Deciphering the Brain's Visual Language : Natural Image Reconstruction using Deep Generative Models from fMRI Signals

Furkan Ozcelik

RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

Université de Toulouse

Doctorat de l'Université de Toulouse

préparé à l'Université Toulouse III - Paul Sabatier

Déchiffrer le langage visuel du cerveau : Reconstruction d'images naturelles à l'aide de modèles génératifs profonds à partir de signaux IRMf

---

Thèse présentée et soutenue, le 30 avril 2024 par

Furkan OZCELIK

École doctorale
EDMITT - Ecole Doctorale Mathématiques, Informatique et Télécommunications de Toulouse

Spécialité
Informatique et Télécommunications

Unité de recherche
CERCO - Centre de Recherche Cerveau et Cognition

Thèse dirigée par
Rufin VANRULLEN

Composition du jury
Mme Gemma ROIG, Présidente, Goethe-Universität
M. Huseyin BOYACI, Rapporteur, Bilkent University
M. Lars MUCKLI, Examinateur, University of Glasgow
M. Rufin VANRULLEN, Directeur de thèse, CNRS Occitanie Ouest

Deciphering the Brain's Visual Language: Natural Image Reconstruction using Deep Generative Models from fMRI Signals

A DISSERTATION PRESENTED
BY
FURKAN OZCELIK
TO
CERCO, CNRS UMR 5549

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF COMPUTER SCIENCE

UNIVERSITÉ DE TOULOUSE, PAUL SABATIER
TOULOUSE, FRANCE

APRIL 2024

Thesis advisor: Rufin VanRullen Furkan Ozcelik

Deciphering the Brain's Visual Language: Natural Image Reconstruction using Deep Generative Models from fMRI Signals

Abstract

The great minds of humanity were always curious about the nature of mind, brain, and consciousness. Through physical and thought experiments, they tried to tackle challenging questions about visual perception. As neuroimaging techniques were developed, neural encoding and decoding techniques provided profound understanding about how we process visual information. Advancements in Artificial Intelligence and Deep Learning areas have also influenced neuroscientific research. With the emergence of deep generative models like Variational Autoencoders (VAE), Generative Adversarial Networks (GAN) and Latent Diffusion Models (LDM), researchers also used these models in neural decoding tasks such as visual reconstruction of perceived stimuli from neuroimaging data.

The current thesis provides two frameworks in the above-mentioned area of reconstructing perceived stimuli from neuroimaging data, particularly fMRI data, using deep generative models. These frameworks focus on different aspects of the visual reconstruction task than their predecessors, and hence they may bring valuable outcomes for the studies that will follow. The first study of the thesis (described in Chapter 2) utilizes a particular generative model called IC-GAN to capture both semantic and realistic aspects of the visual reconstruction. The second study (mentioned in Chapter 3) brings new perspective on visual reconstruction by fusing decoded information from different modalities (e.g. text and image) using recent latent diffusion models. These studies become state-of-the-art in their benchmarks by exhibiting high-fidelity reconstructions of different attributes of the stimuli.

In both of our studies, we propose region-of-interest (ROI) analyses to understand the functional properties of specific visual regions using our neural decoding models. Statistical relations between ROIs and decoded latent features show that while early visual areas carry more information about low-level features (which focus on layout and orientation of objects), higher visual areas are more informative

about high-level semantic features. We also observed that generated ROI-optimal images, using these visual reconstruction frameworks, are able to capture functional selectivity properties of the ROIs that have been examined in many prior studies in neuroscientific research.

Our thesis attempts to bring valuable insights for future studies in neural decoding, visual reconstruction, and neuroscientific exploration using deep learning models by providing the results of two visual reconstruction frameworks and ROI analyses. The findings and contributions of the thesis may help researchers working in cognitive neuroscience and have implications for brain-computer-interface applications.

Résumé en Français

Les grands esprits de l'humanité ont toujours été curieux de la nature de l'esprit, du cerveau et de la conscience. Par le biais d'expériences physiques et mentales, ils ont tenté de répondre à des questions difficiles sur la perception visuelle. Avec le développement des techniques de neuro-imagerie, les techniques de codage et de décodage neuronaux ont permis de mieux comprendre la manière dont nous traitons les informations visuelles. Les progrès réalisés dans les domaines de l'intelligence artificielle et de l'apprentissage profond ont également influencé la recherche en neuroscience. Avec l'émergence de modèles génératifs profonds tels que les autoencodeurs variationnels (VAE), les réseaux adversariaux génératifs (GAN) et les modèles de diffusion latente (LDM), les chercheurs ont également utilisé ces modèles dans des tâches de décodage neuronal telles que la reconstruction visuelle des stimuli perçus à partir de données de neuro-imagerie.

La présente thèse fournit deux bases théoriques dans le domaine de la reconstruction des stimuli perçus à partir de données de neuro-imagerie, en particulier les données IRMf, en utilisant des modèles génératifs profonds. Ces bases théoriques se concentrent sur des aspects différents de la tâche de reconstruction visuelle que leurs prédécesseurs, et donc ils peuvent apporter des résultats précieux pour les études qui suivront. La première étude dans la thèse (décrite au chapitre 2) utilise un modèle génératif particulier appelé IC-GAN pour capturer les aspects sémantiques et réalistes de la reconstruction visuelle. La seconde étude (décrite au chapitre 3) apporte une nouvelle perspective sur la reconstruction visuelle en fusionnant les informations décodées à partir de différentes modalités (par exemple, le texte et l'image) en utilisant des modèles de diffusion latente récents. Ces études sont à la pointe de la technologie dans leurs domaines de référence en présentant des reconstructions très fidèles des différents attributs des stimuli.

Dans nos deux études, nous proposons des analyses de régions d'intérêt (ROI) pour comprendre les propriétés fonctionnelles de régions visuelles spécifiques en utilisant nos modèles de décodage neuronal. Les relations statistiques entre les régions d'intérêt et les caractéristiques latentes décodées montrent que les zones visuelles précoces contiennent plus d'informations sur les caractéristiques de bas niveau (qui se concentrent sur la disposition et l'orientation des objets), tandis que les zones visuelles supérieures sont plus informatives sur les caractéristiques sémantiques de haut niveau. Nous avons également observé que les images optimales de ROI générées à l'aide de nos techniques de reconstruction visuelle sont capables de capturer les propriétés de sélectivité fonctionnelle des ROI qui ont été

examinées dans de nombreuses études antérieures dans le domaine de la recherche neuroscientifique.

Notre thèse tente d'apporter des informations précieuses pour les études futures sur le décodage neuronal, la reconstruction visuelle et l'exploration neuroscientifique à l'aide de modèles d'apprentissage profond en fournissant les résultats de deux bases théoriques de reconstruction visuelle et d'analyses de ROI. Les résultats et les contributions de la thèse peuvent aider les chercheurs travaillant dans le domaine des neurosciences cognitives et avoir des implications pour les applications d'interface cerveau-ordinateur.

