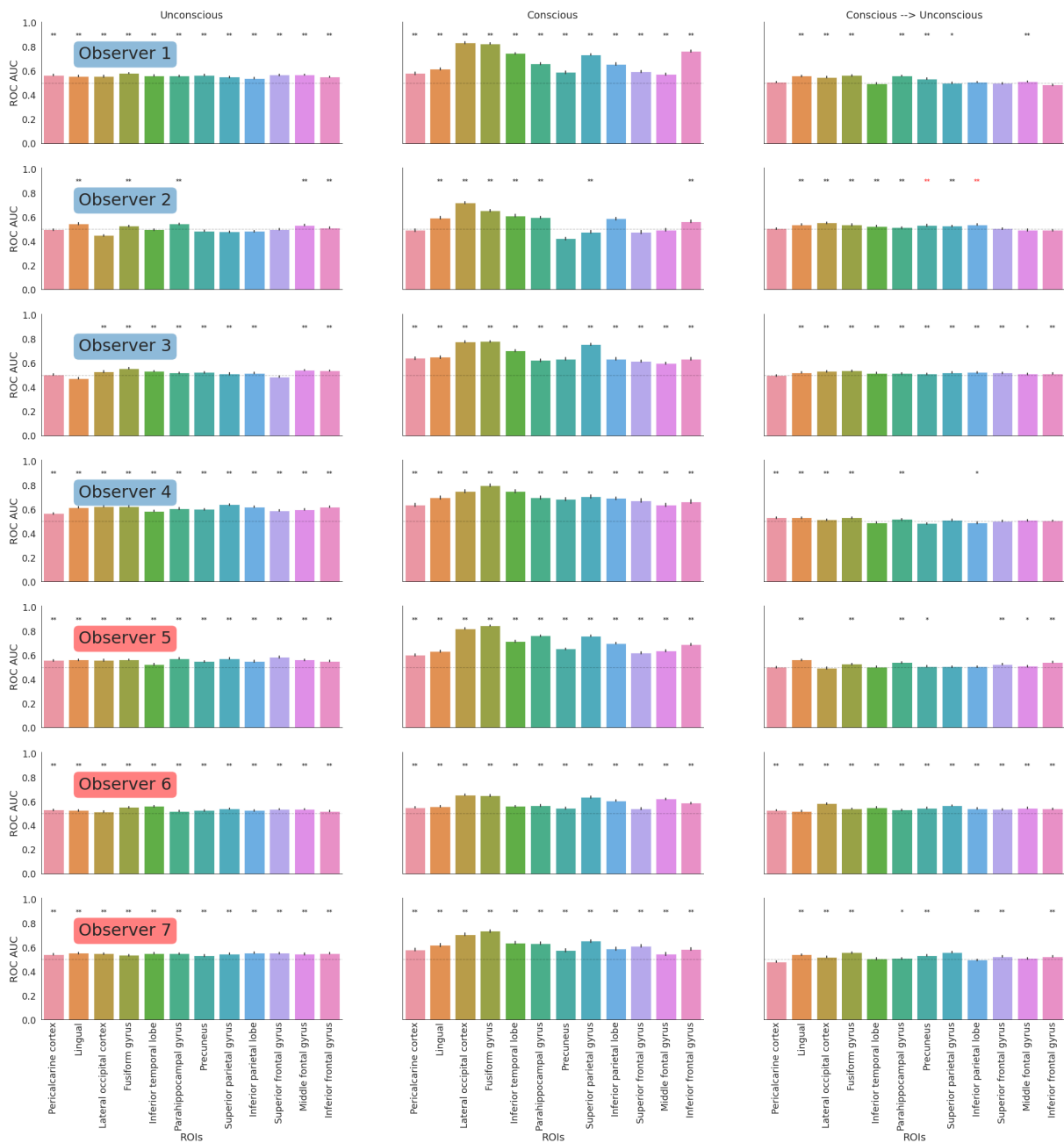


FIGURE 2.7: Decoding performance of the SVM with L1 regularization ( $C = 5$ ) using the previous out-of-sample generalization cross-validation partitioning scheme: \*:p < 0.05, \*\*:p < 0.01, \*\*\*:p < 0.001; one-tailed p-value, after multiple comparison correction for the number of ROIs tested for each subject. Error bars represent the standard error of the mean. Red asterisks indicate those ROIs in which the cross-awareness state generalization appeared above chance but the decoding in the conscious condition was at chance

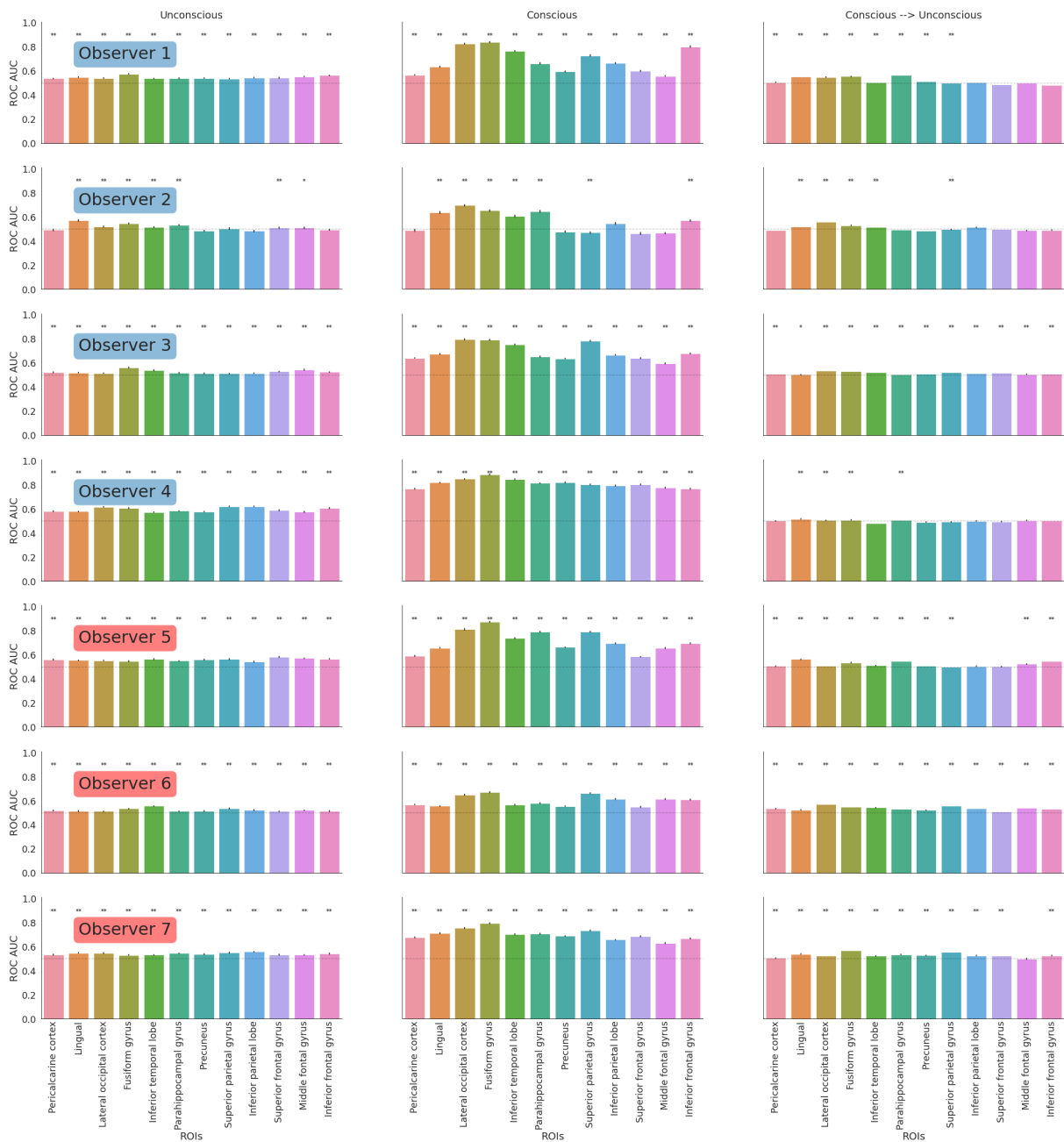


FIGURE 2.8: Decoding performance of the SVM with the default L1 regularization ( $C = 1$ ) using the stratified random shuffle cross-validation partitioning scheme. During each cross-validation within each state of awareness, 95% of the data was used for training, and 5% of the data was used for testing. This cross-validation scheme was repeated 1000 times. During the cross-awareness cross-validation, 95% of the trials in the conscious condition were randomly selected for training and 100% of the trials in the unconscious condition were used for testing: \*:  $p < 0.05$ , \*\*:  $p < 0.01$ , \*\*\*:  $p < 0.001$ ; one-tailed p-value, after multiple comparison correction for the number of ROIs tested for each subject. Error bars represent the standard error of the mean.



## 2.4 Discussion

In this chapter, we presented the results of a high-precision, within-subject fMRI-based decoding experiment to assess the representation of unconscious and conscious items using a combination of subjective and objective measures of awareness (i.e., null perceptual sensitivity on trials in which participants reported no awareness).

Isolating the brain representation of unconscious contents has been difficult to achieve systematically and reliably in previous work (Michel, 2019), with low numbers of trials and signal detection theoretic constraints (Macmillan, 1986) not allowing to decisively discard conscious perception (Fang & He, 2005; Gayet et al., 2020; Hesselmann et al., 2011; Ludwig et al., 2015; Stein et al., 2021). High-precision fMRI decoding paradigms can thus provide a richer information-based approach (Kriegeskorte et al., 2006) to reveal meaningful feature representations of unconscious content, and that otherwise would be missed. And critically, when unconscious content could be decoded, this was restricted to the visual cortex (also see Ludwig & Hesselmann, 2015; Stein et al., 2021).

Here we showed that even considering those subjects that had absolutely no awareness of the target item (null perceptual sensitivity), a linear SVM could decode the image categories from the fusiform gyrus and the middle frontal gyrus robustly through different modifications of decoding pipelines. High-precision fMRI decoding paradigms can thus provide a richer information-based approach (Kriegeskorte et al., 2006) to reveal meaningful feature representations of unconscious content, and that otherwise would be missed.

Remarkably, the fMRI decoding results indicate that the neural representation of conscious and unconscious contents overlap in the ventral visual pathway, also including parietal and even to some extent in prefrontal areas in a substantial part of the observers. Previous studies using lowly sampled fMRI designs could not reveal evidence consistent with this view (Schurger et al., 2010; Sterzer et al., 2008). Visual consciousness may be associated with neural representations that are more stable across different presentations

---

of the events (Schurger et al., 2010), but our data indicate that the underlying representational patterns in terms of perceptual content are to some degree generalizable across awareness states, despite the non-linear dynamic changes in the intensity of the neural response that occur in the frontoparietal cortex during conscious processing, (Beck et al., 2001; Dehaene & Changeux, 2004; Dehaene et al., 2001; Haynes et al., 2005; Kranczioch et al., 2005; Pessoa & Ungerleider, 2004), and even despite the visible items were here presented for a longer duration. This observation points to revisions to the neuronal global workspace model (Dehaene, 2014).

# 3 Assessing the representation of unseen contents using deep neural networks

## 3.1 Introduction: Artificial deep neural network as a model of the human visual system at the represen- tational level

The first neural network model, the perceptron (Rosenblatt, 1960), was first proposed as an abstract mathematical model of a biological neuron. The extension of multilayer perceptron and the different activation functions (Leshno et al., 1993) allowed perceptron models to map complex input and output patterns (Schäfer & Zimmermann, 2006; Sonoda & Murata, 2017). One of the classes of neural network models is called the feedforward convolutional neural network (FCNN) (Fukushima, 1980; Hinton et al., 2015; LeCun & Bengio, 1995; McFee et al., 2018). FCNNs are reasonably good model candidates for studying the human vision system (Lindsay, 2021; Richards et al., 2019). They have shown excellent performance in image recognition as well as predicting primate brain activities (Hinton et al., 2015; Kietzmann et al., 2019a; Kriegeskorte & Douglas, 2018; Schrimpf et al., 2020). The convolutional layers are thought to ensemble local features (i.e., edges and line segments) to become more robust features (i.e., eyes and nose on a face) for the higher-order layers (Qin et al., 2018).

A typical FCNN contains multiple convolutional layers and pooling layers, followed by

a few fully-connected layers before the last output prediction layer (Leshno et al., 1993). For instance, an image, with a height and width of 256 and 3 color channels, is processed by a moving two-dimensional (2D) kernel (e.g.  $3 \times 3$ ) separately for each channel. The 2D kernel moves sequentially from the left to right and from top to bottom within each channel to extract "features" for the next layer. The extracted feature at each iteration by the 2D kernel could be maximized or averaged to represent the features at a higher level (i.e., pooling, Scherer et al., 2010). This processing is repeated for several convolutional layers that contain these 2D kernels. After multiple convolutions and pooling processing, the feature representations of the original image become more and more abstract and the "field of view of network becomes wider" (Fukushima, 1980). Following the convolutional layer, the abstract "feature representations" were pooled or resized and then connected to fully-connected layers. These fully-connected layers are often called "hidden layers" in cognitive computational neuroscience literature (Kriegeskorte, 2015). In the end, an output prediction layer with a softmax (Equation 3.2) produces the probability of the input belonging to one of the categories. In order to train a FCNN to classify image categories, the images are passed forward through the model, and then the predicted categories and the true categories are compared by a loss function, for instance, cross-entropy. The difference between the predicted and true categories, namely the loss, is backpropagated from the last layer to the previous layers. The backpropagated error signals are the gradients. An optimizer, the stochastic gradient descent, for instance, uses the gradients to modify the weights in the layers. Weights in the convolutional layers are the weights in the kernels and the weights in the fully-connected layers are the connections between the artificial neurons between two layers.

## 3.2 Methods

### 3.2.1 A feedforward convolutional neural network simulation of unconscious processing at the representational level

In this chapter, we had deep artificial neural networks from computer vision to perform a similar visual discrimination task to that given to our human observers. We sought to model our observation of hidden informative neural representations of items associated with null perceptual sensitivity. Note that the computer simulation approached the question at the representational level and not at the implementation level (Marr and Vision, 1982, i.e., the network simulation did not aim to provide a biological model of the system). Our modeling goal, therefore, was to demonstrate that the hidden layer representation of the neural network contains information that allows for decoding of a noisy visual stimulus even when artificial neural networks performed image classification at chance level.

The FCNN was pretrained with the ImageNet dataset (Deng et al., 2009) and then adapted to our experiment via transfer learning procedures (Yosinski et al., 2014). Due to a large number of model testing and limited computational power, this chapter focused on decoding the last hidden layer of the FCNN. The rationale was based on previous studies that also modeled visual recognition using FCNN (Güçlü & van Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014). These studies demonstrated that the last hidden layer of the FCNN contains a condensed summarization of the relevant object properties by assembling feature information from the previous convolutional layers. These hidden layers have representational spaces that are similar to the inferior temporal cortex (IT) (Güçlü & van Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014). IT is considered one of the regions that ensembles low-level visual inputs and produces high-level features for recognizing objects (Kriegeskorte et al., 2008b). Therefore, we focused our representational modeling approach on the very last hidden layer of the FCNN in the current chapter. We hypothesized that similar to the human observers, when the FCNN displays null sensitivity in image classification, a linear classifier could nevertheless read out information from

the hidden layer of the FCNN that can predict the content of the stimulus presented. However, we also predicted that when the signal-to-noise ratio of the image was too high, neither the FCNN nor the linear classifier could classify the object categories.

We used the convolutional layers from different pretrained FCNNs, such as the AlexNet (Krizhevsky et al., 2012), the VGGNet (Simonyan & Zisserman, 2015), the ResNet50 (He et al., 2016), the MobileNet (Howard et al., 2017), and the DenseNet169 (Huang et al., 2017), implemented in PyTorch V1.8.0 and torchvision V0.9.0 (Paszke et al., 2017; Paszke et al., 2019) to perform the simulations. These models were downloaded from PyTorch in May 2022. The FCNNs first learned to perform the same visual discrimination task as the human observers with clear images of animate and inanimate items with no background. “Soft labels” were used to make the network less sensitive to the noise added during testing (Nguyen et al., 2014). One example composed of random noise was added to each batch of animate and inanimate images (batch size = 8) during the training phase. The noise was sampled by a normal distribution with the mean and standard deviation of the images of the same batch. The probability associated with this noise instances was "0.5". These random noise training examples are important to improve the robustness of the FCNNs when dealing with animate and inanimate images embedded in noise, and they are important to reduce the bias of the FCNNs in predicting the image categories when the embedded noise was very high. The FCNNs only predict one category, either animate or inanimate, when the images are embedded in high levels of noise if they were not trained with the random noise examples. This method is inspired and simplified from Jin et al. (2015) research results.

The FCNNs were then tested under different levels of noise in the image. The goal here was to emulate the pattern observed in the fMRI study (i.e., decoding of the noisy image in the absence of perceptual sensitivity). To control for the initialization state of the FCNNs, we fine-tuned some of the popular FCNN pretrained models, AlexNet (Krizhevsky et al., 2012), VGG19 (Simonyan & Zisserman, 2015), ResNet50 (He et al., 2016), MobileNetV2 (Howard et al., 2017), and DenseNet169 (Huang et al., 2017), which were pretrained using the ImageNet dataset (Deng et al., 2009) and then were adapted to

our experiment using a transfer learning (fine-tuning) procedure (Yosinski et al., 2014), illustrated by Figure 3.1.

The AlexNet (Krizhevsky et al., 2012) contained six convolutional layers; the VGG19 (Simonyan & Zisserman, 2015) contained five convolutional blocks with an increased size of convolutional processing, resulting 16 convolutional layers; the Resnet50 (He et al., 2016) contained an initial convolutional layer, followed by four convolutional blocks, resulting 50 convolutional layers; the MobileNetV2 (Howard et al., 2017) contained an initial convolutional layer and seven bottleneck blocks, followed by a convolutional layer, a pooling layer, and another convolutional layer, resulting 25 convolutional layers; the DenseNet169 (Huang et al., 2017) contained an initial convolutional layer and four consecutive dense blocks and each was followed by a transition convolutional layer, resulting in 168 convolutional layers.

Pretrained FCNNs were stripped of the original fully-connected layer while weights and biases of the convolutional layers were frozen and not updated further (Yosinski et al., 2014). An adaptive pooling (McFee et al., 2018) operation was applied to the last convolutional layer so that the output of this layer became a one-dimensional vector, and a new fully-connected layer took the weighted sum of these outputs (namely the "hidden layer"). The number of artificial units used in the hidden layer could be any positive integer, but for simplicity, we took 300 as an example and we explored how the number of units (i.e., 2, 5, 10, 20, 50, 100, and 300) influenced the pattern of results. The number of hidden layer units determined the number of new weights,  $w_i$ , for training. The outputs of the hidden layer were passed to an activation function (Specht, 1990), which could be linear (i.e., identical function) or nonlinear (i.e., rectified function). A dropout was applied to the hidden layer during training but not during testing. The dropout function randomly zeros a proportion of outputs of the applied layer, and this procedure forces the layer to learn redundant features to improve robustness. Different dropout rates were explored (i.e., 0, 0.25, 0.5, and 0.75), where a zero dropout rate meant no dropout was applied. The dropout operation was varied to investigate how feature representations could be affected by a simple regularization.

A new fully-connected layer (namely the "classification layer") took the outputs processed by the activation function of the hidden layer to compose the classification layer. The number of artificial units used in the classification layer depended on the activation function applied to the outputs of the layer. If the activation function was sigmoid (Equation 3.1), one unit was used, while if the activation was a softmax function (Equation 3.2), two units were used.

$$\psi(x_i) = \frac{1}{1 + e^{-x_i}} \quad (3.1)$$

EQUATION 3.1: Equation of a sigmoid function.  $x_i$  is the prediction of the  $i^{th}$  categories after the initial pass of the output layer. Under subscripts " $i$ " denotes the  $i^{th}$  output of a given artificial unit.

$$\psi(x_i) = \frac{e^{x_i}}{\sum e^{x_i}} \quad (3.2)$$

EQUATION 3.2: Equation of a softmax function.  $x_i$  is the prediction of the  $i^{th}$  categories after the initial pass of the output layer. Under subscripts " $i$ " denotes the  $i^{th}$  output of a given artificial unit.

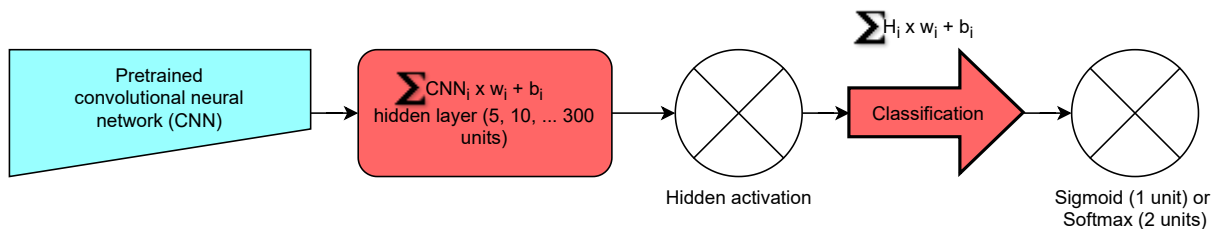


FIGURE 3.1: A simplified scheme of fine-tuning a pre-trained feedforward convolutional neural network model. The task is to classify the living vs. non-living category of the images used in the experiment without noise. The blue architecture was frozen and the weights were not updated, while the weights of the red architectures were updated during training.

The reorganized FCNN was trained on the gray-scaled, augmented (flipped or rotated), and normalized experimental images. The FCNN was then validated on images that were also gray-scaled but with different degrees of augmentation. The loss function was the binary cross entropy (Equation 3.3), a special case of the cross entropy family. The optimizer was Adam (Kingma & Ba, 2014) with a learning rate of  $1e^{-4}$  without decay. The PyTorch implemented binary cross entropy “clamps its log function outputs to be



greater than or equal to -100. This way, we can always have a finite loss value and a linear backward method” (Paszke et al., 2017; Paszke et al., 2019). The validation performance was used to determine when to stop training, and it was estimated every 10 training epochs. The FCNN model was trained until the loss did not decrease for five validation estimations.

$$H_p(q) = -\frac{1}{N} \sum_{i=1}^N y_i \cdot \log(p(y_i)) + (1 - y_i) \cdot \log(1 - p(y_i)) \quad (3.3)$$

EQUATION 3.3: Binary cross entropy loss function used for computing the loss during training and validation of the FCNNs.  $y_i$  is one of the two categories, and  $p(y_i)$  is the probabilistic prediction of the corresponding category.

As noted, after training, the weights and biases of the FCNNs were frozen to prevent the weights from changing during the test phase. During the test phase, Gaussian noise was added to the images to reduce the FCNNs classification performance. Similar augmentations as in the validation set were applied to the testing image sets. The noise added to the images was sampled from a Gaussian distribution centered at zero and different variance ( $\sigma$ ). The level of noise was defined by setting up the variance at the beginning of each test phase.

The noise levels ranged from 0 to 1000 in steps of 51 following a logarithmic trend. For a given noise level, 20 sessions of 96 images with a batch size of 8 were fed to the FCNN model, and both the outputs of the hidden layer and the classification layer were recorded. The outputs of the classification layer were used to determine the "perceptual sensitivity" of the FCNN model, and the outputs of the hidden layer were used to perform subsequent decoding analyses with a linear SVM classifier, in keeping with the fMRI analysis.