Contents

List of Publications

During the course of my Ph.D. I have been involved in various other works with varying degrees of capacity. I list all the publications below:

Reconstruction of Perceived Images from fMRI Patterns and Semantic Brain Exploration using Instance-Conditioned GANs
International Joint Conference on Neural Networks 2022
Furkan Ozcelik, Bhavin Choksi, Milad Mozafari, Leila Reddy, Rufin VanRullen
Link to Code : `https://github.com/ozcelikfu/IC-GAN_fMRI_Reconstruction`

Natural scene reconstruction from fMRI signals using generative latent diffusion
Scientific Reports 2023
Furkan Ozcelik, Rufin VanRullen
Link to Code : `https://github.com/ozcelikfu/brain-diffuser`

Multimodal decoding of human brain activity into images and text
NeurIPS 2023 Workshop - UniReps: Unifying Representations in Neural Models
Matteo Ferrante, Tommaso Boccato, Furkan Ozcelik, Rufin VanRullen, Nicola Toschi

Listing of figures

To my Mom who passed away during COVID-19 Pandemic, and all family and friends who supported me through tough times...

Chapter 1

Introduction

Understanding how the brain works, why we have conscious perception, and how we see and think are some of the most challenging questions in the field of cognitive neuroscience. Researchers have begun to use various techniques, including noninvasive neuroimaging methods like fMRI and EEG, to accelerate research in this area. Recently, advanced deep learning models have enabled the detection of patterns in brain signals with unprecedented complexity. This has led to the possibility of 'reading minds' and has opened up new opportunities for neuroscience research in the areas of neural decoding and visual reconstruction. This thesis presents methods for reconstructing and combining high-level (semantic) and low-level (shape and layout) features from fMRI patterns using deep generative models, following the development of visual reconstruction with deep learning models. The first chapter provides relevant background information for our main studies from various perspectives. First, we provide a brief history of mind, consciousness, and vision. We then review the neuroscientific foundations of vision, various neuroimaging techniques, and early studies on neural decoding and visual recon-

struction. Next, we briefly introduce deep learning models and deep generative models. Finally, we show how deep learning is used in neuroscience studies and visual reconstruction. This thesis presents two studies conducted and published throughout the PhD in Chapters 2 and 3. Chapter 2 proposes a natural image reconstruction framework that utilizes an Instance-Conditioned GAN model to perform accurate semantic reconstruction while preserving low-level details from fMRI patterns on the Generic Object Decoding dataset. Chapter 3 presents the 'Brain-Diffuser' framework, a two-stage scene reconstruction approach that employs latent diffusion models to reconstruct high-complexity images from fMRI signals on the Natural Scenes Dataset. Chapter 4 provides an extended discussion of the studies presented in Chapters 2 and 3. Later, we discuss practical applications and ethical implications of our study and neural decoding research in general, and conclude with a summary and closing thoughts.

1.1 Brief History of Mind, Consciousness, and Vision

> *Why is there something rather than nothing? - Gottfried Wilhelm Leibniz*[133]

One of the most intriguing questions in the universe is why there is something rather than nothing. This fundamental question has guided ontology, the philosophical study of being, for centuries. An equally interesting question is why there are minds in this universe that began to exist and expand 13.7 billion years ago from quantum fluctuations[88], capable of asking 'Why is

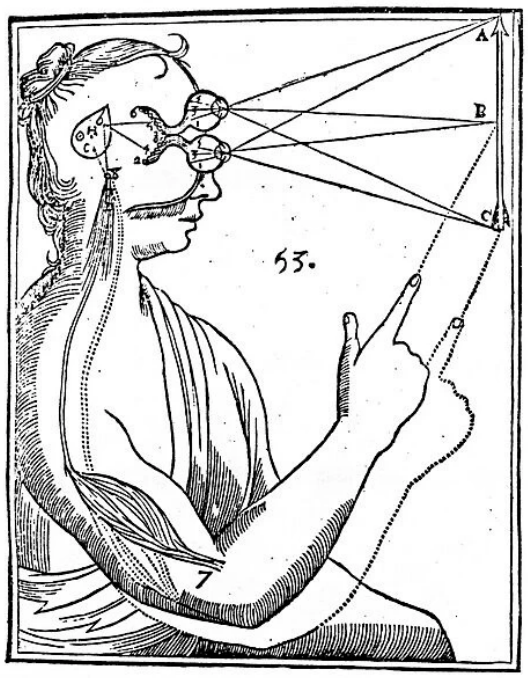

Figure 1.1: A drawing of Descartes' theory of perception. According to the theory, objects are perceived through the eyes and transmitted to the pineal gland, which communicates with the immaterial intellect. Signals are then sent to the muscles to move the arm (Figure from Descartes 1664)[50].

there something rather than nothing?'

The existence of minds has fascinated the brilliant minds throughout history. Philosophers and scientists have studied and discussed the human mind, brain, and consciousness for centuries, yet many puzzling questions remain. Numerous prominent philosophers have delved deeply into the nature of the mind. Plato's concept of the human mind is connected to the idea of the soul. He believed that the human soul consists of three distinct parts: the rational, the emotional, and the appetitive. Plato believed that the rational part is immortal and capable of comprehending the eternal Forms or Ideas[141]. Plato's depiction of the realm of

forms as separate from the physical material world laid the foundation for Western dualism[200]. Aristotle, however, believed that the soul is the form of the body and does not exist independently. He also believed that the intellect consisted of something immaterial in nature. Therefore, his view was not purely materialistic[200]. Through the works of Thomas Aquinas, the idea of different substances for the mind and body continued to influence the Western Christian tradition[200]. Descartes is known to be one of the most prominent proponents of dualism in the late Western philosophical tradition (his drawing of his perceptual theory pictured in Figure 1.1). However, some scholars argue that substances were not the central focus of Descartes' philosophy of mind. Instead, they suggest that his aim was to explain the process of mind-body interaction through the scientific method[37]. Leaving aside the controversial opinions about Descartes' thoughts on the mind, as research in psychology and neuroscience advanced, the concept of dualism became less prevalent among scholars in these fields. Nevertheless, it is known that some prominent philosophers and scientists have defended the idea of dualism on the basis of the authenticity of mental properties, such as the philosopher of science Sir Karl Popper, the Nobel Prize-winning neurobiologist Sir John Eccles, the Oxford philosopher of science and religion Richard Swinburne[180]. During the 20th century, various theories and concepts emerged regarding the nature of the mind and consciousness. Freud's psychoanalytic theory, for example, examined the impact of unconscious processes on behavior and mental states[69]. Behaviorists, including James Watson and B.F. Skinner, attempted to explain all behavior through conditioning, without reference to thoughts or feelings[251,221]. In

contrast to dualism, materialism (or physicalism) is the central idea arguing that everything about the mind can be explained in terms of physical processes and mental states emerge from (or identical to) brain activity. Throughout history, many philosophers have advocated for materialism. Ancient Greek philosophers Democritus and Epicurus are known for their views on materialism. In modern times, materialistic views can be observed in the works of Thomas Hobbes, Julien de La Mettrie, and d'Holbach[222]. Contemporary proponents of materialism include Daniel Dennett and David Armstrong for reductive materialism[49,5] and Richard Rorty, Paul Churchland, and Patricia Churchland for eliminative materialism[191,34]. David Chalmers argued against materialism because it could not account for qualia, or subjective experience[28]. He introduced the 'easy' and 'hard' problems of consciousness. The 'easy' problems can be solved by explaining underlying cognitive functions, behaviors, or mechanisms. On the other hand, the 'hard' problem of consciousness is fundamentally different. It refers to the challenge of explaining why and how subjective experiences arise from physical processes in the brain[27]. Hilary Putnam introduced the concept of functionalism, which emphasizes the function of the cognitive system rather than its internal constitution[182]. Although Putnam later argued against functionalism[219], it still influenced many scholars, such as Daniel Dennett and David Marr, and accommodated the possibility of artificial intelligence having a mind due to its emphasis on multiple realizability. While studies on high-level cognition and behavior have progressed through psychology and cognitive science, there has also been progress in neuroscience describing interactions at the physical and neuronal levels. Researchers