To determine the significance level of the FCNNs performance, the order of the true labels in each session was shuffled while the order of the predicted labels remained the same. The permuted performance was calculated for the 20 sessions. This procedure was repeated 10,000 times to estimate the empirical chance level of the FCNNs. The significance level was the probability that the performance of the FCNNs was greater or equal to the chance level performances (one-tailed test against 0.05). If the p-value is

greater or equal to 0.05, we considered that FCNNs' performance was not different from the empirical chance level.

We then assessed, for a given noisy image, whether the hidden layer of the FCNN (i.e., following the last convolutional layers), contained information that allowed decoding of the category of the image (animate vs inanimate). A linear SVM used the information contained in the FCNN hidden layer to decode the image class across different levels of noise, even when the FCNNs classification performance was at chance level. The outputs and the labels of the hidden layer from the 20 sessions were concatenated. A random shuffle stratified cross-validation procedure was used in the decoding experiments with 50 folds to estimate the decoding performance of the SVM. The statistical significance of the decoding performance was estimated by a different permutation procedure to the FCNNs, which here involved fitting the SVM model and testing the fitted SVM with 50-fold cross-validation in each iteration of permutation, and it was computationally costly <sup>1</sup>. On each permutation iteration, the order of the labels was shuffled while the order of the outputs of the hidden layer remained unchanged before fitting the linear SVM model (Ojala & Garriga, 2010). The permutation iteration was repeated 10,000 times to estimate the empirical chance level. The significance level was the probability of the true decoding score greater or equal to the chance level.

Thus, with five FCNNs (AlexNet Krizhevsky et al., 2012, VGGNet Simonyan and Zisserman, 2015, ResNet50 He et al., 2016, MobileNet Howard et al., 2017, and DenseNet169 Huang et al., 2017) as CNN backbones, with seven hidden layer sizes (2, 5, 10, 20, 50, 100, 300), with six hidden layer activation functions (ReLU Nair and Hinton, 2010, ELU Clevert et al., 2015, SELU Klambauer et al., 2017, Sigmoid, Tangent, Identical), with four dropout rates (0, 0.25, 0.5, 0.75), with two output layer activation functions (Sigmoid and Sotfmax), and with fifty-one noise levels, we had tested 85,388 pairs of FCNN and SVM models in classifying or decoding the image categories.

---

<sup>1</sup>sklearn.model\_selection.permutation\_test\_score

## 3.3 Results

### 3.3.1 General results of the simulation models

Despite the FCNNs failing to classify the image, the animate vs inanimate categories of the stimulus could still be decoded by analyzing the activity patterns of the hidden layer of the network. To test this, a linear SVM was applied to the hidden layer representation for decoding the image class across different levels of noise, even when the FCNN model classification performance was at chance level. Previous studies modeled visual recognition using FCNNs (Güçlü & van Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014), demonstrating that the last hidden layer of FCNNs has representational spaces that are similar to those in high-level regions in ventral visual cortex (Güçlü & van Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014; Kriegeskorte et al., 2008b). Therefore, we focused our analyses on the very last hidden layer of the FCNN in the current chapter, also considering limitations in computational resources due to a large number of simulations<sup>2</sup>.

Figure 3.2<sup>3</sup> shows the classification performance of the FCNNs (black) and also the decoding performance of the SVM decoding the image categories from the hidden layer representation of the FCNNs (blue) as a function of the level of noise and the different factors. When the level of noise was low, FCNNs could classify the category of the images very well, reaching ROC AUC scores higher than 0.9 but performance dropped with the level of Gaussian noise. The observed logarithmic downward trend could be due to the sampling of noise levels in the logarithmic space.

Remarkably, when the FCNNs failed to classify the noisy images ( $p > 0.05$ ;  $N = 56,258$ ), we observed that the hidden layer representation of these FCNNs contained information that allowed a linear SVM classifier to decode the image category above chance levels reliably in 21,188 of the simulations ( $p < 0.05$ , one sample permutation

---

<sup>2</sup>The results were initially reported in Mei et al. (2022), but the models were re-run for checking of robustness and reproducibility on 01/05/2022. The patterns observed are similar between the two runs, but the exact values are different.

<sup>3</sup>High definition figure: <https://tinyurl.com/y6dhls7c>

test). Figure 3.3 illustrates the decoding results based on the hidden layer representation when the FCNN was at chance level.

In order to better interpret the results, we divided the results based on the noise levels. We defined low and high noise by the median of the noise level between 0.01 and 1000 with 50 steps on a logarithm scale. When the noise level was relatively low ( $< 2.8$ ) and the FCNNs failed to discriminate the noisy images ( $N = 16,338$ ), 73.01% of the linear SVMs could decode the FCNN hidden layers, and the difference in decoding performance between SVM and FCNN was significantly greater than zero ( $p < 0.001$ , one sample permutation test, no correction for multiple comparisons).

Remarkably, even when the noise level was higher ( $> 2.8$ ) and the FCNNs classified the images at chance level ( $N = 39,920$ ), 23.19% of the linear SVMs could decode the image category from the FCNN hidden layers. Crucially, the comparison of SVM decoding from the hidden layer and the FCNNs classification performance including those 56,258 cases in which the FCNNs classified noisy images at chance level again showed a significant difference (permutation  $p < 0.05$ ), demonstrating that the hidden layer of the FCNNs contained informative representations despite the FCNN classification performance was at chance level.

MobileNetV2 produced more informative hidden representations that could be decoded by the linear SVMs compared to other candidate model configurations. We also observed that the classification performance of ResNet50 models trained with different configurations (e.g., varying number of hidden units, dropout rates) did not fall to chance level until the noise level was relatively high (closer to the dashed line), and the proportion of SVMs being able to decode FCNN hidden layers was higher compared to other model configurations (39.36% v.s. 32.93% for MobileNetV2, 30.14% for AlexNet, 24.30% for DenseNet, and 24.37% for VGGNet). Additionally, we observed that even when the noise level was high, the ResNet50 models provided a higher proportion of hidden representations that were decodable by the linear SVMs (34.07% v.s. 23.86% for MobileNetV2, 22.72% for DenseNet169, 22.69% for AlexNet, and 12.15% for VGGNet, see Figure 3.3).

In summary, MobileNetV2 and ResNet50 generated the most informative hidden representations. These networks have an intermediate level of depth. By comparison, the deepest network, DenseNet, did not produce better hidden representations. Hence the depth of the network per se does not appear to determine the quality or informativeness of the hidden representations significantly.

Then, we sought to further understand the influence of the components of the FCNN architecture (i.e., convolutional layers, dropout rate, number of hidden units) on decoding performance. We used a random forest classifier to compute the feature importance of the different FCNN components for predicting whether or not the SVM decoded the image class based on the hidden layer representation. The classification performance was estimated by random shuffle stratified cross-validation with 100 folds (80/20 splitting). In each fold, a random forest classifier was fit to predict whether or not the hidden representation was decodable on the training set, and then the feature importance was estimated by a permutation procedure on the test set (Altmann et al., 2010; Fisher et al., 2018). Briefly, for a given component (i.e., hidden layer activation function), the order of instances was shuffled while the order of the instances of other components was not changed, in order to create a corrupted version of the data. The dropped classification performance indicated how important a particular feature was. Figure 3.4 shows that the noise level in the image was the best indicator of whether a hidden representation was decodable, followed by model architecture, followed by the number of hidden units, and then by the type of hidden activation and output activation functions. The least important feature was the dropout rate. We conducted a one-way ANOVA on the feature importances obtained in each cross-validation fold using the attributes in Figure 3.4 as a factor. This allowed us to quantify the effect of the noise level, model architecture, number of hidden units, type of hidden activation function, type of output activation function, and dropout rate, on the feature importance. There were significant differences between the components of the network models tested ( $F(5, 594) = 2215.57$ ,  $p < 0.001$ ,  $\eta^2 = 0.95$ ). Tukey HSD post-hoc tests showed that all the pairwise comparisons were reliable (lowest  $p < 0.015$ , Bonferroni corrected for multiple comparisons, see Table 3.1).

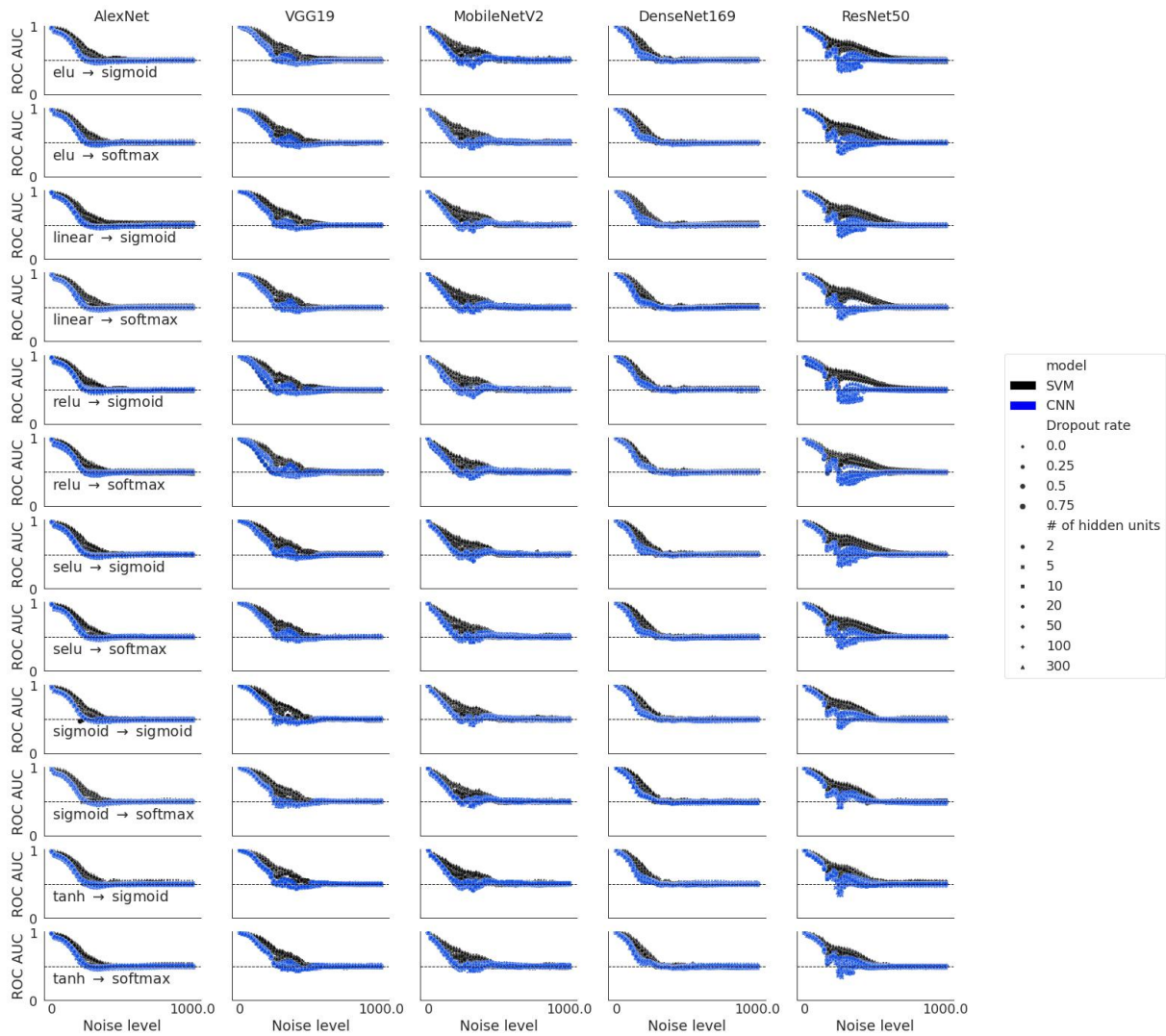


FIGURE 3.2: The black dots illustrate the classification performance of the FCNNs, as a function of noise level, type of pre-trained model configuration (column) and activation functions (row). The blue dots illustrate the classification performance of the linear SVMs applied to the hidden layer of the FCNN model.

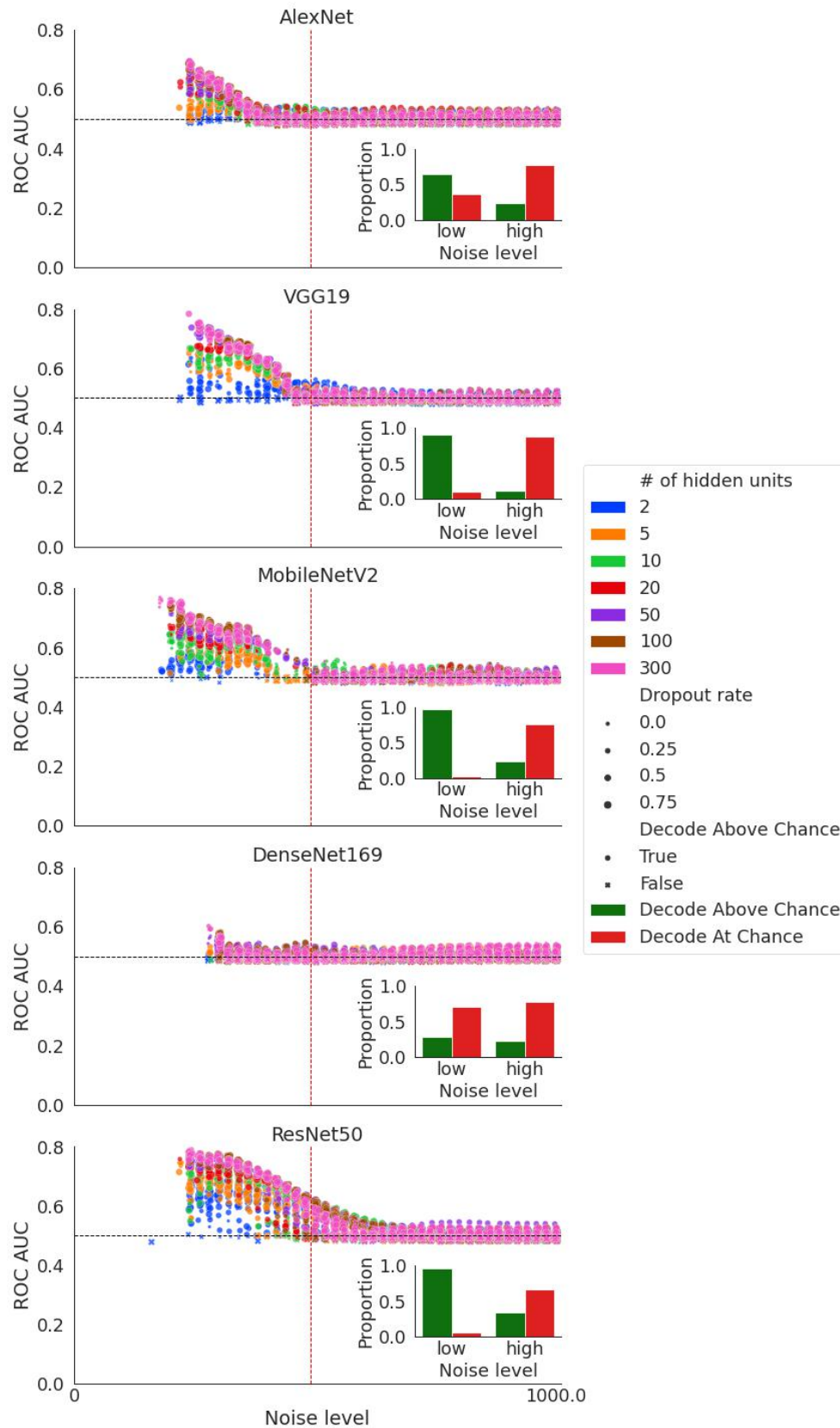


FIGURE 3.3: Image classification performance of the linear SVMs applied to the FCNN hidden layers when a given FCNN failed to discriminate the living v.s. nonliving categories, as a function of the noise level. The superimposed subplot depicts the proportion of times in which the linear SVM was able to decode the FCNN hidden layers as a function of low and high noise levels. The blue bar represents the proportion of linear SVMs being able to decode the FCNN hidden layers, while the orange bar represents the proportion of linear SVMs decoding the FCNN hidden layers at chance level.

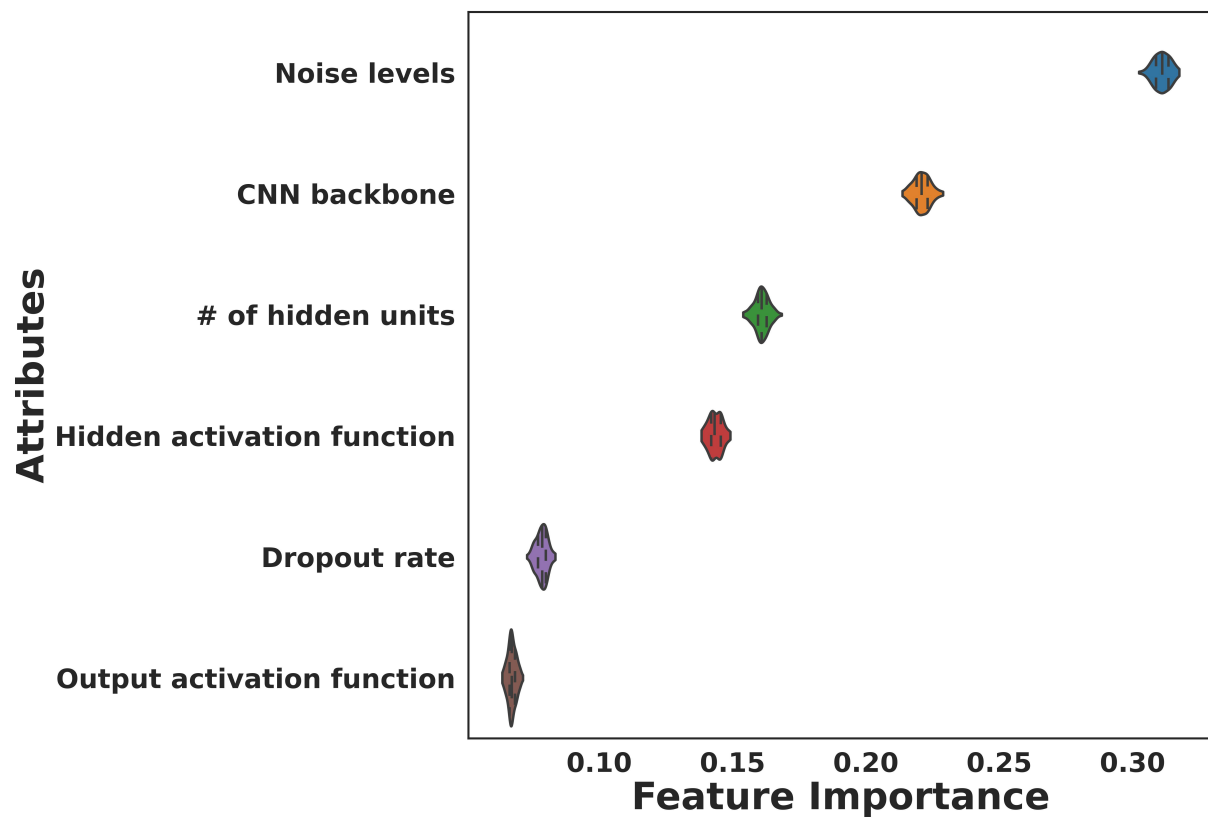


FIGURE 3.4: Feature importance of FCNN components that were manipulated in computational modeling. Feature importance was measured in arbitrary units, including the mean, first and third quartiles. The number of hidden layer units, noise levels, and pretrained configurations influenced the decoding performance of the image class based on the hidden layer of the FCNNs when its classification performance was at chance level.



A	B	$\bar{A}$	$\bar{B}$	$\Delta(A, B)$	SE	T	p-tukey	$\eta^2$
Dropout	Hidden func	0.00767	0.01741	-0.00972	0.00022	-43.49	< 0.001	0.90
Dropout	Hidden units	0.00768	0.01993	-0.01225	0.00022	-54.77	< 0.001	0.94
Dropout	Model architecture	0.00768	0.02444	-0.01676	0.00022	-74.95	< 0.001	0.97
Dropout	Noise level	0.00768	0.02687	-0.01919	0.00022	-85.82	< 0.001	0.97
Dropout	Output func	0.00768	0.01120	-0.00352	0.00022	-15.74	< 0.001	0.55
Hidden func	Hidden units	0.01741	0.01993	-0.00252	0.00022	-11.28	< 0.001	0.39
Hidden func	Model architecture	0.01741	0.02444	-0.00703	0.00022	-31.46	< 0.001	0.83
Hidden func	Noise level	0.01741	0.02687	-0.00946	0.00022	-42.33	< 0.001	0.90
Hidden func	Output func	0.01741	0.01120	0.00621	0.00022	27.75	< 0.001	0.79
Hidden units	Model architecture	0.01993	0.02444	-0.00451	0.00022	-20.18	< 0.001	0.67
Hidden units	Noise level	0.01993	0.02687	-0.00694	0.00022	-31.04	< 0.001	0.83
Hidden units	Output func	0.01993	0.01120	0.00872	0.00022	39.04	< 0.001	0.88
Model architecture	Noise level	0.02444	0.02687	-0.00243	0.00022	-10.87	< 0.001	0.37
Model architecture	Output func	0.02444	0.01120	0.01324	0.00022	59.22	< 0.001	0.95
Noise Level	Output func	0.02687	0.01120	0.01567	0.00022	70.08	< 0.001	0.96

TABLE 3.1: Post-hoc test results regarding the feature importances.  $\bar{A}$ : mean of A;  $\bar{B}$ : mean of B;  $\Delta A, B$ : difference between A and B Dropout: Dropout rate; SE: standard error of the difference between A and B; Hidden func: Hidden activation function; Output func: Output activation function.

### 3.3.2 Robustness of the simulation models under various conditions

Additional simulations were run to test whether similar results are obtained when the FCNNs are trained with images embedded in Gaussian noise. Here we used a similar pipeline to that used with clear images, except that here the images during training were embedded in Gaussian noise sampled from a standard normal distribution (mean centered at zero with unit variance, one). The results are shown in Figures 3.5 and 3.6. To interpret the results, here we defined “low noise” as the noise level was lower than 1 (i.e., the noise level used during training) and “high noise” as the noise level was greater than 1. In the range of low noise, 96.71% of the DenseNet169 configurations (93.20% for ResNet50, 93.40% for VGG19, 77.87% for MobileNetV2, and 75.78% for AlexNet) produced informative hidden representations when the performance of the FCNNs was at chance level ( $p > 0.05$  by a permutation-based t-test). This was due to the FCNN being sensitive to the removal of the noise when it was tested with clearer images. Recall that FCNNs are known to be very sensitive to image perturbations (Kubilius et al., 2019). Critically, when the level of noise was high, 43.72% of the VGG19 configurations (41.67% for ResNet50, 37.35% for DenseNet169, 34.42% for MobileNetV2, and 33.51% for AlexNet) produced informative hidden representations when the performance of the FCNNs was at chance level (Figure 3.7). The performance of the SVM decoding the hidden layer representation was statistically significant overall across all the noise levels ( $p < 0.001$  in a permutation-based t-test), and it was also significant when the noise level was low ( $p < 0.00001$ ) or high ( $p < 0.001$ ).

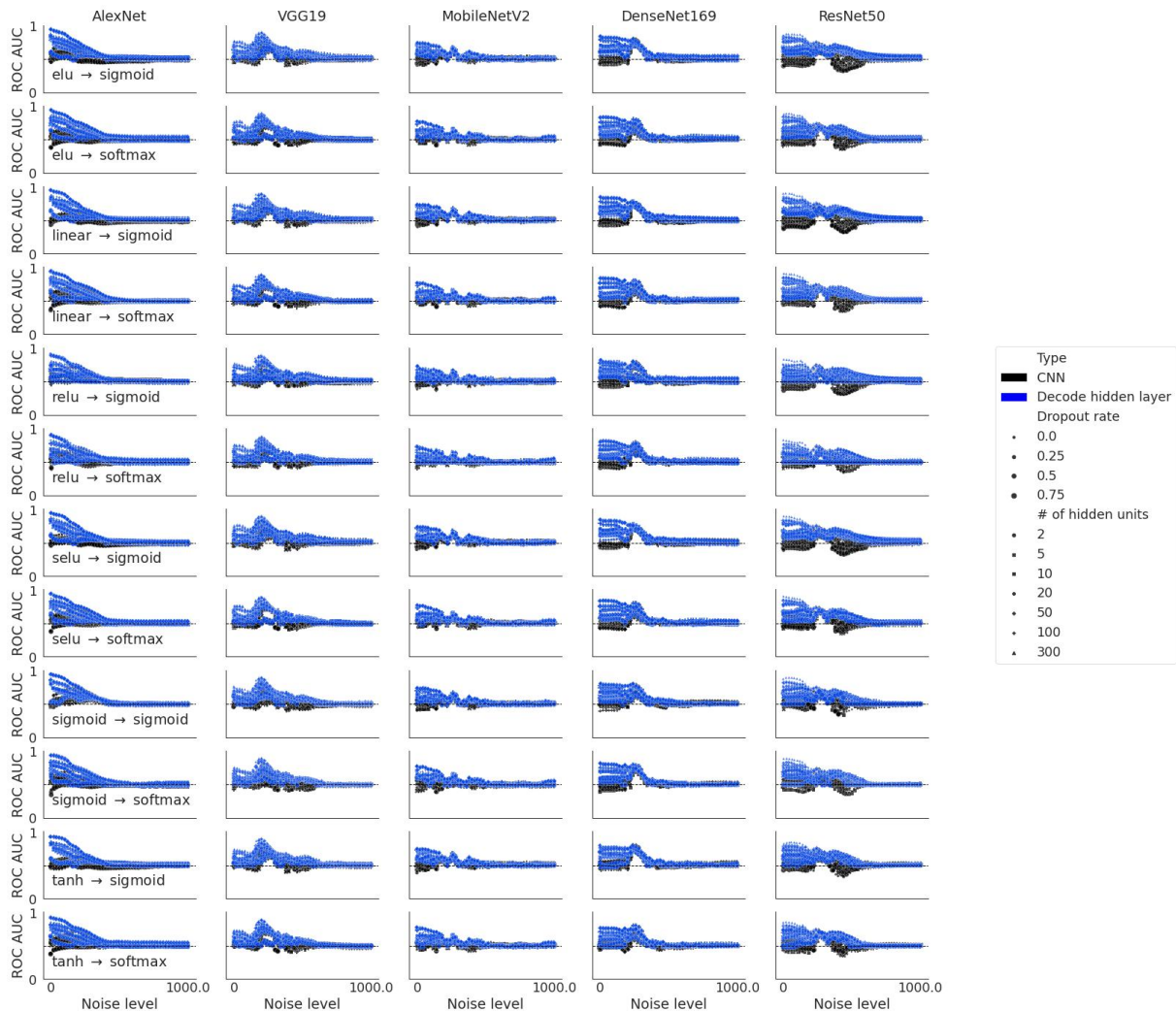


FIGURE 3.5: Performance of the FCNNs and the SVM decoding the hidden representations, when the FCNNs were trained with images embedded in standard Gaussian noise. Color blue depicts the ROC AUC scores of the FCNNs and color black depicts the ROC AUC scores of the SVMs. Different sizes depict the different dropout rates and different shapes depict different numbers of hidden units.

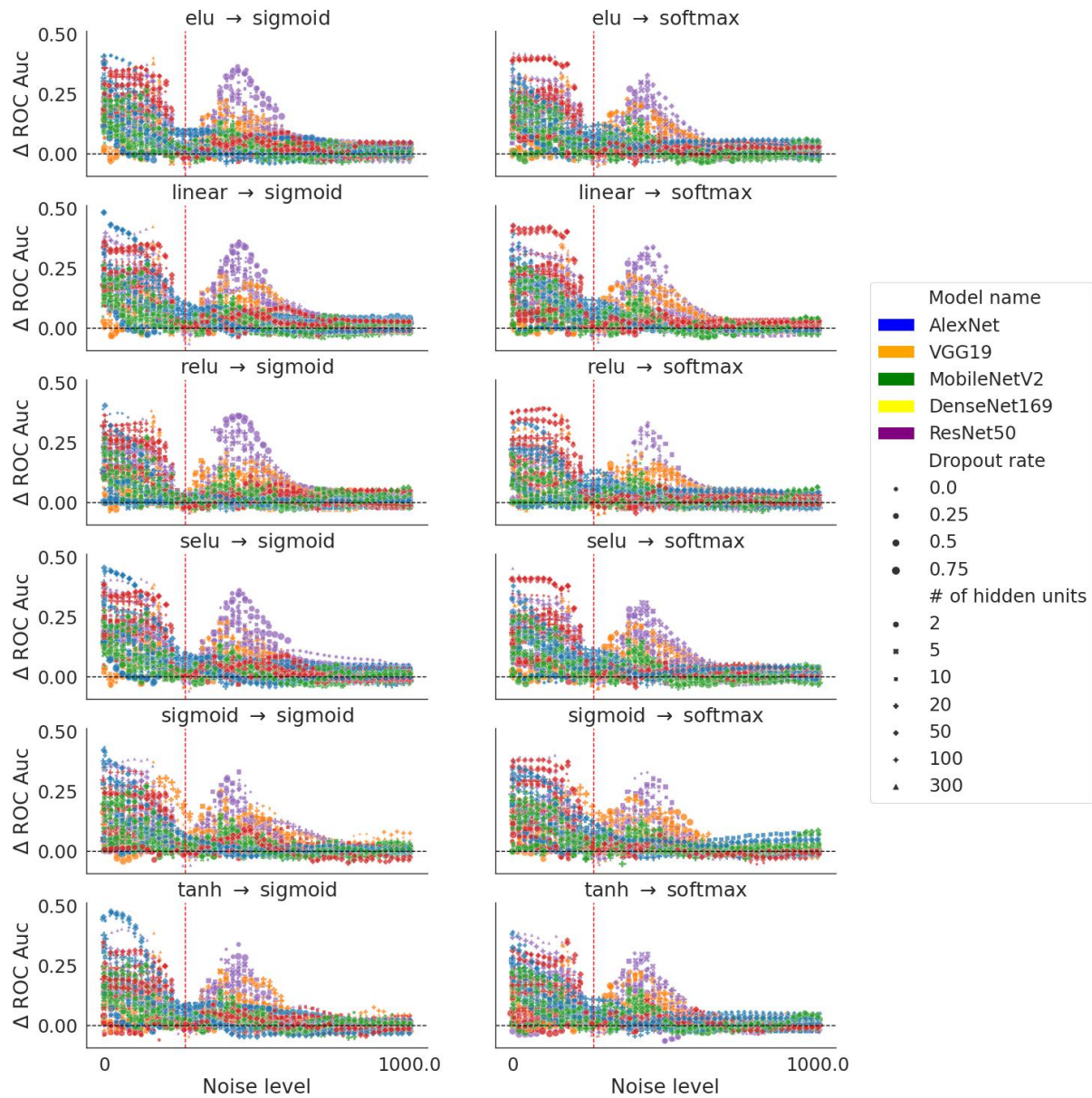


FIGURE 3.6: Difference of performance between the FCNNs and the SVM decoding the hidden layer representation. We subtracted the performance of the FCNNs from the performance of the corresponding SVM. Different colors represented different computer vision models, different sizes represented different dropout rates, and different marker types represented different hidden layer sizes. The small titles on top of each subplot showed the activation functions of the hidden layer and the output layer.

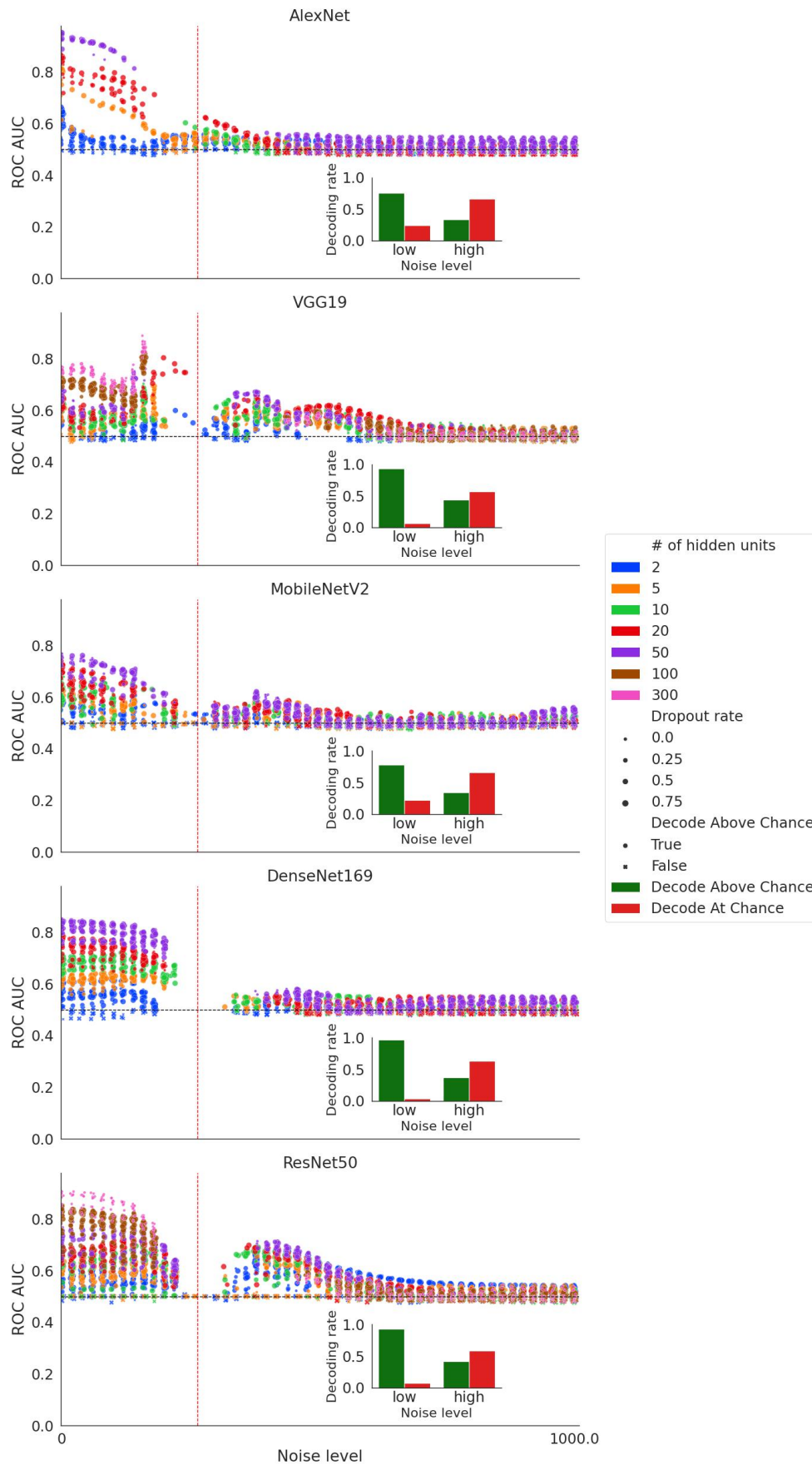


FIGURE 3.7: The performance of the SVM decoding the hidden representations when the FCNNs performance was at chance level. The vertical red dash line depicts the noise level that was used during training. Low and high noise levels were defined based on this. The smaller figure within each subplot shows the proportion of FCNNs configurations that produced informative hidden representations when the noise level was low or high.

Additional analyses were conducted to test whether the image categories could be decoded from the first layer activity pattern. Since the convolutional layers were frozen during the simulation, testing the decoding of the image categories from the first layer was redundant across different FCNN configurations (i.e., across different numbers of hidden units or activation functions). For the analysis of the first layer, we simply added a new output layer on top of the original model that was trained on the ImageNet dataset. We trained the newly added output layer, which mapped 1000 outputs (ImageNet dataset categories) to 2 outputs (animate v.s. inanimate) on the clear images. The training only affected the weights associated with the last layer of the original CNN models and the newly added output layer. The output layer activation function was softmax and the loss function for training was the binary cross entropy. After the training, we froze the model, and then, this was tested with images of different levels of noise following the pipeline of the first simulations. During the testing, the activity patterns of the first layer were recorded for further analyses. The image sizes were  $128 \times 128 \times 3$  and the output of the first layer for the VGG19 was  $127 \times 127 \times 64$  (i.e., with a dimensionality of over one million features after flattening). Relative to the number of trials (i.e.,  $96 \times 20 = 1920$ ), the decoding matrix for a linear SVM classifier with an L2 regularization implemented by scikit-learn was computationally intensive. In a unit test, it took the classifier 20 minutes to train on 80% of the data and test on 20% of the data. For a standard 50-folds cross-validation, it took about 16 hours. To reduce the computational cost, we applied a principal components analysis (PCA) implemented by scikit-learn to reduce the dimensionality of the data. The PCA algorithm was set to keep the components that represented 90% of the variance. Then a linear SVM classifier with L2 regularization was applied to decode the image categories (animate vs inanimate) in 50-fold cross-validation. The discrimination performance of the FCNN and the SVM decoding performance based on the first layer activity patterns were compared to their corresponding empirical chance levels to assess the statistical significance using the same non-parametric t-test conducted for the analyses of the hidden layer. The results showed that the image categories could be decoded from the first layer and decoding performance decreased with increasing levels

of noise in the image. Remarkably, even when the FCNN performed at chance level with noisy images, the image category could be decoded from the first layer activity patterns. These results are presented in Figure 3.8.

Similar patterns were observed: when the FCNNs failed to classify the noise image embedded in a high level of noise, a linear SVM could decode the image categories from the first convolutional layer representations from these FCNNs. The results imply that the FCNNs pretrained on the ImageNet dataset have learned organized patterns that could extract informative representations from noisy images even from the very first layer. These representations are insensitive to noise.

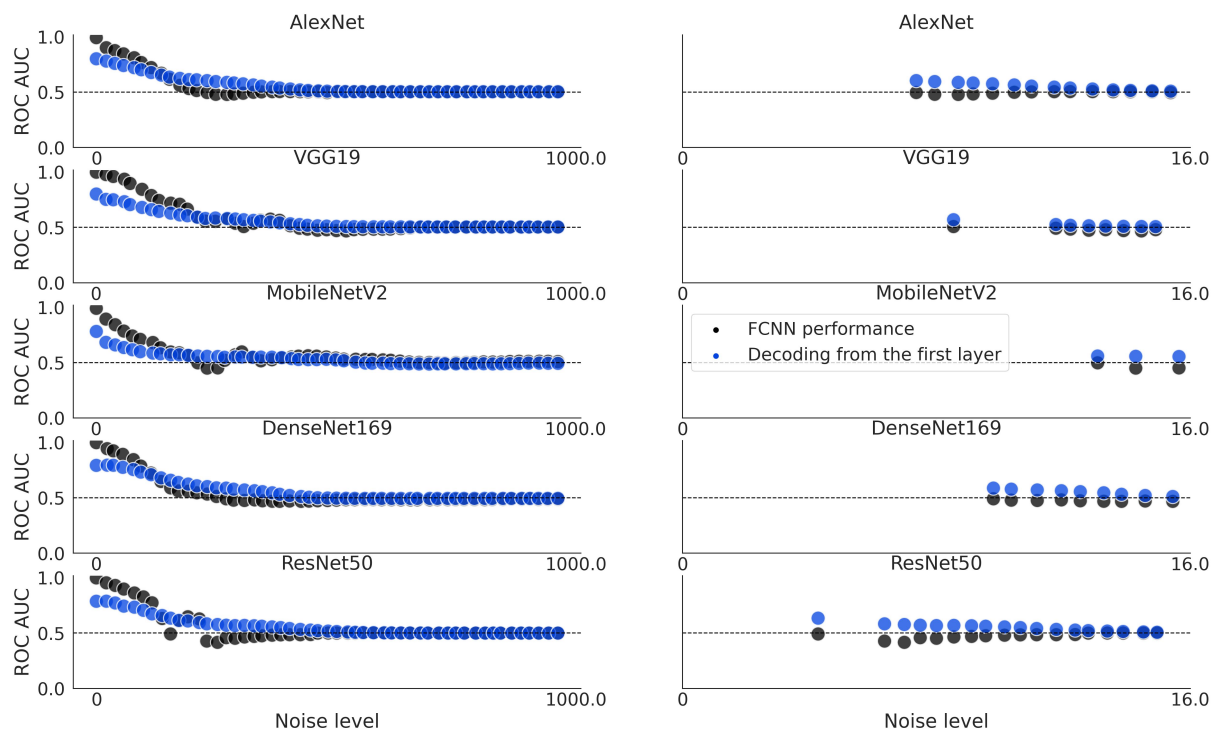


FIGURE 3.8: The left part depicts the FCNN performance (black dots) relative to the SVM decoding performance based on the first layer activity patterns (blue dots). The right part illustrates the range of noise at which the FCNN is at chance level but the SVM could decode significantly above chance the image categories.

### 3.4 Discussion

In the current chapter, we conducted a computational simulation of the fMRI experiment using FCNNs at the representational level. We observed that when the FCNNs failed to

classify the noisy images, we could still decode the noisy image categories from the hidden layer representations of the FCNNs. Additionally, this was replicated with FCNNs trained with adversarial training (i.e., trained with noisy images).

FCNN is arguably one of the many computer vision models that may explain the computations of the ventral visual pathway of the brain (Kriegeskorte, 2015; LeCun et al., 2015; Lindsay, 2021). The results of this chapter provide a simplified computational modeling approach to studying the processing of unseen contents. The simulation approach allowed us to explicitly formulate the properties of the visual unconscious processing (e.g. model backbones, hidden layer size) and test how different features of the properties influenced the processing of unseen items. We tested a combination of five FCNN model backbones, seven hidden layer sizes, six hidden layer activation functions, four dropout rates, and two output layer activations under fifty-one noise level conditions. The wide range of simulation space allowed us to observe how these components of the FCNNs influenced the capacity of the FCNNs' hidden representations to retain information for decoding. The results of the feature importance test of these components showed that besides the noise levels, the CNN backbone model determined the capacity of the hidden layer to retain information for decoding when the FCNN performed at chance level. Kriegeskorte (2015) argued that FCNNs can map complex image features to the goal-driven decision outputs (i.e., living v.s. nonliving). If the FCNNs' hidden layer had a sufficient number of hidden units, the FCNNs could learn the representations of the concepts. The FCNNs mapped the complex image features to the outputs by capturing the different, increasingly abstract levels of representation of the information through the different layers of the network, and importantly, these dynamics are considered similar to the different levels of the ventral visual pathway (Yamins et al., 2013; Yamins & DiCarlo, 2016). Schrimpf et al. (2020) established the brain-score competition for researchers to use many different models to predict different levels of the vision system in primates (human and nonhuman). The representations captured by each layer of the FCNNs were often good predictors of brain representations. These results supported the FCNNs being a reasonable choice of model to simulate the brain representation of visual contents. The



---

simulations conducted in this chapter provided a computational approach to studying the representation of unseen (i.e unrecognized) images. When the FCNNs classified the noisy images at chance level and the noisy image categories could be decoded from the hidden layer representations, the information contained in the hidden layer was not used for the last layer predictions. In other words, information was not read-out by the last layer. The results provided a computational account for the phenomena we observed in the fMRI experiment. When the subjects were unconscious of the image class and their perceptual sensitivity was at chance level, we could decode the image categories from their fusiform gyrus and the middle frontal gyrus. This could also be due to a read-out problem between perceptual information processes at a non-conscious level and how decision mechanisms in the brain use the information to achieve categorical decisions.

# 4 An information-based approach to study neural signatures of conscious and unconscious contents

## 4.1 Introduction: using information-based methods to study conscious and unconscious processing

Kriegeskorte et al. (2006) proposed that researchers should take into greater consideration of the larger, distributed spatial patterns of brain activity, in order to increase the sensitivity in modeling neural activity. By combining a local sphere searchlight method, the multivariate statistics improve the sensitivity of the modeling. A searchlight is a procedure in which a classifier is fit and tested separately for a subset of voxels defined by a moving sphere, and this results in a whole brain decoding accuracy map. Instead of a classifier, a representational similarity analysis using a computational model (see below) can also be used. This approach is now in the class of information-based approaches.

Representational similarity analysis (RSA) quantifies the geometry of the multi-voxel representations using a collection of pairwise item dissimilarity measurements among BOLD responses in a given region of interest (Diedrichsen & Kriegeskorte, 2017; Kriegeskorte et al., 2008a). RSA characterizes the information patterns that come from different aspects (Kriegeskorte et al., 2008a). The activity-pattern information could be measured in the so-called representational similarity analyses by distance-based measurements such as Euclidean distance (Kriegeskorte et al., 2006). The distance-based approach compares

the distance between pairs of activity patterns relative to a computational model. The distances indicate the relative similarities among pairs of activity patterns, including computational model patterns. Thereby allowing comparison across different levels of analyses and different principles of information processing.

The deep convolutional neural network is one of the classes of deep neural network models in computer vision that had been used to explain brain activity (Güçlü & van Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014; Kriegeskorte, 2009; Seeliger et al., 2018). In this chapter, in order to quantify the voxel-wise responses to the masked images, we compared the information patterns captured by the state-of-the-art computer vision models (Schrimpf et al., 2020) and the information patterns captured by the fMRI. Thus, we conducted a standard RSA and an encoding-based RSA, which are explained below.

## 4.2 Methods

### 4.2.1 Representations of the images from the state-of-the-art computer vision models

The information patterns captured by the computer vision models were quantified by the representational dissimilarity matrix (RDM, namely model RDM). First, we fine-tuned the pretrained models with the Caltech101 dataset (Fei-Fei et al., 2004) to control for the size of the hidden representations (i.e., to match the number of units across models for the RSA). A 2D adaptive average pooling was applied to the last convolutional layer of the FCNNs so that the outputs became one-dimensional vectors (He et al., 2015)<sup>1</sup>. The pooled activity patterns of the Alexnet and the VGG19 had 512 units, MobilenetV2 and DenseNet169 had 1024 units, and ResNet50 had 2048 units. Then, a fully-connected layer with 300 units was added to the outputs of adaptive pooling, followed by a scaled exponential linear unit function (SELU, Klambauer et al., 2017)<sup>2</sup>. This allowed us to rescale the representational dimensions of different models to the same dimension. And

---

<sup>1</sup>`torch.nn.AdaptiveAvgPool2d((1,1))`

<sup>2</sup>`torch.nn.SELU(inplace=False)`

SELU was chosen to speed up the training of the new components and scale the output activity patterns for better propagating information to the next layer.