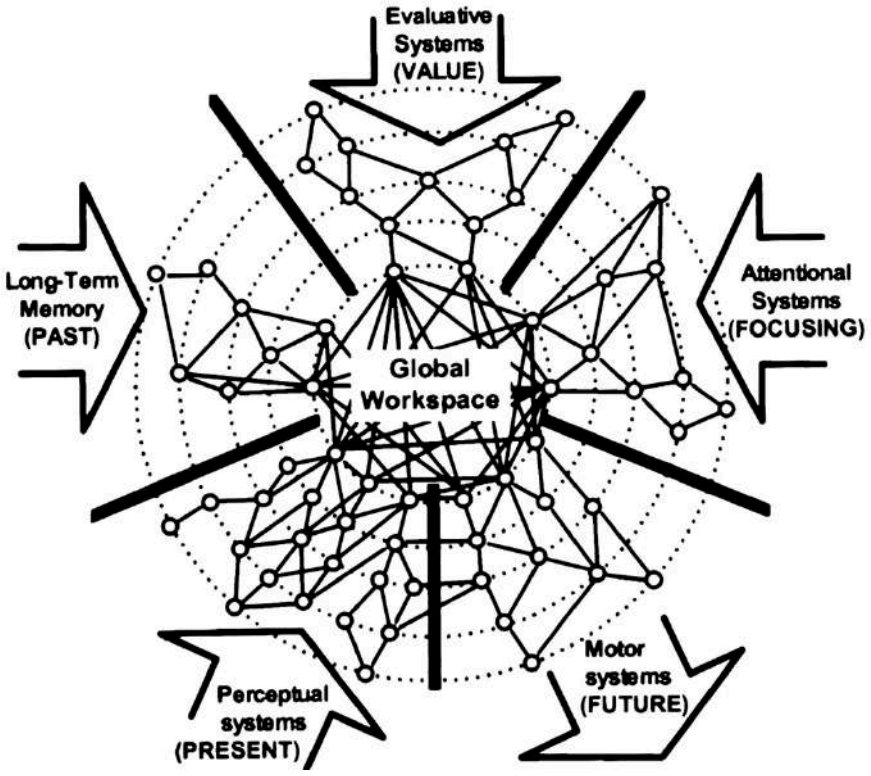

Figure 1.2: Structure of Global Workspace. Information is integrated and broadcasted throughout the system (Figure from Dehaene et al. 1998)[47].

have explored ways to bring these cognitive, behavioral, and neuroscientific studies together through different theories[35]. We began to observe studies and theories on cognition, behaviour and consciousness that are more neuroscientifically plausible such as Neural Correlates of Consciousness[41], Global Workspace Theory[7] (later developed as Neuronal Global Workspace Theory[47], illustrated in Figure 1.2), Integrated Information Theory[238]. Human vision and perception have always been central to behavioral, cognitive, and neuroscientific studies and have played a pioneering role in the study of the senses.

Although we have discussed the historical journey of mind and consciousness

above, we will now highlight some of the key historical milestones in the research of vision and perception. While philosophers such as Plato, Aristotle, and Empedocles provided varying explanations about vision, the first mathematical explanation of vision was observed in Euclid's Book of Optics (Optica)[139]. Following Euclid's work, the most notable studies in these fields include Ptolemy's Optics and Galen's research on visual anatomy in the second century[102]. These philosophers and scientists advocated for the theory of extramission, which suggests that light is emitted from the eyes to objects, rather than from objects to the eyes (intromission). The only exception is Aristotle, who made statements in favor of both theories, although he is better known as one of the first advocates of the intromission theory of vision[223]. Vision and the eye were important topics in Islamic medieval medicine and philosophy. Al-Kindi, one of the most influential philosophers of the era, studied vision and advocated for the extramission theory of vision. However, objections against extramission theories were first raised in the studies of Rhazes (Abu Bakr Muhammad ibn Zakariya al-Razi)[139]. Avicenna (Ibn Sina) and Averroes (lbn Rushd) also studied Aristotelian vision and supported the intromission theory of vision[139]. Alhazen's Book of Optics (Kitāb al-Manāẓir) is considered one of the most influential works of the era, in which he synthesized experimental observations and mathematics to describe vision more systematically than any previous work[102] (his drawing of human eye shown in Figure 1.3). Along with the studies of Kepler, Alhazen's work has continued to be influential until contemporary researches in vision. It is important to note that vision involves more than just optics and the emission of light.

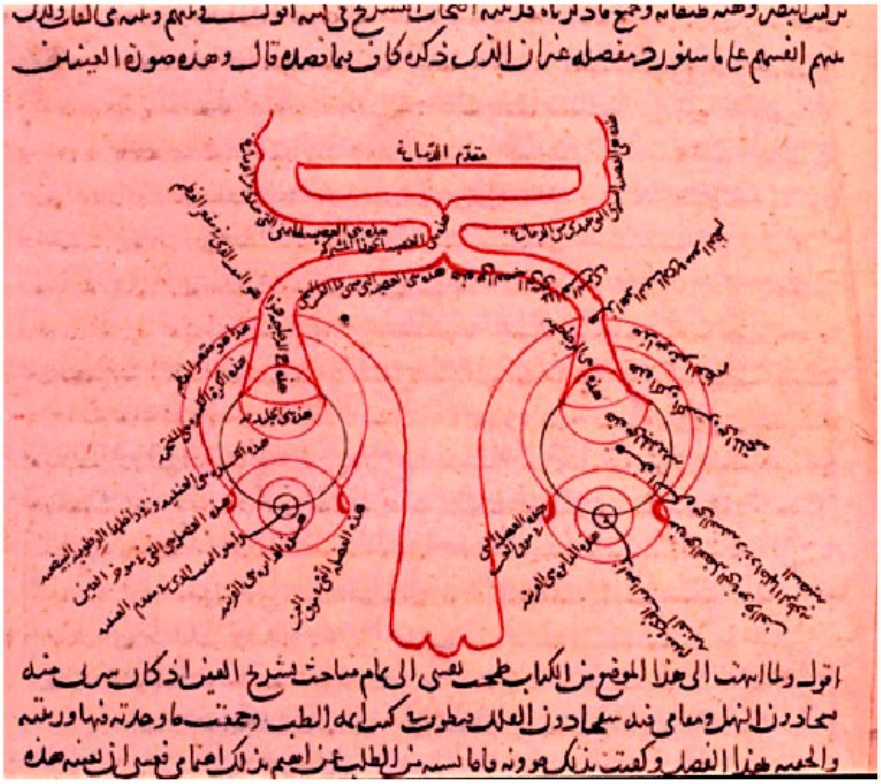

Figure 1.3: A drawing of the structure of the human eye, including the optic chiasm by Alhazen in the Book of Optics (Kitāb al-Manāẓir). Adapted from Daneshfard et al. 2014[43,106]