A fully-connected layer with 96 units (1 per category of the Caltech dataset) was added to the outputs of the SELU layer, followed by a softmax activation function to compute probabilistic predictions of the image categories. The model was trained using 96 unique categories of the Caltech101 images (BACKGROUND\_Google, Faces, Faces\_easy, stop\_sign, and yin\_yang were excluded, Fei-Fei et al., 2004). The convolutional layers were frozen during the training. The loss function was binary cross entropy and the optimizer was stochastic gradient descent. The data was split into train and validation partitions and the training was terminated if the performance on the validation data did not improve for five consecutive epochs.

We then fed the trained FCNNs with exactly the same images used in our experiment but without the noise background, in order to extract the hidden representations of the images. We then average the hidden representations of the trials belonging to the same item (i.e cat) and computed the model RDMs for the AlexNet, the VGGNet, MobileNetV2, the ResNet50, and the DenseNet169 (Figures 4.1, 4.3, 4.2, 4.4, and 4.5 depict the RDMs for these hidden representations), and it appears to be clear clusters for the animate and inanimate image categories. Although it might be difficult to visually inspect for AlexNet, the animate images were highly similar among each other and the inanimate images were highly similar among each other, forming clusters at the lower left and upper right quadrants. The similarities were more consistent among animate images (lower left quadrant) compared to the inanimate images (upper right quadrant).

Using these RDMs, we further conducted a standard RSA (section 4.2.2) and an encoding-based RSA (section 4.2.3) on the fMRI data within each subject.

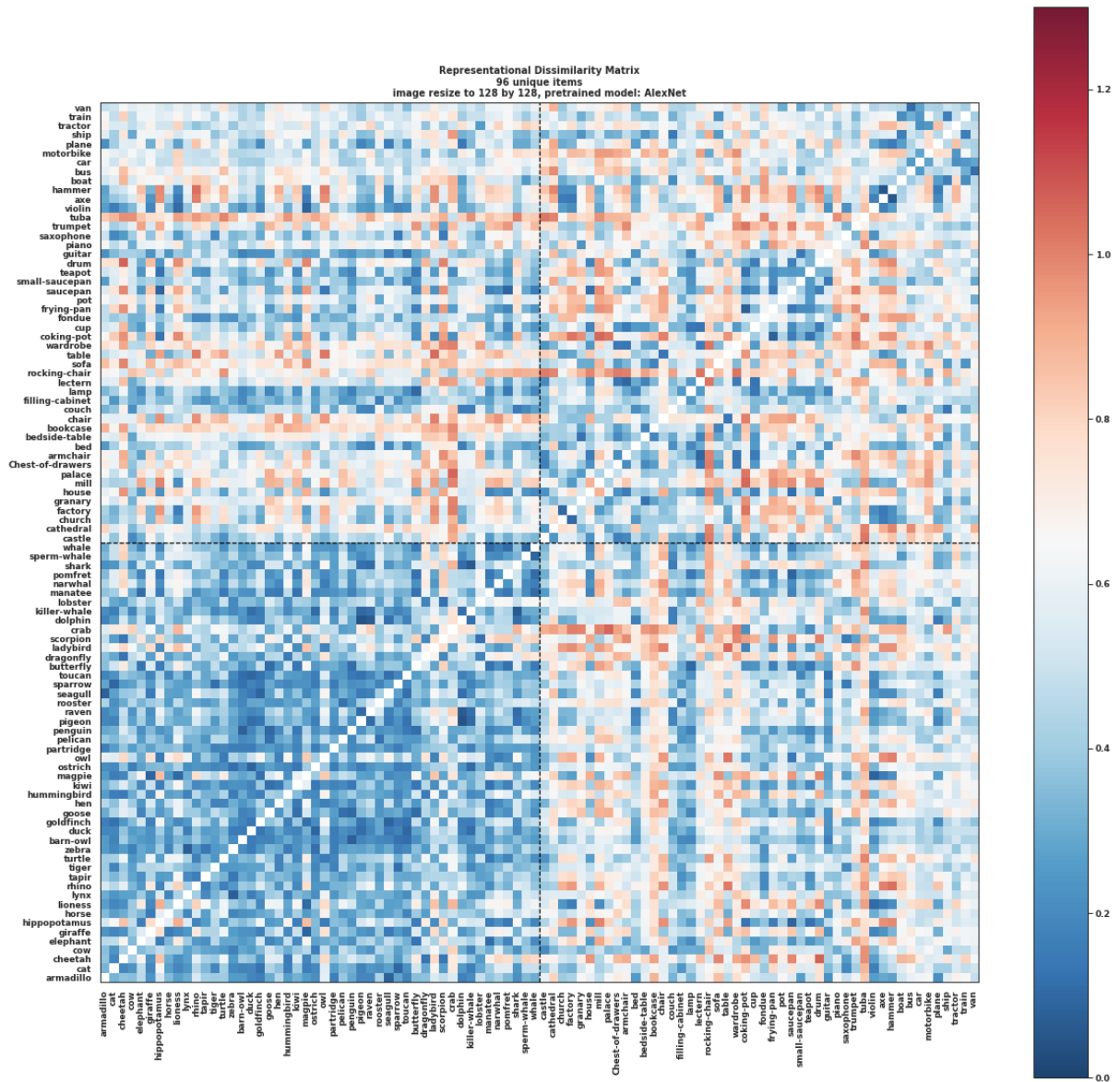


FIGURE 4.1: Representational dissimilarity matrix of the hidden representations of the AlexNet model fine-tuned by Caltech101 dataset. The images were resized to  $128 \times 128 \times 3$  and then passed through the model to obtain the feature representations. The RDM was computed by 1 - Pearson correlations of the feature representations. The first 48 items were animate and the last 48 items were inanimate.

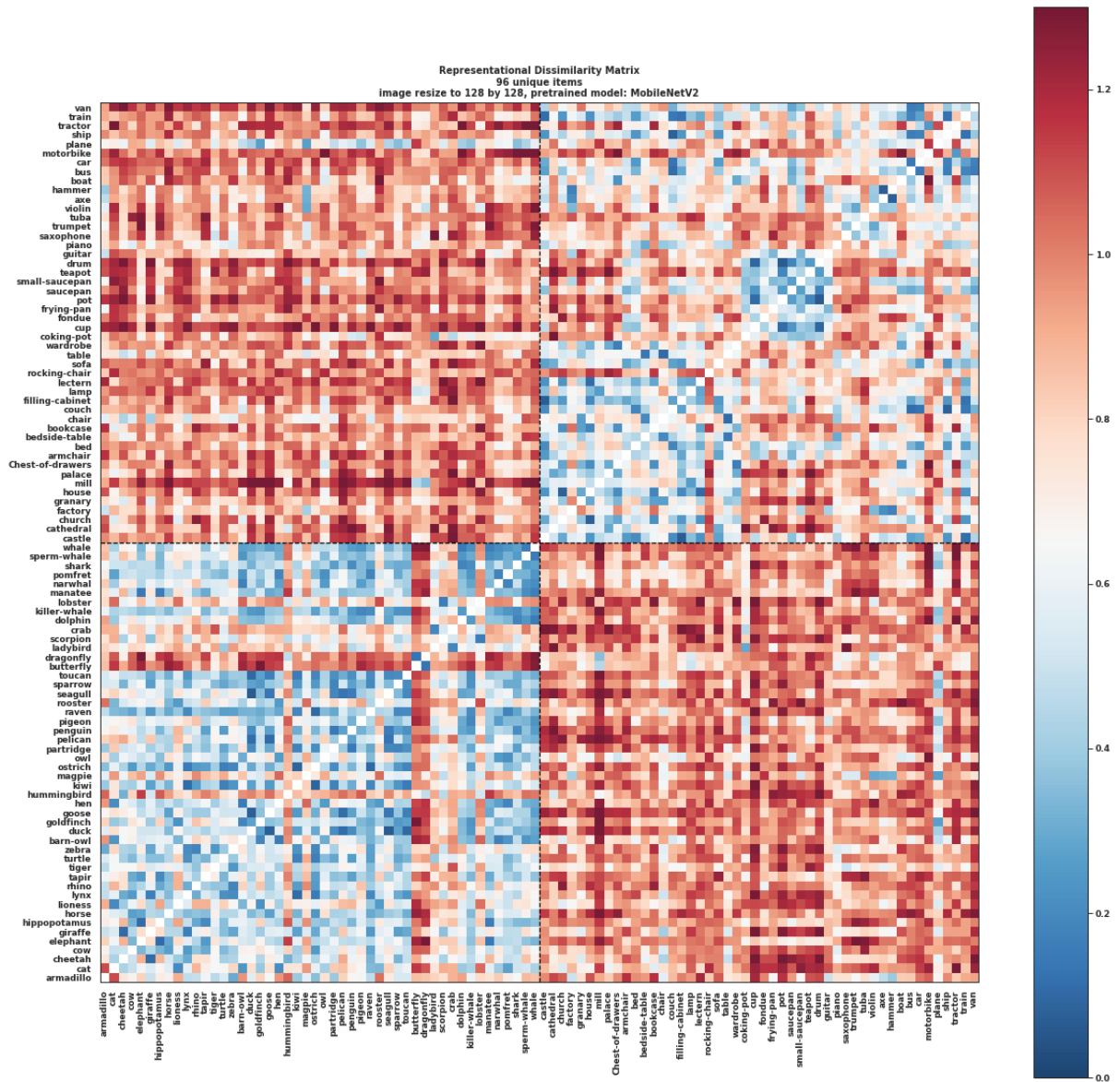


FIGURE 4.2: Representational dissimilarity matrix of the hidden representations of the MobileNetV2 model fine-tuned by Caltech101 dataset. The images were resized to  $128 \times 128 \times 3$  and then passed through the model to obtain the feature representations. The RDM was computed by  $1 - \text{Pearson correlations}$  of the feature representations. The first 48 items were animate and the last 48 items were inanimate.

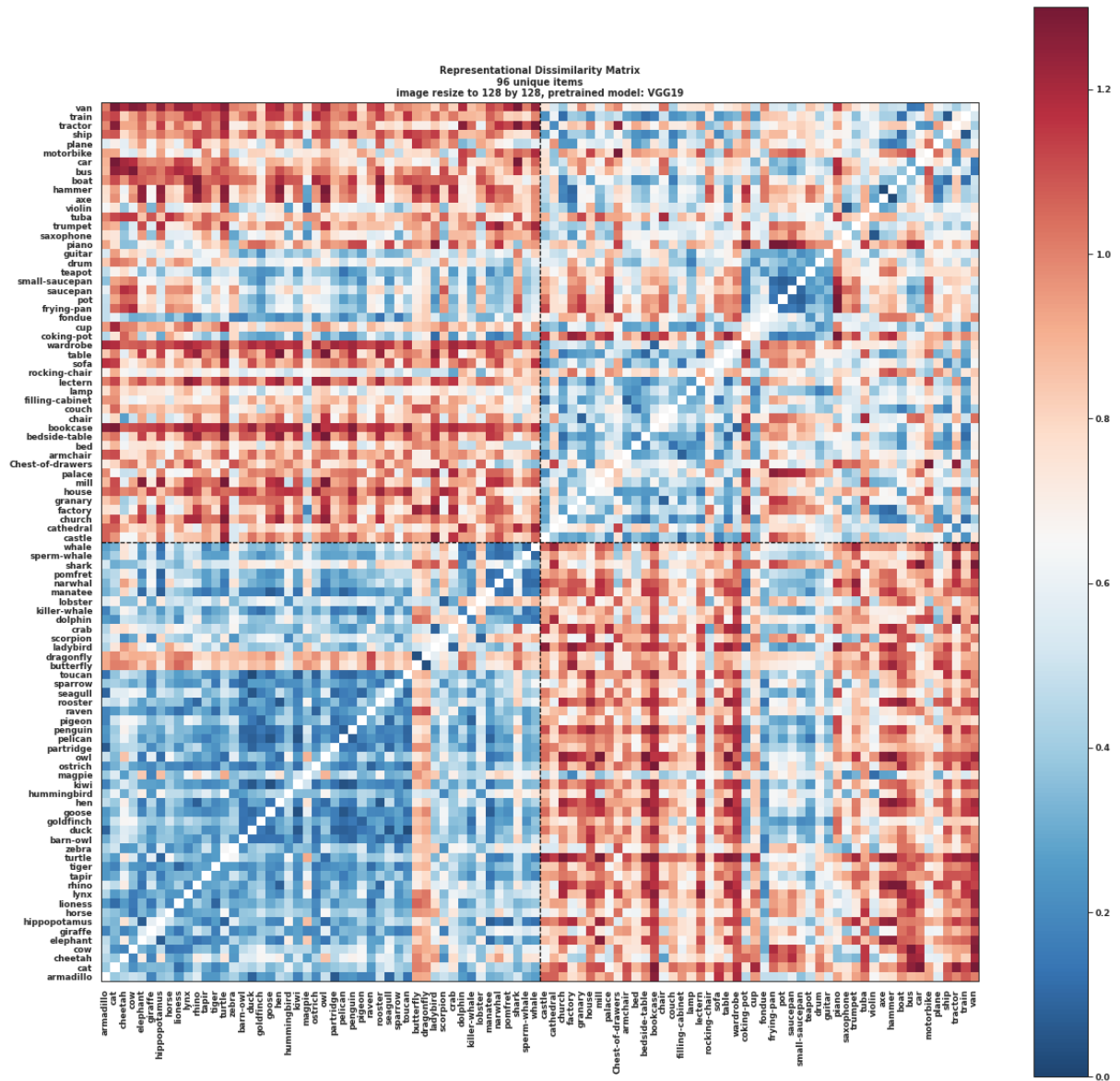


FIGURE 4.3: Representational dissimilarity matrix of the hidden representations of the VGG19 model fine-tuned by Caltech101 dataset. The images were resized to  $128 \times 128 \times 3$  and then passed through the model to obtain the feature representations. The RDM was computed by 1 - Pearson correlations of the feature representations. The first 48 items were animate and the last 48 items were inanimate.

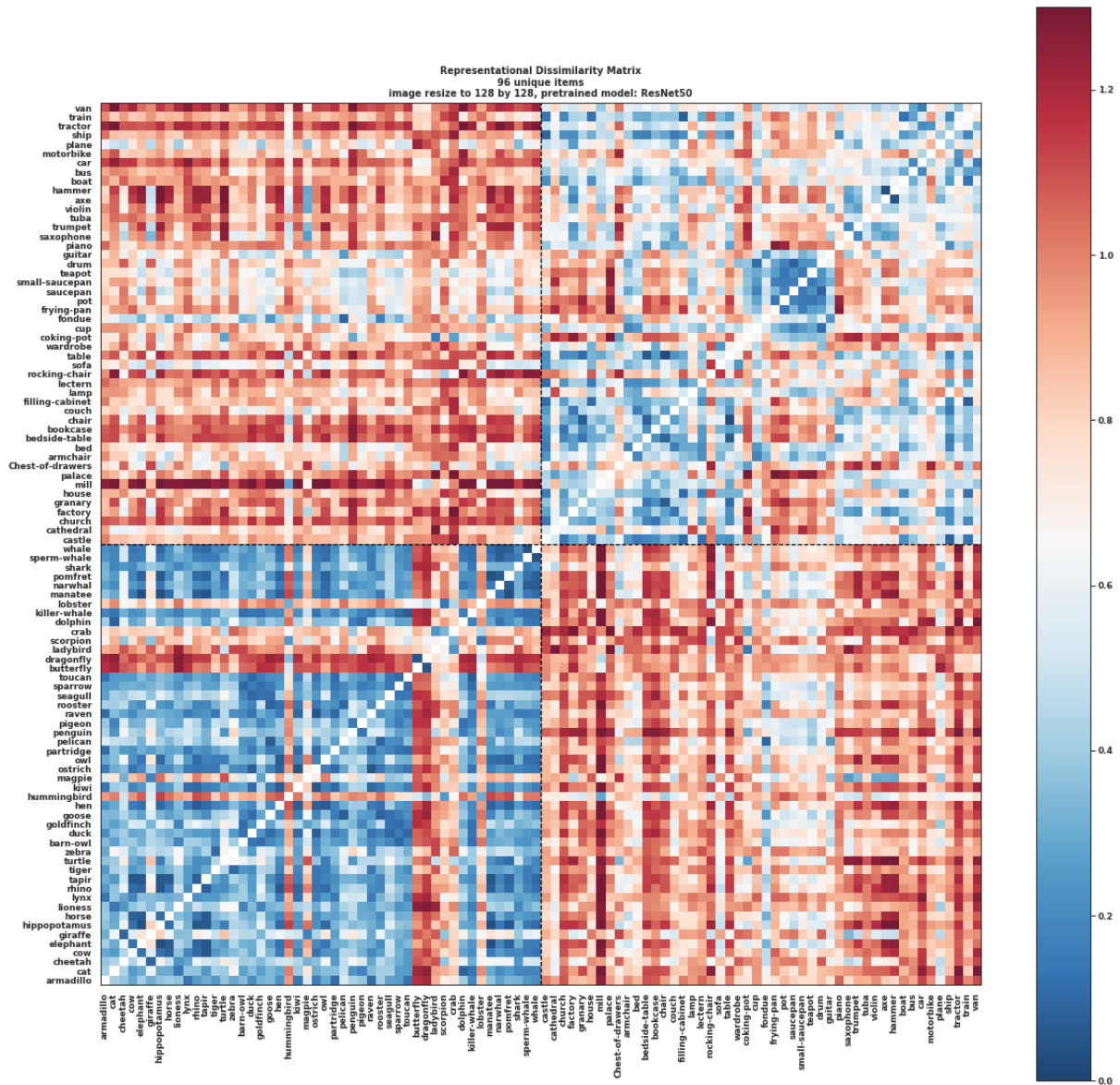


FIGURE 4.4: Representational dissimilarity matrix of the hidden representations of the ResNet50 model fine-tuned by Caltech101 dataset. The images were resized to  $128 \times 128 \times 3$  and then passed through the model to obtain the feature representations. The RDM was computed by 1 - Pearson correlations of the feature representations. The first 48 items were animate and the last 48 items were inanimate.



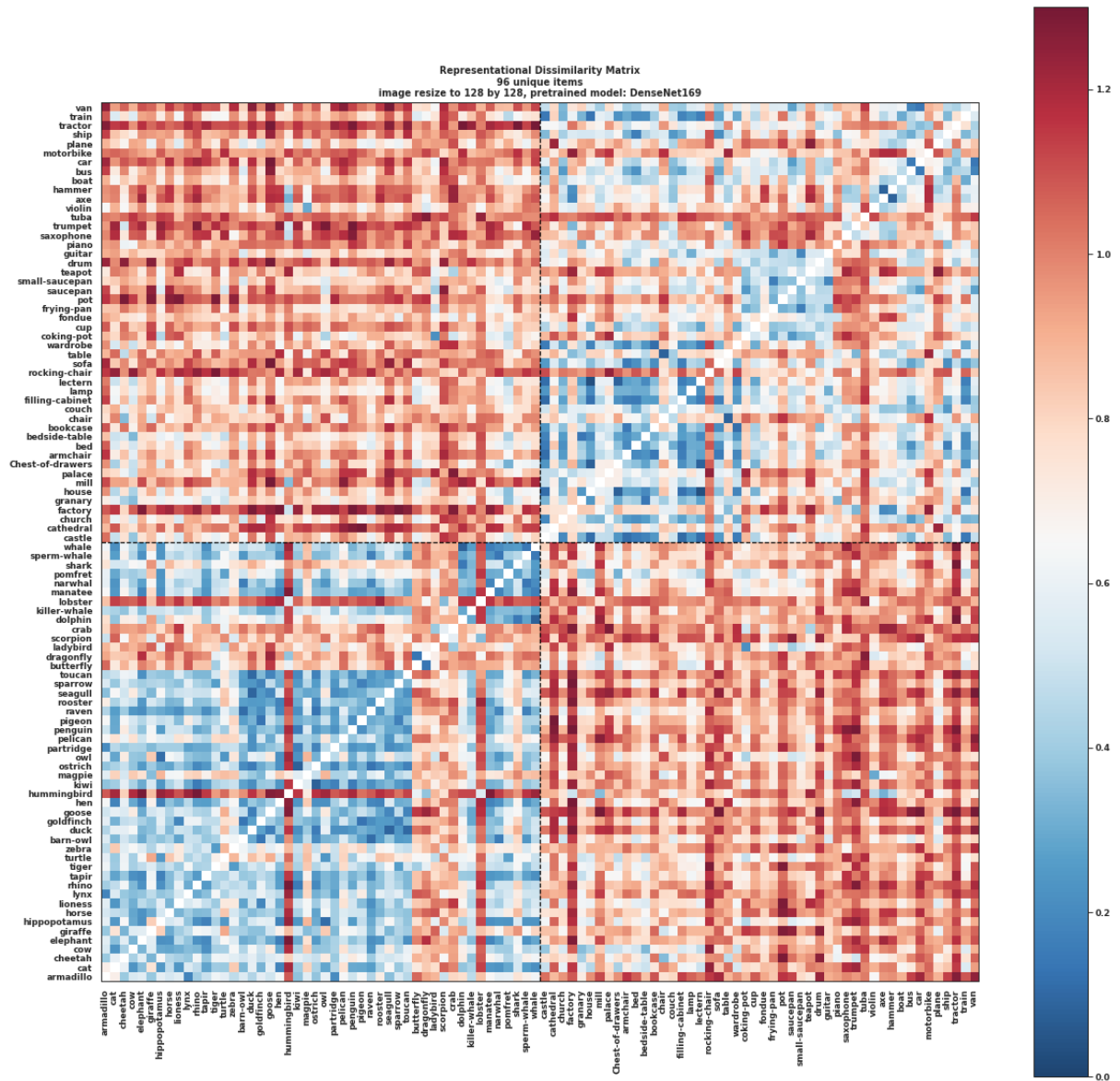


FIGURE 4.5: Representational dissimilarity matrix of the hidden representations of the DenseNet169 model fine-tuned by Caltech101 dataset. The images were resized to  $128 \times 128 \times 3$  and then passed through the model to obtain the feature representations. The RDM was computed by  $1 - \text{Pearson correlations}$  of the feature representations. The first 48 items were animate and the last 48 items were inanimate.

### 4.2.2 Standard whole-brain searchlight RSA

To conduct a whole-brain searchlight RSA, we then extracted the BOLD activity patterns using a mask that contained all the ROIs used for the decoding analyses. We averaged the BOLD signals of trials belonging to the same item (i.e., cat) within a searchlight sphere of 9mm that moved around the brain mask. An RDM was computed for the extracted BOLD signals using 1 - Pearson correlation implemented by Scipy (Virtanen et al., 2020)<sup>3</sup> among each pair of images within each sphere. The lower triangles of the two RDMs were extracted and correlated using Spearman correlation for each sphere of the searchlight. The Spearman correlation coefficient was assigned to the center of each sphere of the whole-brain searchlight. The resulting brain map of FCNNs/brain similarities was converted to standard space for visualization using Nipype (Gorgolewski et al., 2011) and FSL (Jenkinson et al., 2012) tools. This pipeline was run separately in the unconscious and conscious conditions, separately within each participant. Note the fMRI experiment in chapter 2 was designed for within-subject analysis, thus, it was not suitable to further statistical inference on the RSA results.

### 4.2.3 Encoding-based whole-brain searchlight RSA

Konkle and Alvarez (2022) proposed an encoding-based RSA pipeline, adding encoding models on top of the standard RSA pipeline, which can be used to better contextualize the information patterns from the computational model space (i.e., VGG19) to the brain space.