Vision and sight may seem like simple tasks for humans, but in reality, our perception is not a direct reflection of the outside world. Rather, it is the mental representations of the stimuli that we see that shape our perception. Alhazen was aware that there is an unconscious inference process in our visual perception, and the brain concludes more than the available sense data offers[215]. Hermann von Helmholtz coined the term 'unconscious inference' to describe how human vision functions by inferring the best interpretation based on sensory data, as explained in his book 'Physiological Optics'[46]. Similar studies in neuroscience and vision in the nineteenth and twentieth centuries demonstrated how our nervous system represents the external world. Our brain recreates the representation of the world outside as our perception of the world, and it is not directly given to us[46]. Even prior to the study of the neuroscience of perception, Immanuel Kant pointed out that our perception of the world is a subjective interpretation that is formed by our own cognition, rather than a direct reflection of reality[115]. Neuroscientific research on consciousness and perception has shown that vision is a complex process with various aspects. For instance, brain lesions in different areas can affect vision and perception in diverse ways. For example, it can result in visual agnosia, a condition in which the patient is unable to recognize objects, or other types of agnosia, such as prosopagnosia, which is the inability to recognize faces[207]. Another example is blindsight, where patients are able to respond to visual stimuli beyond chance level despite lacking conscious perception due to lesions in the striate cortex[35]. Another interesting case is neglect, which is a deficit in visual consciousness where patients become unaware of things in their left perceptual field after damage, resulting

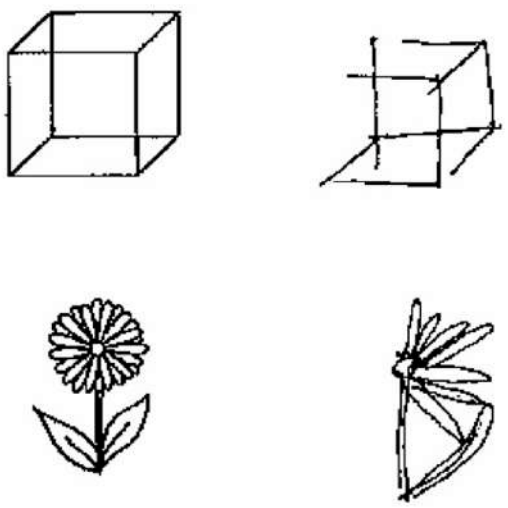

Figure 1.4: Copy drawings of neglect patients from neurophyschological tests showing the loss of left perceptual field due to missing parts of the objects (Figure from Husain 2008)[105].

in the loss of half of their phenomenal space. In neuropsychological tests, they draw a half copy of the presented objects[199] (illustrated in Figure 1.4). These cases of different deficits in visual consciousness demonstrate that the brain has a binding mechanism that integrates various aspects of visual perception. This mechanism is still being discussed among neuroscientists, cognitive scientists, and philosophers of mind[66].

1.2 How Do We See?: Neuroscience of Vision

We will begin by discussing the neuroscience of vision, which plays a crucial role in our perception. The process of seeing begins with our eyes, which are an extension of our central nervous system that faces outward. Light reflected from objects in the outside world passes through the cornea and lens before reaching the photoreceptors in the retina. These photoreceptors are made up of cells known

as rods and cones. Rod cells are more effective for vision in low light conditions, while cone cells are responsible for detecting details and colors. The three types of cone cells respond to wavelengths in the blue, green, and red regions of the color spectrum. The retina contains more rod cells than cone cells, with cone cells being concentrated in the central foveal region. The optic nerve connects the ganglion cells to the brain, transmitting the signals produced by the photoreceptors when they are stimulated by light[55]. Information processing begins at this stage of vision. Humans have approximately 100 million photoreceptors and 1 million ganglion cells, indicating compression in the early stages[153]. The signals from the optic nerves reach the primary visual cortex (V1) by passing through the optic chiasm and then the lateral geniculate nucleus (LGN) in the thalamus via optic radiation[74] (shown in Figure 1.5).

Ganglion cells have receptive fields, meaning they respond to stimulation in specific areas or are selective for certain stimulus locations[87]. For instance, a ganglion cell may respond to light in the lower left corner of the visual field. These receptive fields are composed of two regions: center and surround. Some cells are stimulated by light from the center, while others are stimulated by light from the surround[130,8]. The stimulation relies on the contrast between the center and surround, making the border regions with a sudden change in light more apparent. This center-surround receptive field structure is preserved in the LGN. Moving from the LGN to the primary visual cortex, the structure combines multiple contrast information from the LGN. The primary visual cortex has two cell types: simple cells and complex cells[114]. Hubel and Wiesel discovered that simple cells

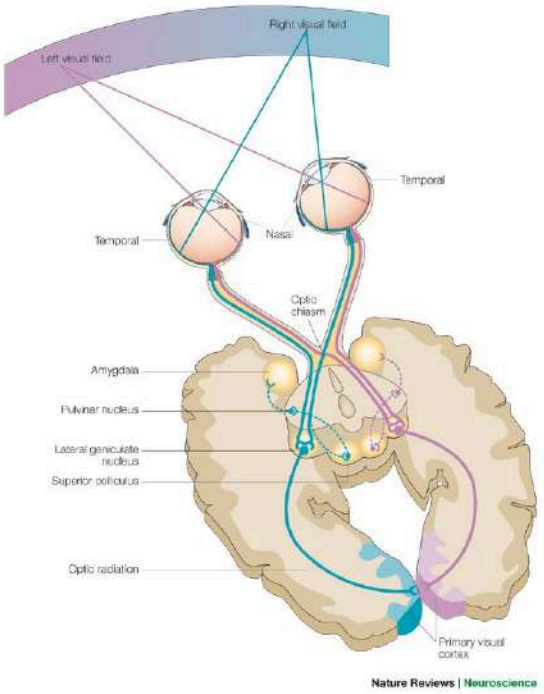

Figure 1.5: Visual pathway from eye to visual cortex (Figure from Hannula et al. 2005)[84].

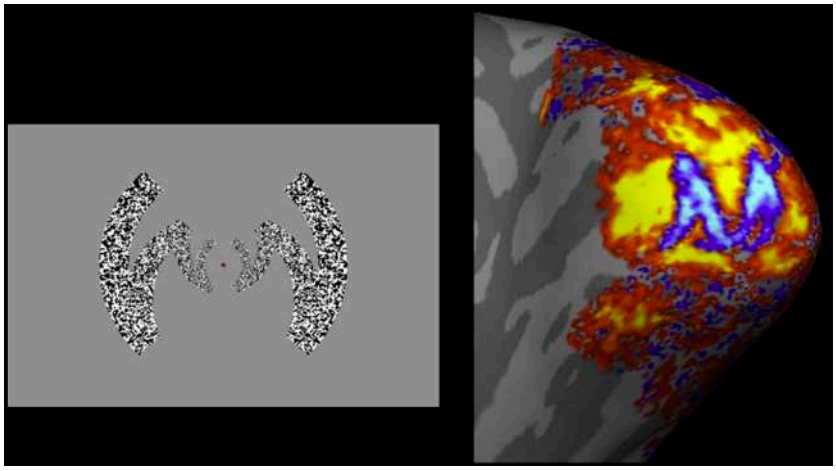

Figure 1.6: The stimulus condition on the left is designed to evoke M-shaped brain activity in fMRI response, as demonstrated on the right (Figure adapted from Polimeni et al. 2010)[179].

are stimulated by edges of a certain orientation, while complex cells have a more abstract stimulation structure[104]. Complex cells can respond to edges over a large area, regardless of the exact location of the stimulus.