To conduct an encoding-based RSA, we split the data into train and test sets by leaving four unique items out. With 96 unique items, we had twenty-four folds for cross-validation. Within the training set, an L2-regularized linear regression model (ridge regression) implemented by scikit-learn (Hilt & Seegrift, 1977; Pedregosa et al., 2011) was trained to predict the voxel values using the image features in the computational model

---

<sup>3</sup>Scipy.spatial.distance.pdist

space. The image features in the computational model space were the information patterns captured by the computer vision models fine-tuned by the Caltech101 (Fei-Fei et al., 2004). The fine-tuning procedure is described in section 4.2.1. Prior to training the ridge regression model, two scalers were trained using only the training set. One scaler was a `MinMaxScaler`<sup>4</sup> that normalized the voxel values, and the other was a `StandardScaler`<sup>5</sup> that normalized the image features. These scalers were trained in the training set and then applied to the test set data for data normalization to improve the analyzing performance and robustness. The ridge regression model was nested in a grid-search algorithm to cross-validate the best L2-regularization term by stratified random shuffle splitting in the training set using twenty folds. The trained ridge regression model predicted the voxel values in the test set. The predicted voxel values were averaged for each unique item. An RDM of the predicted voxel values and an RDM of the original voxel values in the test set were computed using 1 - Pearson correlation implemented by Scipy (Virtanen et al., 2020) among each pair of items within each sphere. The lower triangles of the two RDMs were extracted and correlated using Spearman correlation for each sphere of the searchlight and the correlation coefficients were assigned to the center of the spheres. This pipeline (Figure 4.6) was run for three within- and cross-conscious state conditions similar to sections 2.2.5 and 2.2.6.

These results are however descriptive since currently there are no robust within-subject RSA analytical pipelines to make a within-subject statistical inference.

---

<sup>4</sup>`sklearn.preprocessing.MinMaxScaler`

<sup>5</sup>`sklearn.preprocessing.StandardScaler`

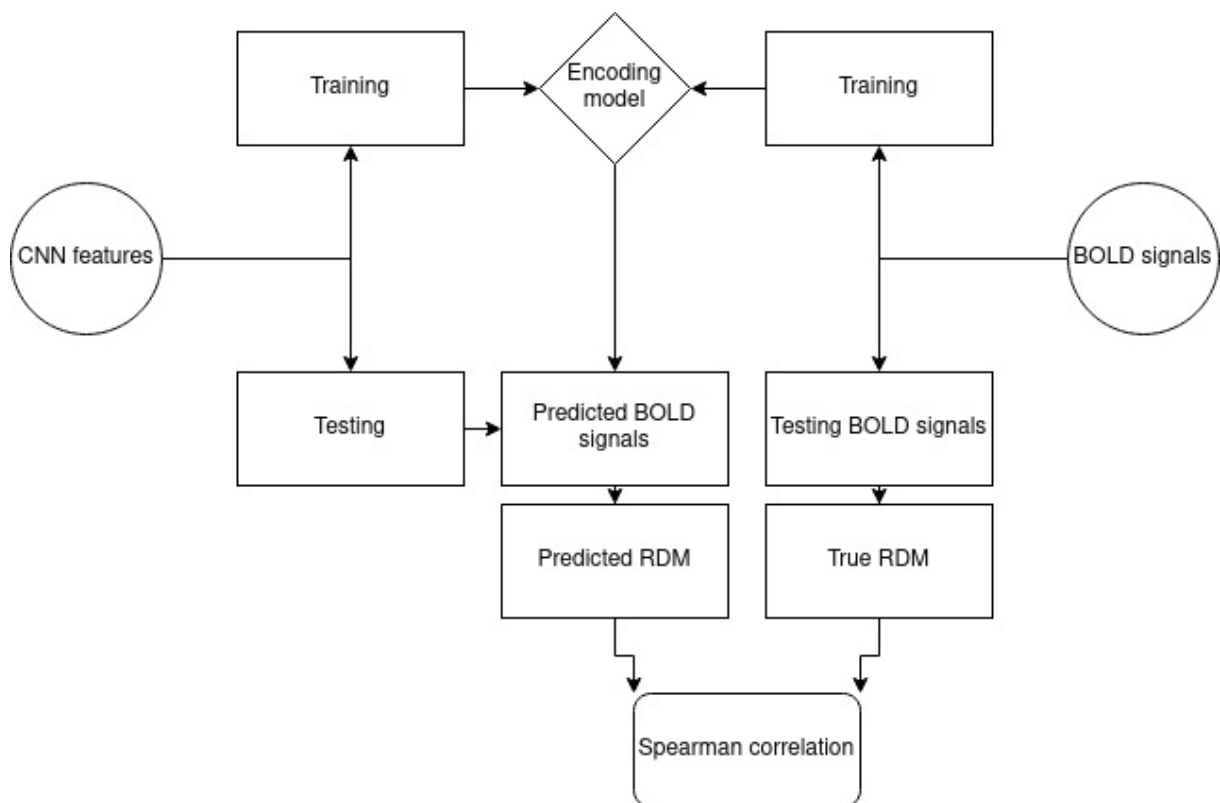


FIGURE 4.6: General diagram of the encoding-based RSA pipeline. The CNN features and the BOLD signals were split into "training" and "testing" sets accordingly for within- and cross-conscious state conditions. The training set was used to train the encoding model (Ridge regression), and the encoding model was used to predict the BOLD signals in the testing set using the testing set CNN features. In the end, we correlated the predicted BOLD signals with the true BOLD signals to obtain the Spearman correlation brain map for further analysis.

## 4.3 Results

### 4.3.1 Standard whole-brain searchlight RSA

We conducted a standard whole-brain searchlight RSA (Kriegeskorte et al., 2008a) to test the information-based approach at the single subject level. We correlated the RDMS of the AlexNet (Figure 4.1), VGGNet (Figure 4.3), MobileNetV2 (Figure 4.2), ResNet50 (Figure 4.4), and DenseNet169 (Figure 4.5) with the voxel responses for unconscious and conscious conditions separately. Note that the results captured by the VGGNet and the ResNet50 models are likely most informative due to their higher ranking in explaining the variance of the human brain (Schrimpf et al., 2020).

In the unconscious condition (Figure 4.7), an RSA based on the AlexNet to extract the model representations of the images showed that six of seven subjects demonstrated model-relevant representations in both the ventral visual pathway and frontoparietal areas. An RSA based on VGG19 representations showed that four subjects had model-relevant representations in the ventral visual pathway, and a different subset of four subjects showed model-relevant representations in the frontoparietal areas. Using the MobileNet, five subjects showed model-relevant representations in the ventral visual pathway, and four subjects showed model-relevant representations in the frontoparietal areas. ResNet50-based RSA showed model-relevant representations in the ventral visual pathway in five subjects, and six subjects showed model-relevant representations in the frontoparietal areas. Finally, the DenseNet169-based RSA showed that three subjects had model-relevant representations in the ventral visual pathway, and four subjects showed model-relevant representations in the frontoparietal areas.

For the conscious condition (Figure 4.8), all subjects showed model-relevant representations in the ventral visual pathway across all models. Six of seven subjects also showed model-relevant representations in the frontoparietal areas across all models.

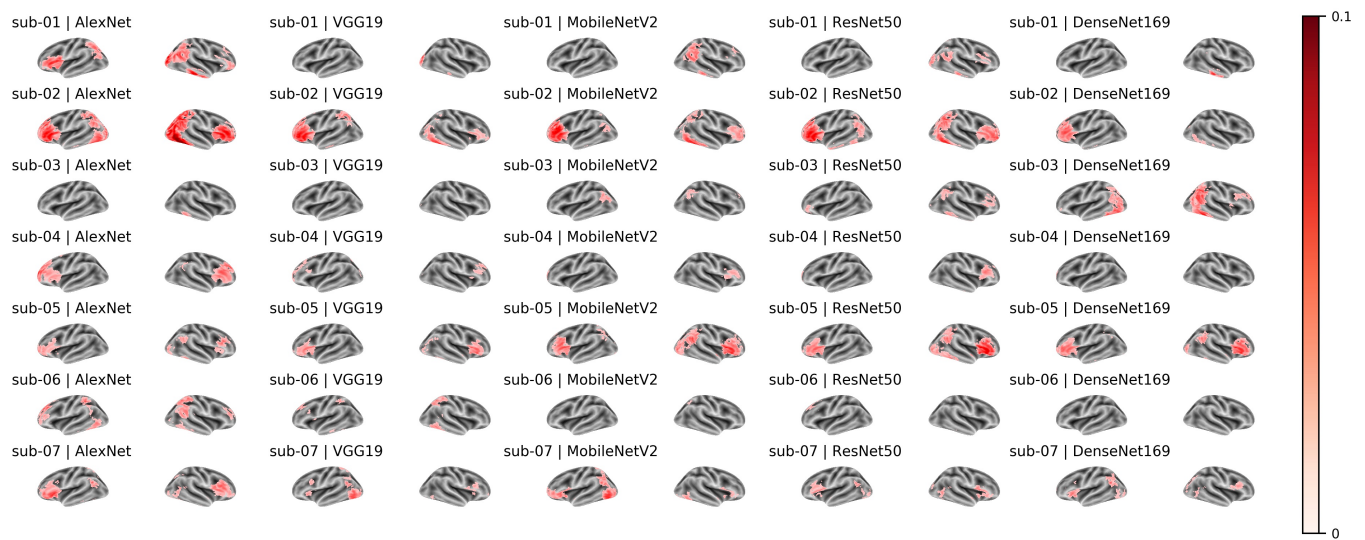


FIGURE 4.7: RSA map of the unconscious condition for the five FCNNs. Each row showed the left and right hemisphere RSA maps of the same subject, and every two columns showed the left and right hemisphere RSA maps of the same computer vision model.

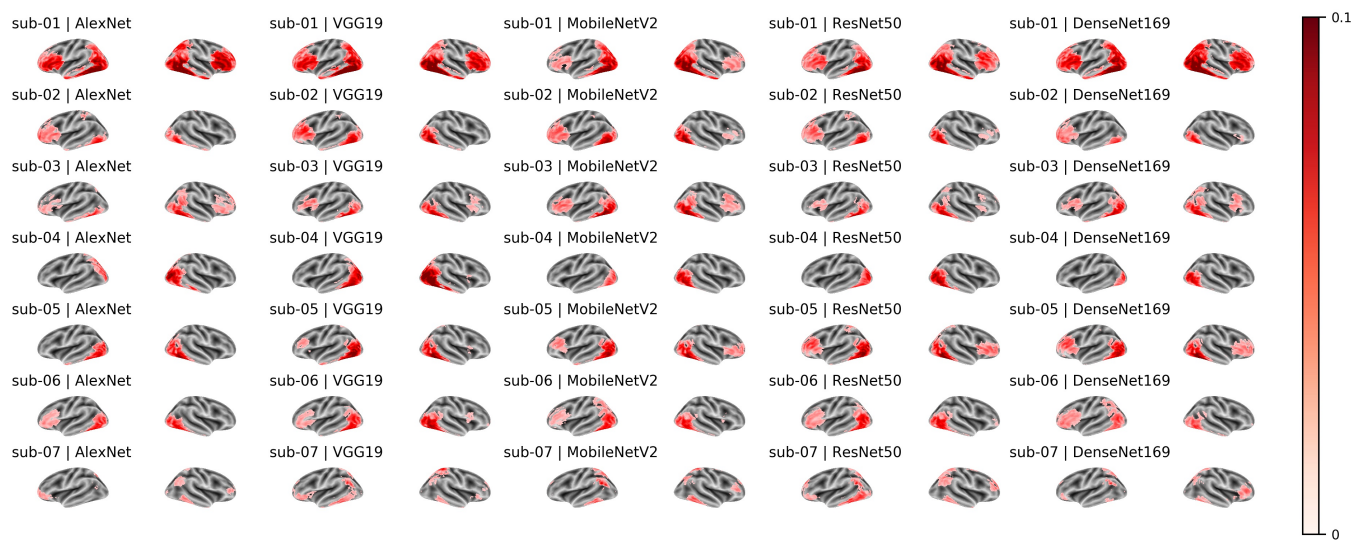


FIGURE 4.8: RSA map of conscious condition for the five FCNNs. Each row showed the left and right hemisphere RSA maps of the same subject, and every two columns showed the left and right hemisphere RSA maps of the same computer vision model.

### 4.3.2 Encoding-based whole-brain searchlight RSA

We also conducted the encoding-based whole-brain searchlight RSA (Konkle & Alvarez, 2022). Compared to the standard whole-brain searchlight RSA, the encoding-based approach first contextualizes the model representations in the source brain state (i.e., conscious) and then measures the model-relevant representations in the target brain state (i.e., unconscious). Thus, the encoding-based approach allows us to measure the model-relevant representations with one extra dimension and test the representational similarity across brain states.

Across different models within the unconscious condition, the computer vision model feature representations were positively correlated to brain representations in the visual areas in four out of seven subjects (column 1 of Figures 4.9 - 4.13). Additionally, we observed that the model feature representations of the AlexNet and DenseNet169 were negatively correlated to the brain representations in the frontal areas for five of seven subjects. The brain patterns correlated to the model were much stronger and more consistent across subjects in the conscious condition. Negative correlations in RSA are difficult to interpret, thus, we only gave a descriptive and speculative interpretation for the moment. The brain representations in the frontal areas of these subjects were similar to the computer vision models, but the brain representations were projected from the visual areas, therefore, the re-represented information "pointed to" a different high-dimensional direction compared to the visual area representations, resulting in a negative correlation. This was also observed within the conscious condition (column 2 of Figures 4.9 - 4.13), where positive correlations were observed in the visual areas and negative correlations were observed beyond the visual areas. As we trained the encoding model in the conscious condition and then transferred the encoded model features to the unconscious condition (column 3 of Figures 4.9 - 4.13), the RSA results could be an enhanced pattern of representations of the unconscious condition, which was contextualized by the conscious brain states. The patterns observed in the transferred encoding-based RSA pipeline from the conscious

condition to the unconscious condition were very similar to those observed within the unconscious condition. This could be due to that the signal-to-noise ratio in the unconscious condition was very low so the univariate encoding models did not enhance the RSA for better pattern representations.

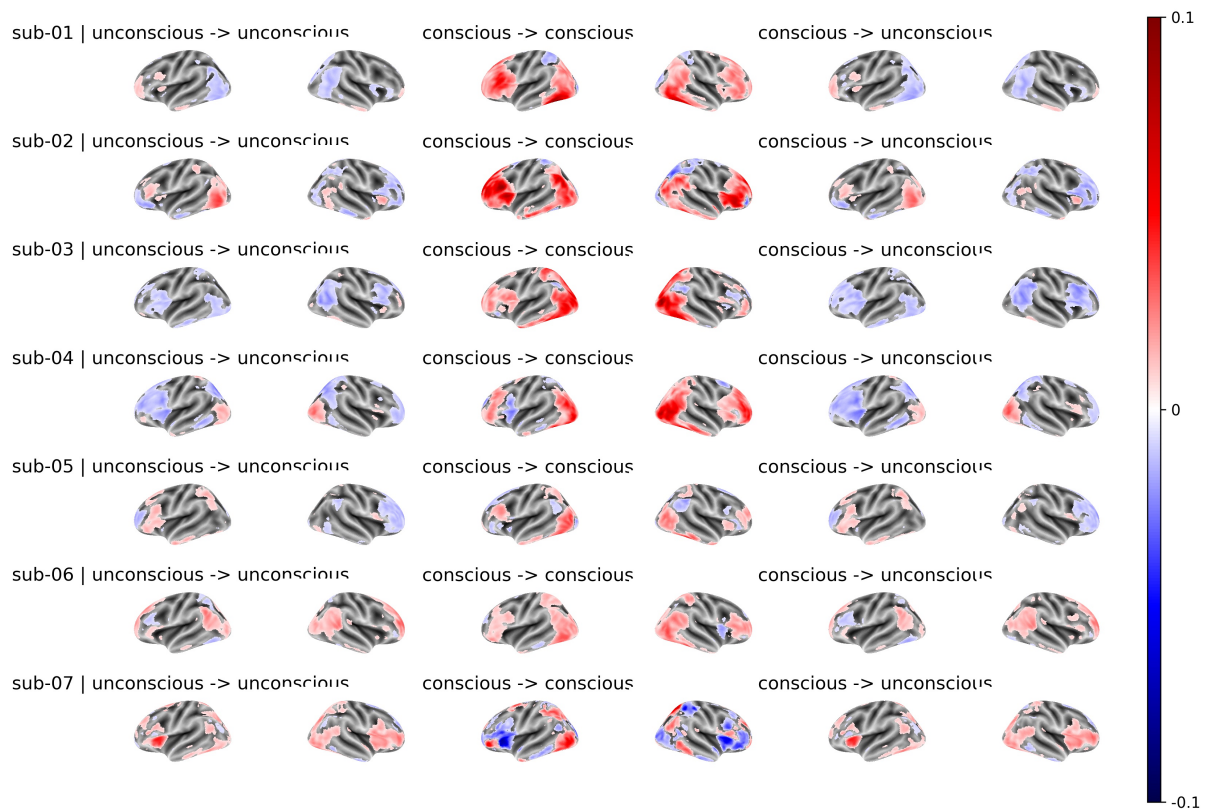


FIGURE 4.9: Encoding-based RSA map of the AlexNet model. The first and second columns showed the RSA maps within the unconscious condition, the third and fourth columns showed the RSA maps within the conscious condition, and the last two columns showed the RSA maps that were contextualized in the conscious condition and then generalized to the unconscious condition.



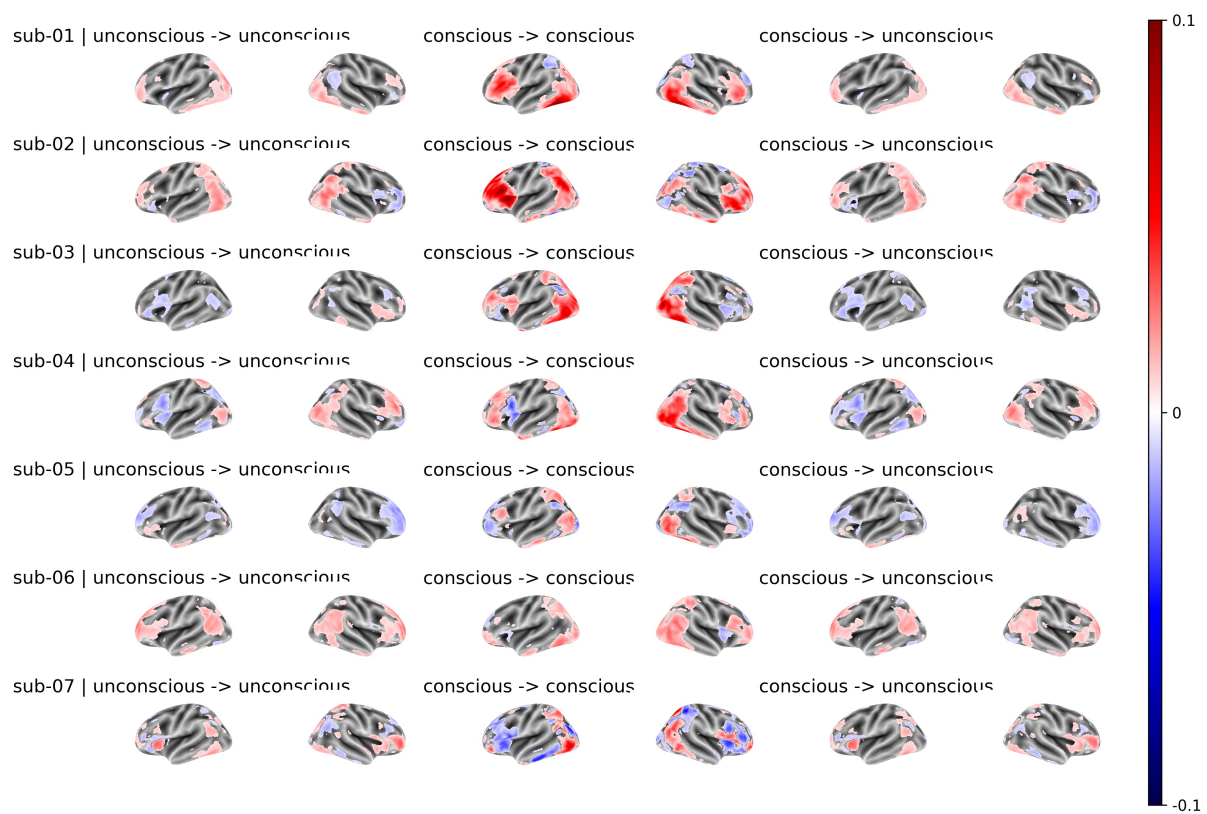


FIGURE 4.10: Encoding-based RSA map of the VGG19 model. The first and second columns showed the RSA maps within the unconscious condition, the third and fourth columns showed the RSA maps within the conscious condition, and the last two columns showed the RSA maps that were contextualized in the conscious condition and then generalized to the unconscious condition.

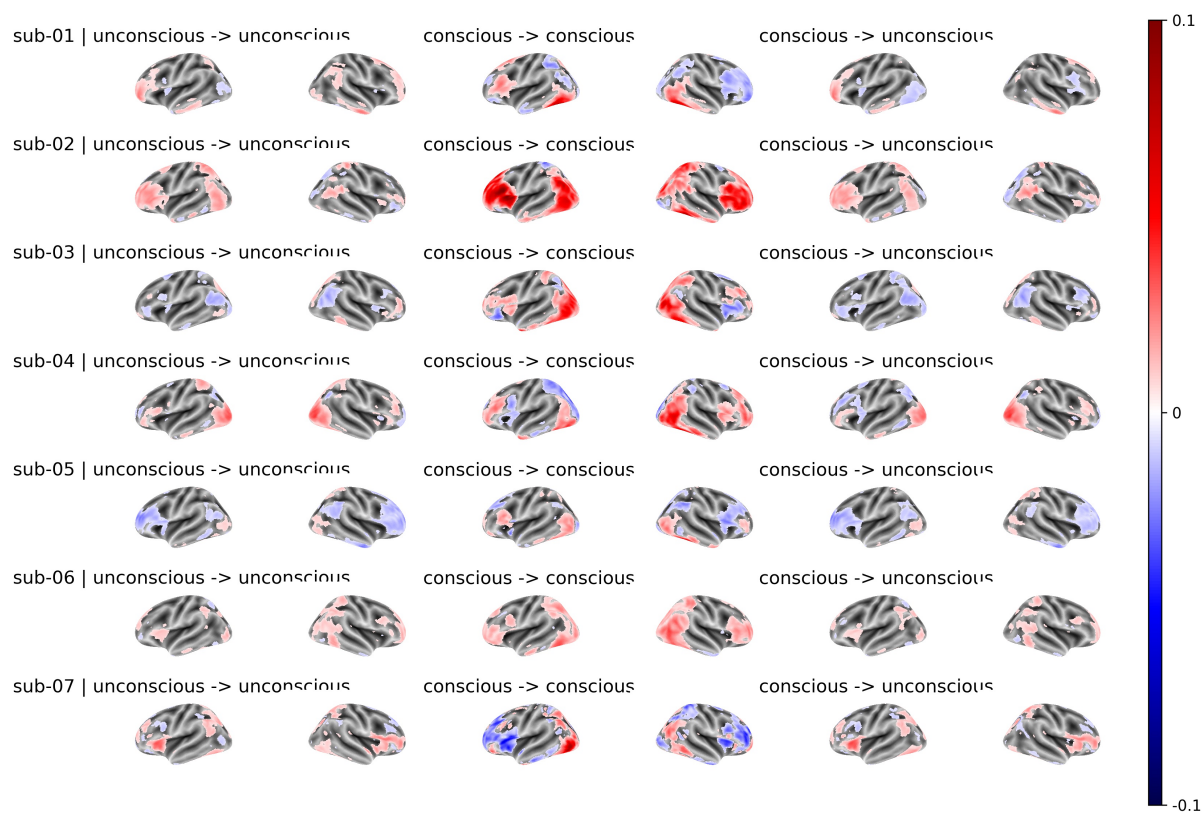


FIGURE 4.11: Encoding-based RSA map of the MobileNetV2 model. The first and second columns showed the RSA maps within the unconscious condition, the third and fourth columns showed the RSA maps within the conscious condition, and the last two columns showed the RSA maps that were contextualized in the conscious condition and then generalized to the unconscious condition.

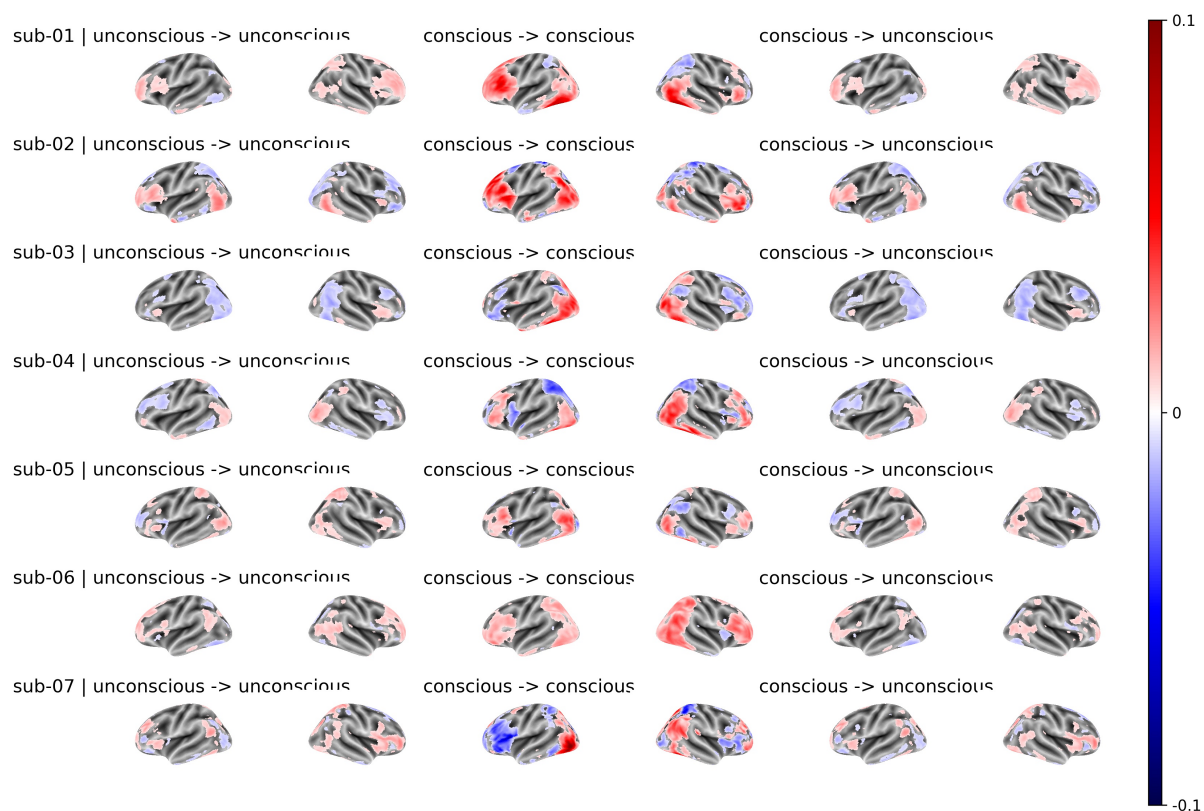


FIGURE 4.12: Encoding-based RSA map of the ResNet50 model. The first and second columns showed the RSA maps within the unconscious condition, the third and fourth columns showed the RSA maps within the conscious condition, and the last two columns showed the RSA maps that were contextualized in the conscious condition and then generalized to the unconscious condition.

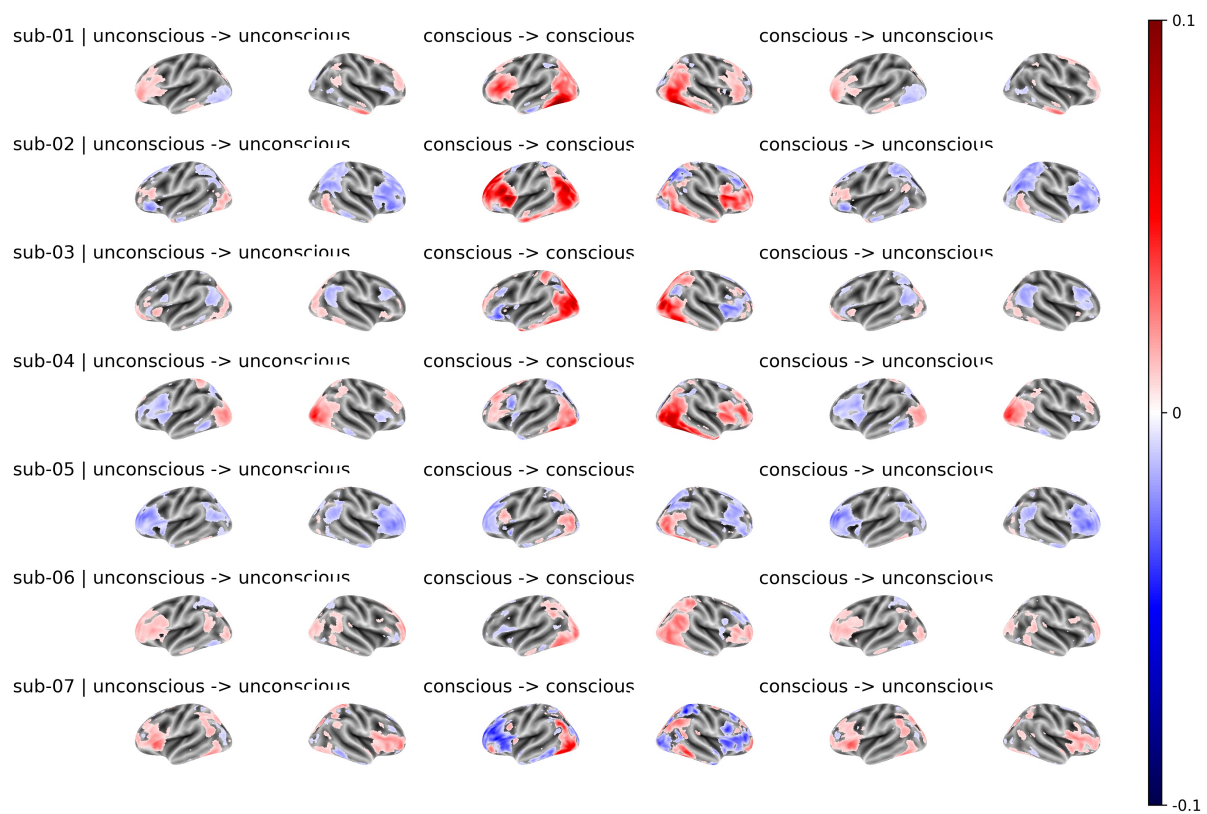


FIGURE 4.13: Encoding-based RSA map of the DensNet169 model. The first and second columns showed the RSA maps within the unconscious condition, the third and fourth columns showed the RSA maps within the conscious condition, and the last two columns showed the RSA maps that were contextualized in the conscious condition and then generalized to the unconscious condition.

## 4.4 Discussion

In this chapter, we explored information-based approaches to investigate the properties of the unconscious and conscious contents represented in the brain. We presented the results of the standard RSA pipeline (Kriegeskorte et al., 2008a) and the encoding-based RSA pipeline (Konkle & Alvarez, 2022). Due to the small sample size of the subject pool and both RSA pipelines were not designed for within-subject design experiments, we offered descriptive results from these RSA pipelines.

The standard RSA results showed that three of the seven subjects' brain activity patterns were correlated to the computational model, including the left prefrontal areas and the occipital areas (i.e., VGG19 and ResNet50) even when the subjects were unconscious of the images. However, the pattern was not robust across all five FCNNs in all subjects. When the subjects were conscious of the images, six of the seven subjects' brain activity patterns were correlated to the computational model in the ventral visual pathway and the frontoparietal areas, and the pattern was consistent across all five models. However, the results of the encoding-based RSA in the unconscious condition were mixed and difficult to interpret. There were negative correlations in the visual areas for some subjects in some of the models (i.e., AlexNet and DenseNet169), and there were negative correlations in the frontal areas for most of the subjects in many of the models (i.e., ResNet50 and DenseNet169). We interpret the negative correlations as the brain representations were associated with the model, but the brain representations of the images have a different structure compared to the computer vision models in terms of how the perceptual input (the images) are decomposed and projected into the model and the brain images are high-dimensional and complex stimuli. In other words, while the hidden representations of the FCNNs of the images are known to explain a large variance of the brain responses (Schrimpf et al., 2020), the way that the brain processes information and encodes relevant features may still be different from the FCNNs. The FCNNs, particularly those trained with supervised methods for image classification (Deng et al., 2009), learn to represent the

images through different layers in a bottom-up fashion, but the brain represents the images not only through bottom-up processes in the visual stream but also through top-down processes, taking information from other higher-order areas (i.e., prefrontal cortex). Thus, representations in the brain could be more visual-oriented or more semantic-oriented in terms of different levels of processing (Popham et al., 2021), which are deeper in the brain than in FCNNs. The higher-order areas of the brain like the prefrontal cortex decode the information projected from the lower-order areas (i.e., ventral visual pathway), and thus the representations become more abstract (Kriegeskorte & Douglas, 2018). The representations in the ventral visual pathway are primarily visual-oriented like the computer vision model, thus, the pattern representations were positively correlated to the FCNNs' hidden representations. On the other hand, when the information was processed and projected to higher-order areas in the frontal cortex, the representations became more semantic-oriented in the brain but less in the model. This may explain that the brain representations in the prefrontal areas were negatively correlated to the FCNNs' hidden representations. The results of the encoding-based RSA are in line with the suggestion that there are representational boundaries between visual-oriented representations and semantic-oriented representations of images in the brain (Popham et al., 2021) and this may not be the case in the FCNNs because it does not incorporate recurrent connections. Additional work is needed to make further determinations.

# 5 General Discussion

## 5.1 Summary of findings and general conclusions

A very important criterion to determine unconscious visual processing is to establish null sensitivity using both subjective measures of awareness collected on a trial-by-trial basis and also objective measures derived from signal detection theory (Soto et al., 2019). Previous fMRI studies of consciousness performed group-level statistical inference from a few subjects and each participant had a low number of trials. These studies were underpowered to test the null sensitivity assumption and hence could not pinpoint the brain representation of the unconscious representation reliably within each subject (Macmillan & Creelman, 2004).

In this thesis, we first aimed to develop sensitive, within-subject neural markers of unconscious and conscious information processing. The high-precision, highly-sampled, within-subject approach allowed us to investigate the properties of the brain representations associated with unconscious perceptual contents. A secondary aim of the thesis was to provide an information-based approach, using machine learning algorithms, pattern matching, and representational similarity analysis using computer vision models. These information approaches allowed us to investigate neural representations of conscious and unconscious contents using both model-/hypothesis-based and data-driven approaches. The third aim of the thesis was to provide a computational simulation of the representations of unseen contents using state-of-the-art FCNNs (Kriegeskorte, 2015; Lindsay, 2021). The simulation approach allowed us to provide a representational model simulation of visual processing/representation associated with null perceptual sensitivity using a simplified approximation of the human brain in which all parameters could be well

controlled.

### **5.1.1 High-precision, high-sampling fMRI data associated with null sensitivity reveal representations of unconscious contents**

Even in those subjects who showed null perceptual sensitivity, unconscious contents could be decoded in the visual, parietal, and frontal areas including the fusiform and the middle frontal gyrus. Additionally, the brain representations of the unconscious contents and the conscious contents were similar to a certain degree. A classifier trained to discriminate the visual categories of the items in the conscious trials generalized to the unconscious state in different regions, including the inferior parietal lobe, inferior temporal lobe, lingual, middle frontal gyrus, and superior parietal gyrus in six out of seven subjects. These patterns were robust across machine learning algorithms defined by different hyperparameters. As discussed in Chapter 2, the high-precision, highly-sampled, within-subject decoding paradigm provided a richer information-based approach to studying the representations of unconscious contents (Kriegeskorte et al., 2006). The paradigm improved the data-driven sensitivity and the statistical power for detecting unconscious brain representations associated with null perceptual sensitivity.

### **5.1.2 A simulation of the fMRI results using feedforward convolutional neural network models**

It has been well documented that FCNNs can explain a great amount of variance of the neural response patterns in the ventral visual pathway (Fukushima, 1980; Güçlü & van Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014; Kriegeskorte & Douglas, 2018; LeCun et al., 2015; Lindsay, 2021; Lindsay et al., 2019; Schrimpf et al., 2020; Yamins et al., 2013; Yamins & DiCarlo, 2016). Thus, it is a reasonable choice to utilize FCNNs to simulate the fMRI results. Accordingly, we had FCNNs perform the image categorization task as in the fMRI experiment, and we manipulated the level of Gaussian noise added



to the images during categorization. The performance of the FCNNs decreased as the noise level increased. When the noise level was high enough to make the FCNNs perform at chance, the categories of the noisy images in the hidden layer could be read-out by a linear machine learning algorithm but could not be read-out by the last output layer of the model.

It was expected that the image categorization performance of the FCNNs was impaired by the addition of noise (Jin et al., 2015), but it was surprising that the information failed to guide the decision of the output decision layer, which resulted in incorrect classifications, could nevertheless be read-out and decoded by an independent classifier (i.e., a linear machine learning algorithm). This means there remain informative traces of the categories of the noisy images in the FCNNs' layers, but the information did not support categorization decisions and the agent's classification performance was at chance level. Thus, information processing of unseen contents in the visual domain, including processing without sensitivity, can lead to meaningful but hidden representational states that are ubiquitous in brains and biologically plausible models based on deep artificial neural networks (Kriegeskorte, 2015; Miconi, 2017; Nonaka et al., 2021; Taherkhani et al., 2020).

### 5.1.3 Neural signatures of conscious and unconscious contents from an information-based perspective

To provide a richer representational account of the unconscious information processing in the human brain, we conducted a standard whole-brain searchlight RSA (Kriegeskorte & Diedrichsen, 2019; Kriegeskorte et al., 2006) and an encoding-based RSA (Konkle & Alvarez, 2022) to investigate how the FCNNs' hidden layer representations of the experiment images correlated to the brain representations.

The standard whole-brain searchlight RSA revealed that the FCNNs' hidden layer representations correlated to the brain representations of the unconscious contents in the ventral visual pathway in four of seven subjects. Particularly for models that ranked high in explaining the variance of the visual cortex (i.e., based on brain-score, Schrimpf et al.,

2020, i.e., VGGNet and ResNet50), five of seven subjects showed model-correlated patterns in the left ventral visual pathway and the left frontoparietal areas in the unconscious condition. In the conscious condition, all the subjects showed model-correlated patterns in the bilateral visual cortex, and six of the seven subjects showed model-correlated patterns in the unilateral frontoparietal areas. Similar to the standard RSA analysis, the encoding-based RSA revealed that the brain representations of the unconscious contents in both the occipital and the frontal areas correlated to the FCNNs' hidden layer representations. For instance, in the VGG19 model-based RSA, six of the seven subjects showed positive correlations to the model in the visual cortex for unconscious contents. Three subjects showed positive correlations in the frontoparietal areas, and two subjects showed negative correlations to the model in the frontoparietal areas. The brain patterns correlated to the model were much stronger and more consistent across subjects in the conscious condition.

When we encoded the FCNNs' hidden representations from the conscious contents, these contextualized representations in the encoding-based RSA were also correlated to the brain representations of the unconscious contents in both the occipital and the frontal areas. The encoding-based RSA allowed us to investigate how the representations of the computational models are generalized in different conscious states. Taking together the encoding and the decoding results, we suggest that the representations of the conscious contents were enhanced patterns of similar representations associated with the unconscious contents, but additional work is needed to make further determinations.

## 5.2 Integration of different results

The goal of the current thesis was to test the scope of information processing and representation of items associated with null perceptual sensitivity in both brains and artificial neural networks. The extent to which information processing occurs for unconscious stimuli remains highly debated in consciousness science. Previous evidence of unconscious information processing in the brain and behavior suffers from confounds associated

with the use of subjective measures to assess awareness (i.e., criterion biases) and the lack of reliability and sensitivity of the measures used for assessing awareness (Newell & Shanks, 2014). We approached this question using a framework that aimed to carefully control for objective null perceptual sensitivity at the individual level, and exploit information-based approaches (Kriegeskorte et al., 2006) to characterize the brain representation of unconscious items (Soto et al., 2019) using a within-subject design, high-precision, highly-sampled neuroimaging approach. Here we demonstrated that even when subjects display null perceptual sensitivity, distributed information patterns related to the semantic category of unconscious items can be reliably decoded from BOLD multivoxel patterns. Evidence of unconscious information processing was largely distributed in the brain across the ventral visual pathway, involving prefrontal substrates in most subjects. Additionally, there were associations between the computer-vision model representations and the brain patterns in both the ventral visual pathway and the frontoparietal areas in the unseen trials, mainly using the standard representational similarity analysis pipeline. This result has implications for models of visual consciousness which suggest that unconscious information processing is local and restricted to the visual cortex (Dehaene et al., 2006), relative to conscious information processing which is widespread in large-scale frontoparietal systems. These data suggest this does not need to be the case.

Our results extend previous neuroimaging studies of unconscious visual processing. Studies have shown that the identity of unconsciously processed items such as grating patterns (King et al., 2016), locations of small line segments (Salti et al., 2015), location of letters (Trübutschek et al., 2017), and human faces (Axelrod et al., 2015) can be decoded from brain activity patterns. However, these studies used subjective measures of visual awareness to split unconscious and conscious trials, and these are known to be unreliable and subject to decision biases. Also, as noted in the introduction, brain responses to unconscious items associated with null perceptual sensitivity have been difficult to detect in a reliable and replicable manner (see Fang & He, 2005; Gayet et al., 2020; Hesselmann et al., 2011; Ludwig & Hesselmann, 2015; Ludwig et al., 2015) and when information could be decoded from multivoxel patterns this was restricted to the primary visual cortex (Stein

et al., 2021).

Intriguingly, our results indicate that the multivariate activity patterns associated with unconscious items associated with null perceptual sensitivity generalize to the representation of conscious items for which perceptual sensitivity was well above chance. This indicates that the brain representations of visual items may be invariant across states of conscious awareness. Schurger et al. (2010) showed that classifiers trained to discriminate the category of invisible items did not generalize to predict the category of conscious visible items. However, unlike the present well-powered within-subject study with over 1300 trials per subject ( $\sim 700$  for the unconscious condition), Schurger et al. (2010) only used a limited number of trials ( $\sim 300$ ) per subject which might have been insufficient for training a classifier that generalized across awareness states.

According to the Global Workspace Model of consciousness (Baars, 1997; Dehaene & Naccache, 2001), consciousness depends on "active neural firing" across large-scale frontoparietal networks (Dehaene & Changeux, 2011, 2004). while unconscious contents only activate primary sensory cortices. However, the GWT needs to be accommodated to explain two features of our results: (i) that the categories of the unseen images could be decoded from the multivoxel patterns of the prefrontal areas (i.e., middle frontal gyrus) and (ii) the brain patterns were similar between the unconscious and conscious conditions. Both the decoding and to some extent the encoding-based RSA results suggested that when the model was trained with the trials of the conscious condition, the model could be generalized from the conscious condition to the unconscious condition. There appear to be invariant properties of visual representations expressed in a spectrum involving unconscious and conscious states. This was not predicted by the Global Workspace model.

According to the local recurrent process model of consciousness, the rapidly presented and masked images trigger a "feedforward pass" through the brain from the occipital to higher-order areas of the ventral visual stream. However, this "feedforward pass" does not lead to conscious awareness (Lamme, 2000, 2020; Lamme et al., 1998). The fact that we observed that unconscious information can be decoded from the prefrontal activity is somehow in line with the recurrent theory, although the theory does not specifically

predict that unconscious contents may be decoded from the prefrontal cortex. Additionally, according to the high-order theory of consciousness (Gennaro, 1996; Rosenthal & Weisberg, 2008; Rosenthal, 2004), conscious awareness occurs when the higher-order brain areas (i.e., the prefrontal areas) re-represent the information projected from the lower-order brain areas (i.e., the visual cortex). Thus, when the re-representation fails, stimuli are unconscious. The decoding and the encoding-RSA results in the unconscious condition show that information about the image categories could be decoded in both the visual cortex and the prefrontal areas, but the encoded representations were not the same in the prefrontal areas (even if the patterns were decodable; see Discussion of Chapter 4). This suggests that the re-representation of the visual content by the prefrontal cortex may be operating at a non-conscious level and this would challenge the higher-order theory. However, recent evidence indicates that long-range feedback connections from the prefrontal cortex, rather than local feedback loops in the visual cortex, are more critical for visual consciousness (Huang et al., 2020). Based on this recent observation, our results may still be accommodated by the higher order theory, if we assume that there can be non-conscious decoding of visual content in the prefrontal in the absence of strong top-down feedback connections to the visual cortex (Rosenthal & Weisberg, 2008; Rosenthal, 2004).

Our simulations provided a computational approach to studying the processing of unseen contents. The information of the noisy image categories in the hidden layer could not be used to guide the network's last readout layer. The simulation results are similar to the observation that areas of the ventral visual pathway contain multivariate patterns of activity that allow for decoding of the image class even when observers display null perceptual sensitivity to the target category. Our results provide new insight into how images are represented in FCNNs and how information is used to make the classification decision under noisy conditions. Moreover, these results provide a formal demonstration that even under conditions associated with null perceptual sensitivity there remains meaningful information in a hidden state of the feedforward network. Recurrent processing could be the underlying mechanism that links processing in the visual cortex, temporal, and

parietal areas (Beck et al., 2001; Fahrenfort et al., 2007; Grill-Spector et al., 2000). However, when visual signals are embedded in noise (i.e., visually masked as in our current experiment) they are more likely to trigger feedforward information processing only without recurrent feedback (Fahrenfort et al., 2007). Accordingly, it has been proposed that visual representations may be preserved during feedforward processing but the absence of recurrent feedback leads to the absence of conscious experience (Bullier, 2001; Lamme & Roelfsema, 2000; Pascual-Leone & Walsh, 2001). The neural network models we used in the computer simulation of the visual task were all feedforward (DiCarlo et al., 2012; Yamins & DiCarlo, 2016), which may lack the capacity to preserve visual features across higher-order layers, so that any useful information might be left to local processes within each layer (Nayebi et al., 2018) and, therefore, the last readout layer may not fully exploit the information from previous layers to guide the perceptual decision.