Both V1 and LGN exhibit retinotopic characteristics. The light entering the retina follows a topographic structure that corresponds to the patterns of activity in these areas[13]. The retinotopic map structure can be observed when a stimulus is applied through the retina and the signals in these regions are examined. Due to the cross-connection that occurs prior to transmission to the primary visual cortex, the right part of our visual field is formed in V1 of the left hemisphere and the left part of our visual field is formed in V1 of the right hemisphere[13]. However, the order in which light enters the retina is preserved which indicates the retinotopy in these regions (An example of retinotopic structure in V1 shown in Figure 1.6).

Various features, including ocular dominance, orientation selectivity, retino-

topy, color, and spatial frequency, are already beginning to be processed in V1[177]. Moving hierarchically through visual cortex, we observe V2, V3, and V4. V2 is involved in more complex shape analysis than V1. Approaching V3, the visual system is divided into two paths, dorsal and ventral. In this region, angle and orientation are analyzed. In the region known as V3a, information about movement and direction is processed[177]. In the V4 region, distinctions are made based on colors and shapes[259,258]. In the literature, the ventral and dorsal visual processing pathways are commonly referred to as the 'what' and 'where' pathways. Some argue that the dorsal pathway encodes action-related information and should be labeled as the 'how' pathway instead of the 'where' pathway[177]. The ventral pathway originates from the visual cortex and extends to the inferior temporal cortex. It plays a major role in object categorization. The dorsal pathway originates from the visual cortex and extends to the posterior parietal cortex. It is responsible for processing information related to the object's location[13].

The regions involved in visual processing are not limited to V1-V4. As we advance through the visual pathway, we encounter specialized regions that process various visual features or are selective to different categories, such as scenes, faces, bodies, hands, words, numerals, and tools[173]. The identification of these specialized functions emerged with lesion studies and was later supported by neuroimaging techniques. For instance, the Lateral Occipital Complex (LOC) is crucial in recognizing objects. Lesions in this area can cause agnosia, which is the inability to recognize objects. The neurons in this area show stronger activation in response to images of objects than to scrambled controls. Additionally, these neurons exhibit

long response latencies, spatially clustered shape selectivity, large and bilateral receptive fields, and 3D-structure selectivity[78,45]. The Colour Centre (CC) region, located within V4, is responsible for color processing and lesions in this region can cause achromatopsia, a form of color blindness[142,9]. The Visual Medial Temporal (MT) area, also known as V5, is composed of neurons that are sensitive to movement. If this region is damaged, patients may experience akinetopsia, a motion blindness disorder that impairs their ability to perceive motion[3,17]. The Fusiform Face Area (FFA) is responsible for detecting and identifying faces. Studies using fMRI have shown that the FFA responds more strongly to face stimuli than to non-face stimuli. Lesions in the FFA can result in prosopagnosia, the inability to recognize faces[116,77]. The Parahippocampal Place Area (PPA) is a brain region that encodes information about the layout of local space and responds strongly to scenes with spatial layout[62,94]. Damage to the PPA can can cause memory issues with topographical information and difficulties with navigating unfamiliar environments[61]. The Extrastriate Body Area (EBA) is a body-selective region located in the lateral occipitotemporal cortex that responds strongly to images of human bodies and body parts in comparison with other classes of stimuli[56]. It also contributes to planning goal-directed actions[264]. The Visual Word Form Area (VWFA) is a specialized region located in the left occipitotemporal sulcus that plays a crucial role in the recognition of written words[147]. It is connected to the language system, specifically Wernicke's area, and is specialized for processing real words lexically[230]. As mentioned above, certain areas of the visual system show functional sensitivity, and studies have shown that even individual neurons

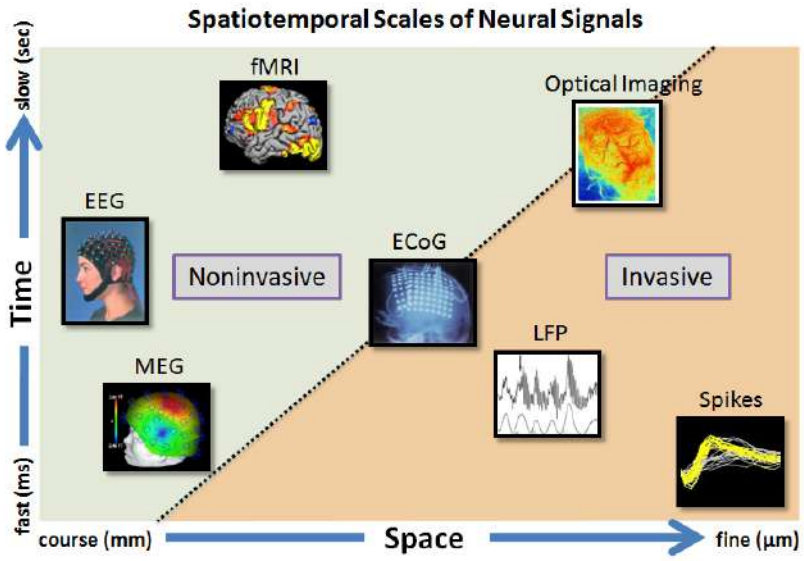

Figure 1.7: Neuroimaging methods (EEG, MEG, fMRI, ECoG, LFP, Optical Imaging and Spikes) are presented according to their properties on spatial resolution, temporal resolution and invasiveness (Figure from Thakor 2012)[236].

can be selective for particular concepts. Quiroga et al. conducted a study on single-cell recordings of neurons in the human medial temporal lobe (MTL). The study found that certain neurons had an invariant representation of specific concepts, places (such as the Sydney Opera), or persons (such as Jennifer Aniston or Halle Berry). These neurons were activated even when the images of the same concept were shown with different viewing angles and luminance, and even when presented as text[184]. This study indicates that the brain contains detailed representations of objects and concepts. The question is how to approach these representations, which is where neuroimaging techniques and neural decoding come into play.

1.3 Neuroimaging

Until the twentieth century, there was no way to record brain activity. Although researchers could learn about the anatomical structure of the brain from dead brains, acquiring signals from a living brain was a challenge[35]. The development of functional neuroimaging methods has made it possible to detect signals from brain activity. Various neuroimaging methods with distinct properties measure different aspects of brain activity. Firstly, invasive neuroimaging methods, such as single-cell recording and ECoG, can be mentioned. Single-cell recording measures action potentials produced by individual neurons through an electrode placed on a neuronal membrane via surgical operation[74]. Electrocorticography (ECoG) is another invasive method for studying the human brain. This method is useful for studying brain activity at a larger scale. ECoG electrodes are larger than single-cell recording electrodes and measure the activity of a population of neurons. The signal is clear and has a sufficient spatial and temporal resolution for many tasks, with minimal noise or distortion[74]. However, non-invasive methods such as EEG, MEG, PET, fMRI, and fNIRS are preferred over invasive methods like single-cell recording and ECoG due to their lower cost and risk[19]. Similar to ECoG, Electroencephalography (EEG) measures electrical potential. However, electrodes are placed non-invasively on the scalp instead[227]. EEG signals are weaker than ECoG signals, but they offer the advantage of acquiring high temporal resolution signals without the need for surgery. It is possible to obtain data on how a particular task affects brain activity. The response signals of movements