Recent research has tried to overcome these issues and experimental data indicates that recurrent neural network (RNN) models provide better representations than FCNNs in object recognition at different levels of image noise (Spoerer et al., 2017; Zwickel et al., 2007). Instead of feedforward connections, RNN models propose feedback connections to incorporate different levels of information. Spoerer et al. (2017) and Kietzmann et al. (2019c) proposed RNN models with local and cross-layer feedback connections and they encode image features that better explain brain activity compared to FCNNs (Kietzmann et al., 2019c; Nayebi et al., 2018; Shi et al., 2018; Spoerer et al., 2017). It will be relevant for future studies to test whether adding recurrent processes to the FCNN model improves the read-out of the hidden layer representations for recognizing noisy images, hence increasing classification performance. For now, we conclude that feedforward processing can lead to meaningful hidden representational states that are ubiquitous in brains and artificial neural network models.

## 6 Resumen amplio en castellano

El modelo de la conciencia humana basado en un espacio neural de trabajo global (Dehaene, 2014) propone que la conciencia está asociada con la actividad sostenida en redes de asociación a gran escala que involucran la corteza frontoparietal, haciendo que la información sea accesible globalmente a los sistemas involucrados en la memoria de trabajo y el control del comportamiento (Dehaene, 2014). Se cree que el procesamiento visual inconsciente, por otro lado, es transitorio y opera localmente en sistemas perceptivos de bajo nivel (Lamme, 2020). Sin embargo, estudios recientes han confrontado este punto de vista con datos interesantes que sugieren que el procesamiento de información inconsciente está implicado en operaciones de orden superior asociadas con el control cognitivo (Van Gaal & Lamme, 2012), el comportamiento guiado por la memoria (Chong et al., 2014; Rosenthal et al., 2016; Soto et al., 2011; Trübtschek et al., 2017; Wuethrich et al., 2018), y el lenguaje (Hassin, 2013). Sin embargo, trabajos posteriores no apoyan este punto de vista (Rabagliati et al., 2018), e incluso la evidencia de procesamiento semántico inconsciente se ha cuestionado recientemente (Kouider & Dehaene, 2007; Stein et al., 2020). Quedan por determinar los límites y el alcance del procesamiento inconsciente.

Es probable que esta controversia se origine por la falta de un marco sólido para aislar el procesamiento de información inconsciente (Soto et al., 2019). Los estudios a menudo se basan sólo en medidas subjetivas para definir la ausencia o presencia de conciencia (Overgaard et al., 2010) y para identificar los marcadores neuronales del procesamiento inconsciente, pero estas medidas son sensibles a sesgos en el criterio de decisión para informar acerca de la presencia o ausencia de conciencia (Peters & Lau, 2015). Esto hace imposible determinar si la “invisibilidad subjetiva” está realmente asociada con el procesamiento inconsciente. La literatura de antecedentes relevante se resume en el Capítulo

1.

Un objetivo clave de esta tesis es desarrollar marcadores neurales que sean sensibles dentro de cada participante del alcance del procesamiento inconsciente. Aquí utilizamos un enfoque de análisis computacional dentro de cada participante, de alta precisión y altamente muestreado con muchos ensayos por participante, para definir las propiedades de las representaciones cerebrales asociadas con el contenido perceptivo inconsciente, definido por la ausencia de sensibilidad perceptiva nula. El segundo objetivo es proporcionar enfoques basados en técnicas de aprendizaje automático para la codificación y la decodificación de representaciones inconscientes en patrones de actividad cerebral. El tercer objetivo es proporcionar una simulación computacional de la representación de contenidos no vistos (i.e. inconscientes) utilizando modelos de redes neuronales profundas de última generación (Kriegeskorte, 2015; Lindsay, 2021). Este enfoque nos permitirá comprender mejor si las representaciones de los contenidos invisibles difieren de los contenidos visibles tanto en el cerebro humano como en los modelos de inteligencia artificial del cerebro humano.

El Capítulo 2 describe el experimento de resonancia magnética funcional de alta precisión tal y como se describe en el párrafo de arriba. Reclutamos a siete participantes para que ser escaneados con fMRI durante seis días. Las imágenes presentaban muy brevemente por unos milisegundos y se enmascararon antes y después. Se pidió a los sujetos (i) que identificaran y respondieran a la categoría de la imagen (es decir, animado versus inanimado) y (ii) que calificaran el estado de conciencia visual asociado con la imagen (es decir, inconsciente, breve destello y consciente). Luego llevamos a cabo un análisis de comportamiento fuera de línea para asegurarnos de que los sujetos tenían una sensibilidad perceptiva nula acerca de la categoría de las imágenes cuando informaban que eran "inconscientes". Y luego llevamos a cabo un análisis de patrones multivariado para decodificar las categorías de imágenes de las señales de resonancia en varias regiones de interés (ROI), es decir, giro lingual, corteza pericalcarina, corteza occipital lateral, giro fusiforme, giro parahipocampal, lóbulo temporal inferior, lóbulo parietal inferior, precúneo, giro parietal superior, giro frontal superior, giro frontal medio y giro frontal inferior. Esto se hizo para cada sujeto y cada uno de los estados de conciencia. Usamos un support



vector machine con regularización L1, anidado con la eliminación de vóxeles invariantes y escalado de características entre 0 y 1 como pasos de preprocesamiento. Los datos se dividieron en conjuntos de entrenamiento y prueba dejando un objeto animado y otro inanimado como conjunto de prueba. El rendimiento de la tubería instalada se estimó usando el área bajo la curva operativa del receptor (ROC AUC) en el conjunto de prueba. Esto se denominó generalización fuera de la muestra y se llevó a cabo dentro del estado inconsciente y consciente de cada sujeto y cada ROI.

Para los resultados de comportamiento, todos los sujetos mostraron una sensibilidad perceptiva superior al azar tanto cuando los participaban informaban de ver un destello o ser conscientes. Cuatro de los siete sujetos mostraron una sensibilidad perceptiva nula en aquellos ensayos en los que los sujetos informaron ser inconscientes de las categorías de las imágenes. Dos de los otros tres sujetos mostraron sensibilidad nula marginal. En los ensayos inconscientes, los patrones cerebrales en la circunvolución fusiforme y la circunvolución frontal media contenían categorías de imágenes que podían decodificarse en seis de los siete sujetos. Las categorías de las imágenes inconscientes también se decodificó a partir de la actividad en la circunvolución frontal inferior en cinco sujetos, y en cuatro sujetos del lóbulo parietal inferior, el lóbulo temporal inferior, la corteza occipital lateral, la circunvolución parahipocampal y la circunvolución parietal superior. Este resultado tiene implicaciones para los modelos de conciencia visual que sugieren que el procesamiento de información inconsciente es local y está restringido a la corteza visual (Dehaene et al., 2006). Estos datos sugieren que este no tiene por qué ser el caso.

También encontramos que los patrones en la circunvolución fusiforme, la corteza occipital lateral y el precúneo podían generalizarse de consciente a inconsciente en todos los sujetos. Además, los patrones de actividad en el lóbulo parietal inferior, el lóbulo temporal inferior, la circunvolución frontal media lingual y la circunvolución parietal superior podrían generalizarse de consciente a inconsciente en seis sujetos. Estos resultados fueron sólidos incluso con diferentes términos de regularización o procedimientos de validación cruzada. Esto indica que las representaciones cerebrales de elementos visuales pueden ser invariantes a través de los estados de conciencia.

El objetivo del Capítulo 3 es utilizar redes neuronales artificiales convolucionales feed-forward (FCNN) para proporcionar una simulación de los resultados obtenidos en el experimento fMRI. La hipótesis es que la representación de la capa oculta de la red neuronal puede contener información que permite la decodificación de un estímulo visual ruidoso incluso cuando las FCNN realizan la clasificación de imágenes al azar (sensibilidad nula en la discriminación).

La FCNN fue entrenada previamente y adaptado a nuestro experimento a través de procedimientos de aprendizaje por transferencia (Yosinski et al., 2014). Nuestras FCNN contenían sólo una capa oculta después de las capas convolucionales previamente entrenadas (es decir, VGGNet, Simonyan and Zisserman, 2015). Las FCNN primero aprendieron a realizar la misma tarea de discriminación visual que los observadores humanos con imágenes claras de elementos animados e inanimados. Luego, las FCNN realizaron la tarea con diferentes niveles de ruido en la imagen. Al igual que en el estudio de fMRI con los observadores humanos, usamos un máquina de vector de soporte lineal (SVM) para analizar la información contenida en la capa oculta de FCNN y ver si se podía decodificar la clase de imagen bajo diferentes niveles de ruido. Sorprendentemente, cuando las FCNN no podía clasificar las imágenes con ruido, observamos que la representación de la capa oculta de estos FCNN contenía información que permitía que un clasificador (support vector machine lineal) decodificara la categoría de imagen por encima del azar.

El Capítulo 4 utilizó enfoques computacionales basados en la información para estudiar los marcadores neuronales de los contenidos conscientes e inconscientes, usando los datos de fMRI presentados en el primer de los capítulos empíricos. Cuantificamos las respuestas a nivel de vóxel a las imágenes enmascaradas dentro de las regiones del cerebro analizadas y comparamos los patrones de información capturados por los modelos de visión por computadora (Schrimpf et al., 2020) y los patrones de información capturados por el fMRI. Por lo tanto, realizamos análisis de la similaridad de las representación (RSA) (Kriegeskorte et al., 2006), lo que nos permitió cuantificar la similaridad entre el modelo de computación visual y los patrones de fMRI.

Los resultados mostraron que los patrones de actividad cerebral en tres de los siete sujetos estaban correlacionados con el modelo computacional, incluidas las áreas prefrontal izquierda y las áreas occipitales (es decir, VGG19 y RestNet50), incluso cuando los sujetos no eran conscientes de las imágenes. Sin embargo, este patrón no fue sólido en los cinco modelos de FCNN. Cuando los sujetos eran conscientes de las imágenes, los patrones de actividad cerebral de seis de los siete sujetos se correlacionaron con el modelo computacional en la vía visual ventral y en las áreas frontoparietales, y el patrón fue consistente en los cinco modelos.

El Capítulo 5 resume y discute los hallazgos experimentales de la tesis. El objetivo de la tesis fue probar el alcance del procesamiento inconsciente y la representación de los contenidos asociados con una sensibilidad perceptual nula, tanto en el cerebro humano como en redes neuronales artificiales de visión computacional. La evidencia previa de procesamiento de información inconsciente en el cerebro y el comportamiento adolece de varios factores de confusión asociados con el uso de medidas subjetivas para evaluar la conciencia (es decir, sesgos en el criterio de decisión para informar de la ausencia de conciencia visual) y la falta de confiabilidad y sensibilidad de las medidas utilizadas para evaluar la conciencia (Newell & Shanks, 2014). Abordamos este tema utilizando un paradigma novedoso para controlar cuidadosamente la sensibilidad perceptiva nula objetiva a nivel individual y explotar los enfoques basados en la información (Kriegeskorte et al., 2006) y el aprendizaje automático para caracterizar la representación cerebral de los elementos inconscientes (Soto et al., 2019) utilizando un enfoque de neuroimagen altamente muestreado, de alta precisión dentro de cada sujeto.

Los resultados indican que los patrones de actividad multivariados asociados con elementos inconscientes asociados con una sensibilidad perceptiva nula se generalizan a la representación de elementos conscientes para los cuales la sensibilidad perceptiva estaba muy por encima del azar. Esto indica que las representaciones cerebrales de elementos visuales pueden ser invariantes a través de los estados de conciencia. Este resultado no se predice por el influyente modelo de la conciencia humana basado en un espacio neural de trabajo global (Dehaene, 2014). Los resultados también sugieren que la re-representación

de los contenidos codificados en la corteza visual por parte de la corteza prefrontal podría estar operando a un nivel inconsciente y esto desafiaría modelos de conciencia como la teoría de orden superior que proponen un papel fundamental de los procesos de representación en la corteza frontal para la experiencia consciente. Sin embargo, la evidencia reciente indica que las conexiones de retroalimentación (feedback) de largo alcance desde la corteza prefrontal a la corteza visual son más importantes para la conciencia visual que las conexiones de abajo-arriba o feedforward (Huang et al., 2020). En base a esto, nuestros resultados aún pueden ser explicados de acuerdo la teoría de orden superior de la conciencia (Rosenthal & Weisberg, 2008; Rosenthal, 2004).

Los resultados de nuestra simulación brindan una nueva perspectiva sobre cómo se representan las imágenes en las FCNN y cómo usas la información para tomar decisiones en condiciones ruidosas. Además, estos resultados proporcionan una demostración formal de que, incluso en condiciones asociadas con una sensibilidad perceptiva nula, sigue habiendo información significativa en un estado oculto de la red neural. El procesamiento recurrente (feedback o de arriba-abajo) podría ser el mecanismo subyacente que vincula el procesamiento en la corteza visual, las áreas temporal y parietal (Beck et al., 2001; Fahrenfort et al., 2007; Grill-Spector et al., 2000). Sin embargo, cuando las señales visuales están mezcladas con ruido (es decir, visualmente enmascaradas como en nuestros experimentos), es más probable que activen un procesamiento de información de abajo-arriba o feedforward, sin retroalimentación recurrente (Fahrenfort et al., 2007). En consecuencia, se ha propuesto que las representaciones visuales pueden conservarse durante el procesamiento feedforward pero la ausencia de retroalimentación recurrente conduce a la ausencia de experiencia consciente (Bullier, 2001; Lamme & Roelfsema, 2000; Pascual-Leone & Walsh, 2001). Los modelos de redes neuronales que usamos en la simulación por computadora de la tarea visual eran todos feedforward (DiCarlo et al., 2012; Yamins & DiCarlo, 2016), y por tanto, pueden carecer de la capacidad para preservar las características visuales en capas de orden superior, de modo que cualquier información útil puede quedarse en los procesos locales dentro de cada capa (Nayebi et al., 2018) y, por lo tanto, es posible que la última capa de la red neural no aproveche completamente la información de las capas

anteriores para guiar la decisión en la clasificación de la percepción.

Investigaciones recientes han tratado de solventar este asunto y los datos experimentales sugieren que los modelos computacionales basados en redes neuronales recurrentes (RNN) proporcionan mejores representaciones que las FCNNs para el reconocimiento de objetos a través de diferentes niveles de ruido en la imagen (Spoerer et al., 2017; Zwickel et al., 2007). En vez de conexiones puramente feedforward, los modelos RNN models poseen mecanismos de feedback (Kietzmann et al., 2019b; Nayebi et al., 2018; Shi et al., 2018; Spoerer et al., 2017). Será relevante para estudios futuros examinar si la presencia de proceso de feedback recurrente en las redes neurales artificiales mejora la lectura de la información en la capa oculta de la red para el reconocimiento de imágenes en presencia de ruido. Por ahora, concluimos que el procesamiento de abajo-arriba (feedforward) puede dar lugar a representaciones significativas de objetos enmascarados y no vistos que se encuentran en estados ocultos de la red neural, y que están presentes tanto en el cerebro humano como en los modelos artificiales de visión computacional.

# Bibliography

- Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al. (2015). Tensorflow: Large-scale machine learning on heterogeneous systems.
- Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., & Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, 14.
- Altmann, A., Tološi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics*, 26(10), 1340–1347.
- Axelrod, V., Bar, M., & Rees, G. (2015). Exploring the unconscious using faces. *Trends in Cognitive Sciences*, 19(1), 35–45.
- Baars, B. J. (1993). *A Cognitive Theory of Consciousness*. Cambridge University Press.
- Baars, B. J. (1997). In the theatre of consciousness. Global workspace theory, a rigorous scientific theory of consciousness. *Journal of Consciousness Studies*, 4(4), 292–309.
- Baars, B. J., Geld, N., & Kozma, R. (2021). Global workspace theory (GWT) and prefrontal cortex: Recent developments. *Frontiers in Psychology*, 5163.
- Bachmann, T., & Francis, G. (2013). *Visual Masking: Studying perception, Attention, and Consciousness*. Academic Press.
- Beck, D. M., Rees, G., Frith, C. D., & Lavie, N. (2001). Neural correlates of change detection and change blindness. *Nature Neuroscience*, 4(6), 645–650.
- Bhavsar, H., & Panchal, M. H. (2012). A review on support vector machine for data classification. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, 1(10), 185–189.

- Breitmeyer, B., Ogmen, H., Ögmen, H., et al. (2006). *Visual Masking: Time Slices Through Conscious and Unconscious Vision*. Oxford University Press.
- Breitmeyer, B., Ogmen, H., Ramon, J., & Chen, J. (2005). Unconscious and conscious priming by forms and their parts. *Visual Cognition*, *12*(5), 720–736.
- Breitmeyer, B. G. (2007). Visual masking: Past accomplishments, present status, future developments. *Advances in Cognitive Psychology*, *3*(1-2), 9.
- Breitmeyer, B. G., Ogmen, H., & Chen, J. (2004). Unconscious priming by color and form: Different processes and levels. *Consciousness and Cognition*, *13*(1), 138–157.
- Brown, R., Lau, H., & LeDoux, J. E. (2019). Understanding the higher-order approach to consciousness. *Trends in Cognitive Sciences*, *23*(9), 754–768.
- Bruha, I. (2000). From machine learning to knowledge discovery: Survey of preprocessing and postprocessing. *Intelligent Data Analysis*, *4*(3-4), 363–374.
- Bullier, J. (2001). Feedback connections and conscious vision. *Trends in Cognitive Sciences*, *5*(9), 369–370.
- Cheesman, J., & Merikle, P. M. (1986). Distinguishing conscious from unconscious perceptual processes. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *40*(4), 343.
- Cheesman, J., & Merikle, P. M. (1984). Priming with and without awareness. *Perception & Psychophysics*, *36*(4), 387–395.
- Chollet, F. et al. (2018). Keras: The python deep learning library. *The Astrophysics Source Code Library (ASCL)*, ascl-1806.
- Chong, T. T.-J., Husain, M., & Rosenthal, C. R. (2014). Recognizing the unconscious. *Current Biology*, *24*(21), 1033–1035.
- Clevert, D.-A., Unterthiner, T., & Hochreiter, S. (2015). Fast and accurate deep network learning by exponential linear units (ELUs). *arXiv preprint arXiv:1511.07289*.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.
- Dehaene, S. (2014). *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts*. Penguin.

- Dehaene, S., & Changeux, J.-P. (1989). A simple model of prefrontal cortex function in delayed-response tasks. *Journal of Cognitive Neuroscience*, *1*(3), 244–261.
- Dehaene, S., & Changeux, J.-P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, *70*(2), 200–227.
- Dehaene, S., & Changeux, J.-P. (2004). Neural mechanisms for access to consciousness. *The Cognitive Neurosciences*, *3*, 1145–58.
- Dehaene, S., & Changeux, J.-P. (2005). Ongoing spontaneous activity controls access to consciousness: A neuronal model for inattention blindness. *PLoS biology*, *3*(5), e141.
- Dehaene, S., & Changeux, J.-P. (2000). Reward-dependent learning in neuronal networks for planning and decision making. *Progress in Brain Research*, *126*, 217–229.
- Dehaene, S., Changeux, J.-P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: A testable taxonomy. *Trends in Cognitive Sciences*, *10*(5), 204–211.
- Dehaene, S., Charles, L., King, J.-R., & Marti, S. (2014). Toward a computational theory of conscious processing. *Current Opinion in Neurobiology*, *25*, 76–84.
- Dehaene, S., Jobert, A., Naccache, L., Ciuciu, P., Poline, J.-B., Le Bihan, D., & Cohen, L. (2004). Letter binding and invariant recognition of masked words: Behavioral and neuroimaging evidence. *Psychological Science*, *15*(5), 307–313.
- Dehaene, S., Kerszberg, M., & Changeux, J.-P. (1998a). A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the National Academy of Sciences*, *95*(24), 14529–14534.
- Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition*, *79*(1-2), 1–37.
- Dehaene, S., Naccache, L., Cohen, L., Le Bihan, D., Mangin, J.-F., Poline, J.-B., & Rivière, D. (2001). Cerebral mechanisms of word masking and unconscious repetition priming. *Nature Neuroscience*, *4*(7), 752–758.



- Dehaene, S., Naccache, L., Le Clec'H, G., Koechlin, E., Mueller, M., Dehaene-Lambertz, G., van de Moortele, P.-F., & Le Bihan, D. (1998b). Imaging unconscious semantic priming. *Nature*, *395*(6702), 597–600.
- Dehaene, S., Sergent, C., & Changeux, J.-P. (2003). A neuronal network model linking subjective reports and objective physiological data during conscious perception. *Proceedings of the National Academy of Sciences*, *100*(14), 8520–8525.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2009, 248–255.
- Deutsch, R., Gawronski, B., & Strack, F. (2006). At the boundaries of automaticity: Negation as reflective operation. *Journal of Personality and Social Psychology*, *91*(3), 385.
- Diaz, M. T., & McCarthy, G. (2007). Unconscious word processing engages a distributed network of brain regions. *Journal of Cognitive Neuroscience*, *19*(11), 1768–1775.
- DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434.
- Diedrichsen, J., & Kriegeskorte, N. (2017). Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLoS Computational Biology*, *13*(4), e1005508.
- Dutta, A., Shah, K., Silvanto, J., & Soto, D. (2014). Neural basis of non-conscious visual working memory. *Neuroimage*, *91*, 336–343.
- Ekstrom, A. (2010). How and when the fmri bold signal relates to underlying neural activity: The danger in dissociation. *Brain Research Reviews*, *62*(2), 233–244.
- Evett, L. J., & Humphreys, G. W. (1981). The use of abstract graphemic information in lexical access. *The Quarterly Journal of Experimental Psychology*, *33*(4), 325–350.
- Fahrenfort, J. J., Scholte, H. S., & Lamme, V. A. (2007). Masking disrupts reentrant processing in human visual cortex. *Journal of Cognitive Neuroscience*, *19*(9), 1488–1497.

- Fang, F., & He, S. (2005). Cortical responses to invisible objects in the human dorsal and ventral pathways. *Nature Neuroscience*, *8*(10), 1380–1385.
- Fei-Fei, L., Fergus, R., & Perona, P. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. In: *2004 Conference on Computer Vision and Pattern Recognition Workshop*. IEEE. 2004, 178–178.
- Fisch, L., Privman, E., Ramot, M., Harel, M., Nir, Y., Kipervasser, S., Andelman, F., Neufeld, M. Y., Kramer, U., Fried, I., et al. (2009). Neural “ignition”: Enhanced activation linked to perceptual awareness in human ventral stream visual cortex. *Neuron*, *64*(4), 562–574.
- Fisher, A., Rudin, C., & Dominici, F. (2018). All models are wrong but many are useful: Variable importance for black-box, proprietary, or misspecified prediction models, using model class reliance. *arXiv preprint arXiv:1801.01489*.
- Friston, K. J., Ashburner, J., Frith, C. D., Poline, J.-B., Heather, J. D., & Frackowiak, R. S. (1995). Spatial registration and normalization of images. *Human Brain Mapping*, *3*(3), 165–189.
- Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., & Frackowiak, R. S. (1994). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, *2*(4), 189–210.
- Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(4), 193–202.
- Gayet, S., Guggenmos, M., Christophel, T. B., Haynes, J.-D., Paffen, C. L., Sterzer, P., & Van der Stigchel, S. (2020). No evidence for mnemonic modulation of interocularly suppressed visual input. *NeuroImage*, 116801.
- Gennaro, R. J. (1996). *Consciousness and Self-consciousness: A Defense of the Higher-order Thought Theory of Consciousness* (Vol. 6). John Benjamins Publishing.