or external stimuli in these tasks are called event-related potentials (ERP)[152]. Magnetoencephalography (MEG) is a technique similar to EEG. However, MEG measures the magnetic fields produced by the brain's electrical activity. Unlike EEG, which is distorted by the skull and scalp's effects on electrical activity, MEG signals are better preserved and easier to localize[151]. One downside of MEG is its higher cost compared to EEG. Both EEG and MEG measure neural activity through electrical potentials and magnetic fields, while PET, fNIRS, and fMRI measure metabolic changes caused by brain activity[134]. Positron Emission Tomography (PET) is a medical imaging technique that detects radioactive-labeled compounds to indicate brain activity[134]. Functional near-infrared spectroscopy (fNIRS) measures brain activity by using near-infrared light to measure oxygen levels[175]. Although it is portable and affordable, its spatial resolution does not match that of PET and fMRI. Functional Magnetic Resonance Imaging (fMRI) is a commonly used method because it does not require the use of radioactive tracers and it has a high-spatial resolution. fMRI detects changes in blood flow and oxygen levels caused by brain activity through the magnetic field of the scanner[44]. This thesis focuses on datasets recorded using the fMRI neuroimaging technique. Therefore, we will provide a brief explanation.

Seiji Ogawa and his colleagues invented functional magnetic resonance imaging (fMRI), a non-invasive method that visualizes whole-brain activity without the need for injections[164]. When a population of neurons increases its activity, blood flow to the veins near those neurons increases within seconds, providing the necessary oxygen and glucose. Importantly, there is a difference between oxy-

genated and deoxygenated hemoglobin, the latter being paramagnetic and acting as a small magnet[44]. fMRI detects small distortions caused by deoxygenated hemoglobin and measures the ratio of oxygenated to deoxygenated hemoglobin in the blood. This ratio is referred to as blood-oxygen-level-dependent (BOLD) signals. BOLD responses are an indirect measure of neuronal activity since they appear and peak seconds after neuronal activity. While the temporal resolution of fMRI is lower than that of EEG, the spatial resolution is superior[134] (illustrated in Figure 1.7). fMRI is an ideal choice for measuring brain activity for visual tasks such as visual decoding and reconstruction due to its high spatial resolution, sufficient temporal resolution (with an adjusted experiment design), and non-invasive nature.

1.4 Early Studies in Neural Decoding and Visual Reconstruction

> *In principle, you can decode any kind of thought that is occurring in the brain at any point in time…. you can think about this like writing a dictionary. If you were, say, an anthropologist, and you went to a new island where people spoke a language that you had never heard before, you might slowly create a dictionary by pointing at a tree and saying the word "tree," and then the person in the other language would say what that tree was in their language, and over time you could build up a sort of a dictionary to translate between your language and this other foreign language. And we essentially play the same game in neuroscience - Jack Gallant*[178]

Nancy Kanwisher and her colleagues' discovery of selective regions for different objects in the temporal lobe raised questions for some researchers, including Isabel Gauthier. Gauthier disagreed with the strong claims about the localization of face perception and conducted experiments that led her to believe that FFA was not a 'face area' but an 'expertise area.' This area is engaged whenever people see objects that they have a lot of expertise in. However, Kanwisher and her colleagues were not convinced. They criticized Gauthier's experiments and interpretation[178]. To put an end to these debates, James Haxby employed a different method to understand the relationship between brain regions and object categories. Prior to this, researchers had been focusing on the amount of activation in a particular region in one condition compared to another. Haxby designed the decoding framework, which uses a model to predict which object has been shown based on an fMRI pattern. As a result of his initial study, he asserted that the representations of faces and objects in the ventral temporal cortex were widely distributed and overlapping[90]. In a subsequent study, Spiridon and Kanwisher challenged some of Haxby's claims and demonstrated that category-selective regions, such as FFA and PPA, lacked sufficient information to accurately distinguish between two non-preferred stimuli[226]. O'Toole and Haxby later came to a similar conclusion, stating that the regions preferred for faces and houses may not be enough to classify non-preferred objects[168]. Reddy and Kanwisher revised their initial claim after discovering that it is possible to differentiate between two non-preferred objects, such as shoes and cars, by analyzing the signals in FFA and PPA using support vector machines, despite their weakness[194]. However, these results do not

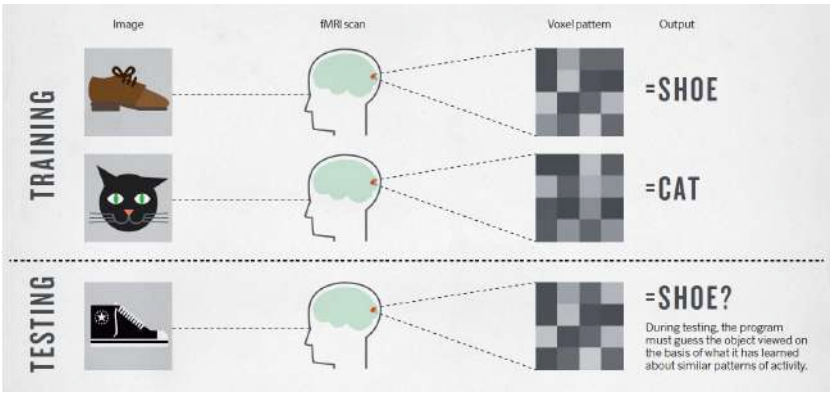

Figure 1.8: A simple visual explanation of neural decoding. During the training phase, voxel patterns are recorded for two types of stimuli: shoes and cats. In the test phase, the type of stimulus is predicted based solely on the voxel pattern. (Figure from Smith 2013)[224].

affect modularity or category selectivity in the mentioned regions. It is important to note that finding weak signals to discriminate between two non-specialized categories does not necessarily mean that the region is not specialized in a particular category. For instance, a neural network trained solely on facial images can still provide features to differentiate between shoes and cars, but its specialization is in facial perception[166]. Although debates and controversies continued regarding the representation of categories in the brain, whether it is distributed or modular, one of the key outcomes has been the invention of neural decoding.

Neural decoding is a method in which a model predicts information about the stimuli, such as category, orientation, and color, from brain signals obtained through neuroimaging methods (simple description presented in Figure 1.8). Its inverse, neural encoding, involves predicting neural activity based on given stimuli. The thesis will primarily focus on decoding, specifically reconstructing images from fMRI signals. Early decoding studies were based on Haxby's analysis of neural

response patterns, which involved distributed representations of categories in the brain, also known as multivariate pattern analysis (MVPA)[89]. Cox and Savoy developed a support vector machine (SVM) model to predict the class of various objects, such as birds, chairs, horses, and teapots, from fMRI signals[40]. Hanson et al. reanalyzed the object recognition data from Haxby et al. (2001) using feed-forward neural networks.[86]. Carlson et al. and O'Toole et al. both utilized linear discriminant analysis on the same data[23,168]. Ken Norman conducted one of the earliest research reviews in this field with James Haxby. They referred to these analyses as Multi-voxel pattern analysis initially, but later changed it to multivariate pattern analysis (MVPA)[163]. In early studies of visual decoding, researchers decoded various properties of visual stimuli. Two studies, one by Haynes and Rees and the other by Kamitani and Tong, focused on decoding edge orientation from oriented and masked gratings[91,113] (depicted in Figure 1.9).