- Goebel, R., Muckli, L., Zanella, F. E., Singer, W., & Stoerig, P. (2001). Sustained extrastriate cortical activation without visual awareness revealed by fmri studies of hemianopic patients. *Vision Research*, *41*(10-11), 1459–1474.
- Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: A flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in Neuroinformatics*, *5*, 13.
- Greenwald, A. G., Draine, S. C., & Abrams, R. L. (1996). Three cognitive markers of unconscious semantic activation. *Science*, *273*(5282), 1699–1702.
- Grill-Spector, K., Kushnir, T., Hendler, T., & Malach, R. (2000). The dynamics of object-selective activation correlate with recognition performance in humans. *Nature Neuroscience*, *3*(8), 837–843.
- Güçlü, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014.
- Hämäläinen, M., Hari, R., Ilmoniemi, R. J., Knuutila, J., & Lounasmaa, O. V. (1993). Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of Modern Physics*, *65*(2), 413.
- Hassin, R. R. (2013). Yes it can: On the functional abilities of the human unconscious. *Perspectives on Psychological Science*, *8*(2), 195–207.
- Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*(5539), 2425–2430.
- Haynes, J.-D. (2009). Decoding visual consciousness from human brain signals. *Trends in Cognitive Sciences*, *13*(5), 194–202.
- Haynes, J.-D., Driver, J., & Rees, G. (2005). Visibility reflects dynamic changes of effective connectivity between v1 and fusiform cortex. *Neuron*, *46*(5), 811–821.
- He, K., Zhang, X., Ren, S., & Sun, J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, 770–778.

- He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(9), 1904–1916.
- Henry, J. C. (2006). Electroencephalography: Basic principles, clinical applications, and related fields. *Neurology*, *67*(11), 2092–2092.
- Hesselmann, G., Hebart, M., & Malach, R. (2011). Differential bold activity associated with subjective and objective reports during “blindsight” in normal observers. *Journal of Neuroscience*, *31*(36), 12936–12944.
- Hilt, D. E., & Seegrift, D. W. (1977). *Ridge, A Computer Program for Calculating Ridge Regression Estimates* (Vol. 236). Department of Agriculture, Forest Service, Northeastern Forest Experiment Station, 1977.
- Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- Horowitz, J. L. The bootstrap. In: *Handbook of econometrics*. Vol. 5. Elsevier, 2001, pp. 3159–3228.
- Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.
- Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. Densely connected convolutional networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, 2017, 4700–4708.
- Huang, L., Wang, L., Shen, W., Li, M., Wang, S., Wang, X., Ungerleider, L. G., & Zhang, X. (2020). A source for awareness-dependent figure–ground segregation in human prefrontal cortex. *Proceedings of the National Academy of Sciences*, *117*(48), 30836–30847.
- Humphreys, G. W., Besner, D., & Quinlan, P. T. (1988). Event perception and the word repetition effect. *Journal of Experimental Psychology: General*, *117*(1), 51.
- Humphreys, G. W., Evett, L. J., & Taylor, D. E. (1982). Automatic phonological priming in visual word recognition. *Memory & Cognition*, *10*(6), 576–590.

- Jenkinson, M., Bannister, P., Brady, M., & Smith, S. M. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, *17*(2), 825–841.
- Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., & Smith, S. M. (2012). FSL. *Neuroimage*, *62*(2), 782–790.
- Jiang, Y., Zhou, K., & He, S. (2007). Human visual cortex responds to invisible chromatic flicker. *Nature Neuroscience*, *10*(5), 657–662.
- Jin, J., Dundar, A., & Culurciello, E. (2015). Robust convolutional neural networks under adversarial noise. *arXiv preprint arXiv:1511.06306*.
- Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260.
- Kanai, R., Walsh, V., & Tseng, C.-h. (2010). Subjective discriminability of invisibility: A framework for distinguishing perceptual and attentional failures of awareness. *Consciousness and Cognition*, *19*(4), 1045–1057.
- Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Computational Biology*, *10*(11).
- Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2019a). Deep neural networks in computational neuroscience. *Oxford Research Encyclopaedia of Neuroscience*. <https://doi.org/10.1093/acrefore/9780190264086.013.46>
- Kietzmann, T. C., McClure, P., & Kriegeskorte, N. Deep neural networks in computational neuroscience. In: *Oxford research encyclopedia of neuroscience*. 2019.
- Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019c). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, *116*(43), 21854–21863.
- Kim, C.-Y., & Blake, R. (2005). Psychophysical magic: Rendering the visible ‘invisible’. *Trends in Cognitive Sciences*, *9*(8), 381–388.

- King, J.-R., & Dehaene, S. (2014). Characterizing the dynamics of mental representations: The temporal generalization method. *Trends in Cognitive Sciences*, *18*(4), 203–210.
- King, J.-R., Pescetelli, N., & Dehaene, S. (2016). Brain mechanisms underlying the brief maintenance of seen and unseen sensory information. *Neuron*, *92*(5), 1122–1134.
- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Klambauer, G., Unterthiner, T., Mayr, A., & Hochreiter, S. Self-normalizing neural networks. In: *Advances in Neural Information Processing Systems*. 2017, 971–980.
- Koechlin, E., Naccache, L., Block, E., & Dehaene, S. (1999). Primed numbers: Exploring the modularity of numerical representations with masked and unmasked semantic priming. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(6), 1882.
- Konkle, T., & Alvarez, G. A. (2022). A self-supervised domain-general learning framework for human ventral stream representation. *Nature Communications*, *13*(1), 1–12.
- Kouider, S., & Dehaene, S. (2007). Levels of processing during non-conscious perception: A critical review of visual masking. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 857–875.
- Kranczioch, C., Debener, S., Schwarzbach, J., Goebel, R., & Engel, A. K. (2005). Neural correlates of conscious perception in the attentional blink. *Neuroimage*, *24*(3), 704–714.
- Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*, 417–446.
- Kriegeskorte, N. (2011). Pattern-information analysis: From stimulus decoding to computational-model testing. *Neuroimage*, *56*(2), 411–421.
- Kriegeskorte, N. (2009). Relating population-code representations between man, monkey, and computational models. *Frontiers in Neuroscience*, *3*, 35.
- Kriegeskorte, N., & Diedrichsen, J. (2019). Peeling the onion of brain representations. *Annual Review of Neuroscience*, *42*, 407–432.

- Kriegeskorte, N., & Douglas, P. K. (2018). Cognitive computational neuroscience. *Nature Neuroscience*, *21*(9), 1148–1160.
- Kriegeskorte, N., & Douglas, P. K. (2019). Interpreting encoding and decoding models. *Current Opinion in Neurobiology*, *55*, 167–179.
- Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, *103*(10), 3863–3868.
- Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008a). Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, *4*.
- Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., & Bandettini, P. A. (2008b). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*(6), 1126–1141.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems*. 2012, 1097–1105.
- Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R., Hong, H., Majaj, N., Issa, E., Bashivan, P., Prescott-Roy, J., Schmidt, K., et al. Brain-like object recognition with high-performing shallow recurrent nets. In: *Advances in Neural Information Processing Systems*. 2019, 12805–12816.
- Lamme, V. A. (2010). How neuroscience will change our view on consciousness. *Cognitive Neuroscience*, *1*(3), 204–220.
- Lamme, V. A. (2000). Neural mechanisms of visual awareness: A linking proposition. *Brain and Mind*, *1*(3), 385–406.
- Lamme, V. A. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences*, *10*(11), 494–501.
- Lamme, V. A. (2020). Visual functions generating conscious seeing. *Frontiers in Psychology*, *11*, 83.
- Lamme, V. A., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feed-forward and recurrent processing. *Trends in Neurosciences*, *23*(11), 571–579.

- Lamme, V. A., Super, H., Landman, R., Roelfsema, P. R., & Spekreijse, H. (2000). The role of primary visual cortex (v1) in visual awareness. *Vision Research*, *40*(10-12), 1507–1521.
- Lamme, V. A., Super, H., & Spekreijse, H. (1998). Feedforward, horizontal, and feedback processing in the visual cortex. *Current Opinion in Neurobiology*, *8*(4), 529–535.
- Lamme, V. A., Van Dijk, B. W., & Spekreijse, H. (1993). Contour from motion processing occurs in primary visual cortex. *Nature*, *363*(6429), 541–543.
- Lamme, V. A., Zipser, K., & Spekreijse, H. (2002). Masking interrupts figure-ground signals in v1. *Journal of Cognitive Neuroscience*, *14*(7), 1044–1053.
- Lau, H. (2008). Are we studying consciousness yet. *Frontiers of Consciousness: Chichele Lectures, 2008*, 245.
- Lau, H. C., & Passingham, R. E. (2006). Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proceedings of the National Academy of Sciences*, *103*(49), 18763–18768.
- LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. *The Handbook of Brain Theory and Neural Networks*, *3361*(10), 1995.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.
- Leshno, M., Lin, V. Y., Pinkus, A., & Schocken, S. (1993). Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, *6*(6), 861–867.
- Lewis-Peacock, J. A., & Norman, K. A. (2014). Multi-voxel pattern analysis of fMRI data. *The Cognitive Neuroscience*, *512*, 911–920.
- Li, M., Zhao, F., Lee, J., Wang, D., Kuang, H., & Tsien, J. Z. (2015). Computational classification approach to profile neuron subtypes from brain activity mapping data. *Scientific Reports*, *5*(1), 1–14.
- Lindsay, G. W. (2021). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of Cognitive Neuroscience*, *33*(10), 2017–2031.
- Lindsay, G. W., Rubin, D. B., & Miller, K. D. (2019). A simple circuit model of visual cortex explains neural and behavioral aspects of attention. *BioRxiv*, 875534.



- Ludwig, K., & Hesselmann, G. (2015). Weighing the evidence for a dorsal processing bias under continuous flash suppression. *Consciousness and Cognition*, *35*, 251–259.
- Ludwig, K., Kathmann, N., Sterzer, P., & Hesselmann, G. (2015). Investigating category- and shape-selective neural processing in ventral and dorsal visual stream under interocular suppression. *Human Brain Mapping*, *36*(1), 137–149.
- Macmillan, N. A. (1986). The psychophysics of subliminal perception. *Behavioral and Brain Sciences*, *9*(1), 38–39.
- Macmillan, N. A., & Creelman, C. D. (2004). *Detection Theory: A User's Guide*. Psychology press.
- Marcel, A. J. (1980). Conscious and preconscious recognition of polysemous words: Locating the selective effects of prior verbal context. *Attention and Performance VIII*, 435–457.
- Marcel, A. J. (1983). Conscious and unconscious perception: Experiments on visual masking and word recognition. *Cognitive Psychology*, *15*(2), 197–237.
- Marcel, T., Katz, L., & Smith, M. (1974). Laterality and reading proficiency. *Neuropsychologia*, *12*(1), 131–139.
- Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of eeg-and meg-data. *Journal of Neuroscience Methods*, *164*(1), 177–190.
- Marr, D., & Vision, A. (1982). A computational investigation into the human representation and processing of visual information. *WH San Francisco: Freeman and Company, San Francisco*.
- McFee, B., Salamon, J., & Bello, J. P. (2018). Adaptive pooling operators for weakly labeled sound event detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(11), 2180–2193.
- Mei, N., Rankine, S., Olafsson, E., & Soto, D. (2020). Similar history biases for distinct prospective decisions of self-performance. *Scientific Reports*, *10*(1), 1–13.
- Mei, N., Santana, R., & Soto, D. (2022). Informative neural representations of unseen contents during higher-order processing in human brains and deep artificial networks. *Nature Human Behaviour*, *6*(5), 720–731.

- Michel, M. (2019). Consciousness science underdetermined: A short history of endless debates. *Ergo, an Open Access Journal of Philosophy*, 6(28).
- Miconi, T. (2017). Biologically plausible learning in recurrent neural networks reproduces neural dynamics observed during cognitive tasks. *Elife*, 6, e20899.
- Moreno-Martínez, F. J., & Montoro, P. R. (2012). An ecological alternative to snodgrass & vanderwart: 360 high quality colour images with norms for seven psycholinguistic variables. *PloS One*, 7(5), e37527.
- Naccache, L., Blandin, E., & Dehaene, S. (2002). Unconscious masked priming depends on temporal attention. *Psychological Science*, 13(5), 416–424.
- Naccache, L., & Dehaene, S. (2001). The priming method: Imaging unconscious repetition priming reveals an abstract representation of number in the parietal lobes. *Cerebral Cortex*, 11(10), 966–974.
- Nagel, T. (1974). What is it like to be a bat. *Readings in Philosophy of Psychology*, 1, 159–168.
- Nair, V., & Hinton, G. E. Rectified linear units improve restricted Boltzmann machines. In: *International Conference on Machine Learning (ICML)*. 2010.
- Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *Neuroimage*, 56(2), 400–410.
- Nayebi, A., Bear, D., Kubilius, J., Kar, K., Ganguli, S., Sussillo, D., DiCarlo, J. J., & Yamins, D. L. Task-driven convolutional recurrent models of the visual system. In: *Advances in Neural Information Processing Systems*. 2018, 5290–5301.
- Newell, B. R., & Shanks, D. R. (2014). Unconscious influences on decision making: A critical review. *Behavioral and Brain Sciences*, 37(1), 1–19.
- Nguyen, Q., Valizadegan, H., & Hauskrecht, M. (2014). Learning classification models with soft-label information. *Journal of the American Medical Informatics Association*, 21(3), 501–508.
- Nonaka, S., Majima, K., Aoki, S. C., & Kamitani, Y. (2021). Brain hierarchy score: Which deep neural networks are hierarchically brain-like? *IScience*, 24(9), 103013.

- Ojala, M., & Garriga, G. C. (2010). Permutation tests for studying classifier performance. *Journal of Machine Learning Research*, *11*(Jun), 1833–1863.
- Overgaard, M., Timmermans, B., Sandberg, K., & Cleeremans, A. (2010). Optimizing subjective measures of consciousness. *Consciousness and Cognition*, *19*(2), 682–684.
- Pascual-Leone, A., & Walsh, V. (2001). Fast backprojections from the motion to the primary visual area necessary for visual awareness. *Science*, *292*(5516), 510–512.
- Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. (2017). Automatic differentiation in pytorch.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, *32*.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, *12*, 2825–2830.
- Peirce, J. W. (2007). Psychopy—psychophysics software in python. *Journal of Neuroscience Methods*, *162*(1-2), 8–13.
- Pereira, F., & Botvinick, M. (2011). Information mapping with pattern classifiers: A comparative study. *Neuroimage*, *56*(2), 476–496.
- Persaud, N., Davidson, M., Maniscalco, B., Mobbs, D., Passingham, R. E., Cowey, A., & Lau, H. (2011). Awareness-related activity in prefrontal and parietal cortices in blindsight reflects more than superior visual performance. *Neuroimage*, *58*(2), 605–611.
- Pessoa, L., & Ungerleider, L. G. (2004). Neural correlates of change detection and change blindness in a working memory task. *Cerebral Cortex*, *14*(5), 511–520.
- Peters, M. A., & Lau, H. (2015). Human observers have optimal introspective access to perceptual processes even for visually masked stimuli. *Elife*, *4*, e09651.

- Popham, S. F., Huth, A. G., Bilenko, N. Y., Deniz, F., Gao, J. S., Nunez-Elizalde, A. O., & Gallant, J. L. (2021). Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nature Neuroscience*, *24*(11), 1628–1636.
- Pournaghdali, A., & Schwartz, B. L. (2020). Continuous flash suppression: Known and unknowns. *Psychonomic Bulletin & Review*, *27*(6), 1071–1103.
- Pruim, R. H., Mennes, M., van Rooij, D., Llera, A., Buitelaar, J. K., & Beckmann, C. F. (2015). ICA-AROMA: A robust ICA-based strategy for removing motion artifacts from fMRI data. *Neuroimage*, *112*, 267–277.
- Qin, Z., Yu, F., Liu, C., & Chen, X. (2018). How convolutional neural network see the world—a survey of convolutional neural network visualization methods. *arXiv preprint arXiv:1804.11191*.
- Rabagliati, H., Robertson, A., & Carmel, D. (2018). The importance of awareness for understanding language. *Journal of Experimental Psychology: General*, *147*(2), 190.
- Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., et al. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, *22*(11), 1761–1770.
- Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025.
- Rosenblatt, F. (1960). Perceptron simulation experiments. *Proceedings of the Institute of Radio Engineers (IRE)*, *48*(3), 301–309.
- Rosenthal, C. R., Andrews, S. K., Antoniadis, C. A., Kennard, C., & Soto, D. (2016). Learning and recognition of a non-conscious sequence of events in human primary visual cortex. *Current Biology*, *26*(6), 834–841.
- Rosenthal, D., & Weisberg, J. (2008). Higher-order theories of consciousness. *Scholarpedia*, *3*(5), 4407.
- Rosenthal, D. M. (1993). Higher-order thoughts and the appendage theory of consciousness. *Philosophical Psychology*, *6*(2), 155–166.
- Rosenthal, D. M. (2004). Varieties of higher-order theory. *Advances in Consciousness Research*, *56*, 17–44.

- Rossion, B., De Gelder, B., Pourtois, G., Guérit, J.-M., & Weiskrantz, L. (2000). Early extrastriate activity without primary visual cortex in humans. *Neuroscience Letters*, *279*(1), 25–28.
- Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.
- Salti, M., Monto, S., Charles, L., King, J.-R., Parkkonen, L., & Dehaene, S. (2015). Distinct cortical codes and temporal dynamics for conscious and unconscious percepts. *Elife*, *4*, e05652.
- Schäfer, A. M., & Zimmermann, H. G. Recurrent neural networks are universal approximators. In: *International Conference on Artificial Neural Networks*. Springer. 2006, 632–640.
- Scherer, D., Müller, A., & Behnke, S. Evaluation of pooling operations in convolutional architectures for object recognition. In: *International Conference on Artificial Neural Networks*. Springer. 2010, 92–101.
- Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., et al. (2020). Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, 407007.
- Schurger, A., Pereira, F., Treisman, A., & Cohen, J. D. (2010). Reproducibility distinguishes conscious from nonconscious neural representations. *Science*, *327*(5961), 97–99.
- Seeliger, K., Fritsche, M., Güçlü, U., Schoenmakers, S., Schoffelen, J.-M., Bosch, S., & Van Gerven, M. (2018). Convolutional neural network-based encoding and decoding of visual object recognition in space and time. *NeuroImage*, *180*, 253–266.
- Sergent, C., Baillet, S., & Dehaene, S. (2005). Timing of the brain events underlying access to consciousness during the attentional blink. *Nature Neuroscience*, *8*(10), 1391–1400.
- Sergent, C., & Dehaene, S. (2004a). Is consciousness a gradual phenomenon? evidence for an all-or-none bifurcation during the attentional blink. *Psychological Science*, *15*(11), 720–728.

- Sergent, C., & Dehaene, S. (2004b). Neural processes underlying conscious perception: Experimental findings and a global neuronal workspace framework. *Journal of Physiology-Paris*, *98*(4-6), 374–384.
- Sheikh, U. A., Carreiras, M., & Soto, D. (2019). Decoding the meaning of unconsciously processed words using fMRI-based MVPA. *NeuroImage*, *191*, 430–440.
- Shi, J., Wen, H., Zhang, Y., Han, K., & Liu, Z. (2018). Deep recurrent neural network reveals a hierarchy of process memory during dynamic natural vision. *Human Brain Mapping*, *39*(5), 2269–2282.
- Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Smith, S. (2002). Fast robust automated brain extraction. *Human Brain Mapping*, *17*(3), 143–155.
- Sonoda, S., & Murata, N. (2017). Neural network with unbounded activation functions is universal approximator. *Applied and Computational Harmonic Analysis*, *43*(2), 233–268.
- Soto, D., Mäntylä, T., & Silvanto, J. (2011). Working memory without consciousness. *Current Biology*, *21*(22), 912–913.
- Soto, D., Sheikh, U. A., & Rosenthal, C. R. (2019). A novel framework for unconscious processing. *Trends in Cognitive Sciences*, *23*(5), 372–376.
- Specht, D. F. (1990). Probabilistic neural networks and the polynomial adaline as complementary techniques for classification. *IEEE Transactions on Neural Networks*, *1*(1), 111–121.
- Spoerer, C. J., McClure, P., & Kriegeskorte, N. (2017). Recurrent convolutional neural networks: A better model of biological object recognition. *Frontiers in Psychology*, *8*, 1551.
- Stein, T., Kaiser, D., Fahrenfort, J. J., & Van Gaal, S. (2021). The human visual system differentially represents subjectively and objectively invisible stimuli. *PLoS Biology*, *19*(5), e3001241.

- Stein, T., Utz, V., & Van Opstal, F. (2020). Unconscious semantic priming from pictures under backward masking and continuous flash suppression. *Consciousness and Cognition*, *78*, 102864.
- Sterzer, P., Haynes, J.-D., & Rees, G. (2008). Fine-scale activity patterns in high-level visual areas encode the category of invisible objects. *Journal of Vision*, *8*(15), 10–10.
- Suzuki, H., & Fukuda, H. (2013). Unconscious nature of insight problem solving: An analysis using subliminal priming by continuous flash suppression. *Cognitive Studies*, *20*, 353–367.
- Taherkhani, A., Belatreche, A., Li, Y., Cosma, G., Maguire, L. P., & McGinnity, T. M. (2020). A review of learning in biologically plausible spiking neural networks. *Neural Networks*, *122*, 253–272.
- Trübutschek, D., Marti, S., Ojeda, A., King, J.-R., Mi, Y., Tsodyks, M., & Dehaene, S. (2017). A theory of working memory without consciousness or sustained activity. *Elife*, *6*, e23871.
- Tsuchiya, N., & Koch, C. (2004). Continuous flash suppression. *Journal of Vision*, *4*(8), 61–61.
- Tsuchiya, N., & Koch, C. (2005). Continuous flash suppression reduces negative afterimages. *Nature Neuroscience*, *8*(8), 1096–1101.
- Van Gaal, S., & Lamme, V. A. (2012). Unconscious high-level information processing: Implication for neurobiological theories of consciousness. *The Neuroscientist*, *18*(3), 287–301.
- Van Gaal, S., Naccache, L., Meuwese, J. D., Van Loon, A. M., Leighton, A. H., Cohen, L., & Dehaene, S. (2014). Can the meaning of multiple words be integrated unconsciously? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1641), 20130212.
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R.,

- Larson, E., . . . SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. <https://doi.org/10.1038/s41592-019-0686-2>
- Weiskrantz, L. (2009). *Blindsight: A Case Study Spanning 35 Years and New Developments*. Oxford University Press.
- Wickens, D. D. (1973). Some characteristics of word encoding. *Memory & Cognition*, *1*(4), 485–490.
- Wright, R. E. (1995). Logistic regression. *Reading and Understanding Multivariate Statistics*, 217—244.
- Wuethrich, S., Hannula, D. E., Mast, F. W., & Henke, K. (2018). Subliminal encoding and flexible retrieval of objects in scenes. *Hippocampus*, *28*(9), 633–643.
- Yamins, D. L., Hong, H., Cadieu, C., & DiCarlo, J. J. Hierarchical modular optimization of convolutional networks achieves representations similar to macaque it and human ventral stream. In: *Advances in Neural Information Processing Systems*. 2013, 3093–3101.
- Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356.
- Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. How transferable are features in deep neural networks? In: *Advances in Neural Information Processing Systems*. 2014, 3320–3328.
- Zhang, J., & Mueller, S. T. (2005). A note on roc analysis and non-parametric estimate of sensitivity. *Psychometrika*, *70*(1), 203–212.
- Zwicker, T., Wachtler, T., & Eckhorn, R. (2007). Coding the presence of visual objects in a recurrent neural network of visual cortex. *Biosystems*, *89*(1-3), 216–226.