Thirion et al. attempted to reconstruct whole stimuli using retinotopic structures of visual areas instead of decoding a particular aspect of stimuli. They were able to reconstruct simple shapes from brain signals[237]. Later on, Miyawaki et al. reconstructed simple geometric and alphabetic shapes using multiscale local image decoders[154] (presented in Figure 1.10). Following the success in visually decoding and reconstructing basic shapes, researchers at the Gallant Lab have taken on the challenge of working with natural images. In their first study, Kay et al. used models to estimate receptive fields for voxels and deployed them to identify test images from a set of candidates.[119] (showed in Figure 1.11). Naselaris et al. created a reconstruction model using the structural and semantic

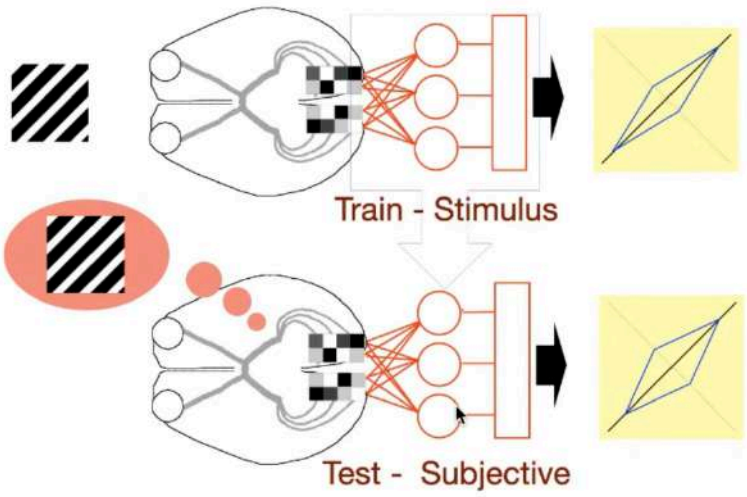

Figure 1.9: Decoding of edge orientation from oriented gratings. Top: Decoding the orientation of the presented gratings. Bottom: Decoding the orientation of the imagined gratings. The model predicts the orientation of the gratings (solid black line) with its corresponding uncertainty (Figure from Neuromatch Conference 2022)[38,113].

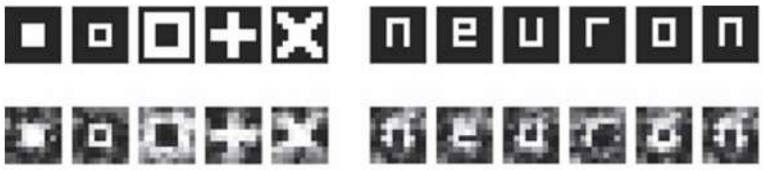

Figure 1.10: Reconstruction examples of simple geometric and alphabetic shapes from Miyawaki et al.[154].Top: Test images that are presented to the subjects while the fMRI signals are being recorded. Bottom: The reconstructed images that were decoded from the fMRI signals of the subjects. (Adapted from Miyawaki et al. 2008)[154].

information of stimuli images within a Bayesian framework, based on the same data used in the previous study[158].Nishimoto et al. established a model to reconstruct natural movies using a Bayesian approach, taking the task complexity even further than static image reconstruction[161]. Researchers have even gone beyond decoding presented stimuli and established models for decoding mental images when participants are awake[195] or asleep[101]. Although the initial visual decoding and reconstruction results appeared promising, certain bottlenecks prevented the reconstruction quality from advancing beyond a certain point. One of these bottlenecks is the inadequate representation of stimuli caused by hand-crafted feature extractors or priors. To mitigate this problem, researchers have begun to use features extracted from deep learning models, which provide richer representations capable of representing complex stimuli at different levels of hierarchy.

1.5 Introduction to Deep Learning

The origins of Deep Learning (DL) and Deep Neural Networks (DNN) can be traced back to the early stages of computational neuroscience and artificial intelligence research. The milestones of the connectionist AI paradigm and machine learning, as opposed to symbolic AI, can be seen as a prequel to deep learning. These works include numerous studies and models, beginning with the McCulloch-Pitts neuron model[148], Hebbian Learning[93], Perceptrons[204], Backpropagation[253], Hopfield Networks[99], Boltzmann Machines[97], Multilayer Perceptrons[205], Support Vector Machines[39], Long Short Term Memory Networks[98], LeNet[132], Deep Belief Networks[96], and others. Although the connectionist AI paradigm has been around

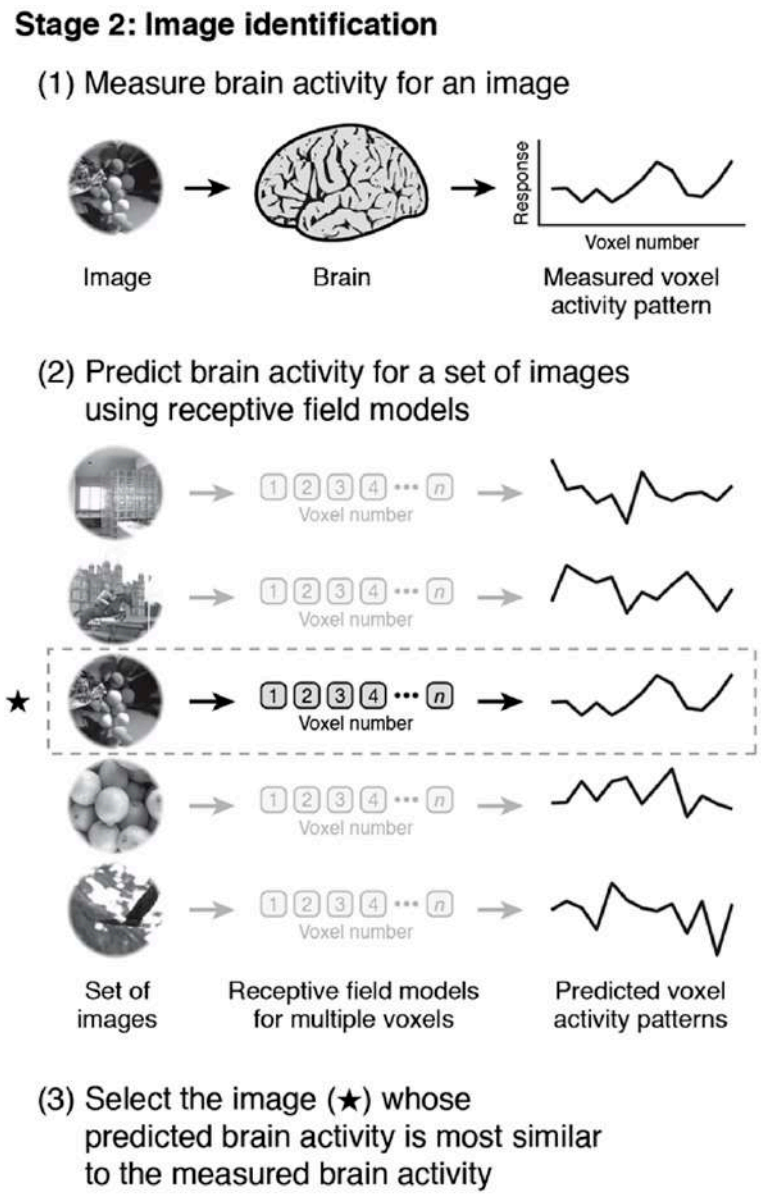

Figure 1.11: Image identification stage of Kay et al.[119] The first step involves recording fMRI responses. In the second step, brain activity predictions are obtained using receptive field models (neural encoding). Finally, the closest image is selected based on the distance in brain activity response (Adapted from Kay et al. 2008)[119].

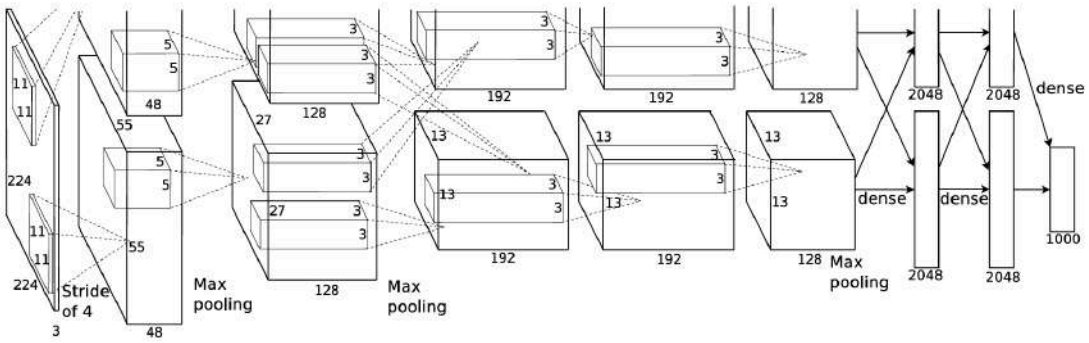

Figure 1.12: The structure of the AlexNet model consists of convolutional, maxpooling, and dense layers. The model is divided into two pathways, which accelerates model training on two GPUs. (Figure from Krizhevsky et al. 2012)[129].

for eight decades, it was not always the main paradigm of AI. Other approaches, such as symbolic AI, were more popular during certain periods in the history of AI. Despite its achievements in important tasks, such as playing chess, planning, and scheduling, symbolic AI was not sufficient for tasks that were relatively easy for humans, such as object recognition. Machine learning models such as support vector machines, multilayer perceptrons, or LeNet-like convolutional neural networks (CNNs) have been used for these tasks. However, their performance did not match that of humans either.

After collecting large datasets such as ImageNet-1M, improving computation with the development of graphical processing units, and designing better models, AlexNet started the era of deep learning in 2012 by winning the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), where it significantly outperformed the second-place competitor, reducing the top-5 error from 26% to 15.3%[129]. AlexNet is a deep CNN model with 60 million parameters. It consists of 8 layers, including 5 convolutional layers and 3 fully connected layers (the struc-

ture of AlexNet is shown in Figure 1.12). This model was purposely designed for GPU training, allowing parallel processing and reducing training time - a critical factor when working with millions of images. Following AlexNet's groundbreaking performance, several CNN models with deeper architectures, more parameters, and improved designs have emerged. These models have consistently demonstrated better performance on the ILSVRC challenge, including VGG[220], GoogLeNet[232], and the first model to pass 100 layers which is ResNet (with 101 and 152 layer variations)[92]. With the success of image classification, researchers have developed CNNs for various tasks, including image detection (e.g. Faster-RCNN[197], YOLO[196]) and segmentation (e.g. DeconvNet[162], U-Net[203]).

Deep learning models have shown promising results in computer vision tasks, particularly discriminative ones as we mentioned above. Researchers began designing image generation models, including Variational Autoencoders (VAE)[122], Generative Adversarial Networks (GAN)[75], and Auto-regressive models[242]. VAE is considered the first generative model of the deep learning era. It is relatively easy to train compared to other generative models. However, its performance is limited, often resulting in blurry and unrealistic image generation. While GANs were popular for their sharp and realistic generations, they have been prone to problems such as mode collapse, which researchers have tried to mitigate through various studies. Although autoregressive models were developed later, they have not gained popularity due to their high computational requirements and slow processing. Various architectures have been developed following the initial models, such as β-VAE[95], VD-VAE[32], VQ-VAE[243] for Variational Autoencoders; Deep Convolutional GAN

(DCGAN)[186], Wasserstein GAN[4], Relativistic GAN[111], Pix2Pix[108], CycleGAN[262] for Generative Adversarial Networks; PixelRNN[242] and PixelCNN[241] for Autoregressive Models.

While deep learning models have shown impressive performance in computer vision tasks, natural language processing (NLP) has not progressed as rapidly in AI research until 2018. The development of transformer models can be considered a significant milestone, as they have revolutionized the field of NLP and have become the foundation for many discriminative and generative models, including CLIP and GPT[246]. Models such as Bidirectional Encoder Representations from Transformers (BERT)[51] and Generative Pretrained Transformers (GPT)[187] demonstrated exceptional performance in NLP tasks, including text classification and generation. Later, the transformer architecture has also been applied to computer vision, as seen in models like Vision Transformer (ViT)[54] and DeiT[239]. Meanwhile, researchers were searching for a variety of techniques to get the models to learn useful features without requiring too much supervision. Contrastive learning techniques have become popular for this purpose. They have been used not only for unimodal models but also for multimodal models that include both text and image modalities. One such model is the Contrastive Language-Image Pretraining (CLIP) model, which processes text and image inputs through two different backbones until they are reduced to a latent representation of the same dimension. In this hidden space, the model compares two latent variables from text and images and is trained using a contrastive loss[185].

Meanwhile, the use of Generative Adversarial Networks (GANs) in image

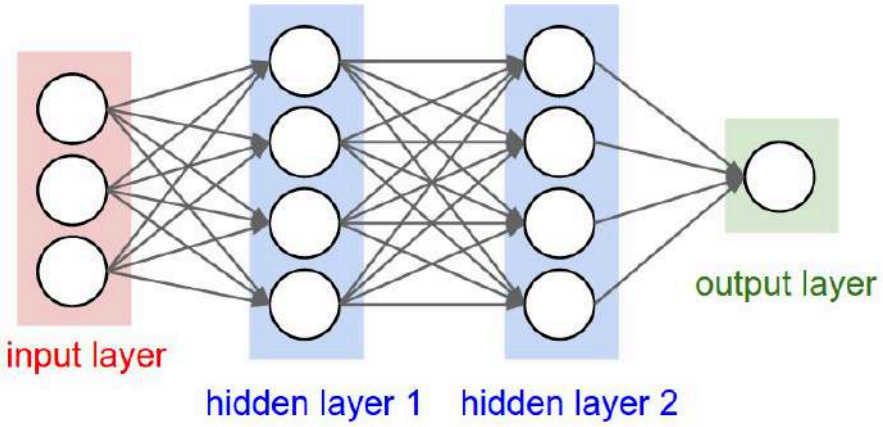

Figure 1.13: Structure of a simple Deep Neural Network with two hidden layers (Figure from Stanford University CS231n Course 2017)[240].

generation research had become saturated. However, models such as Vector Quantized Variational Autoencoder (VQVAE)[243,193], Vector Quantized GAN (VQ-GAN)[63], and DALL-E[190] demonstrated the potential to generate images with greater flexibility without compromising realism. Later, diffusion models have become the latest invention in generative models and are used for image generation (described in section 1.5.3). Diffusion was a groundbreaking technique, but its application to image space can be costly due to its high dimensionality[225]. As an alternative, researchers have developed latent diffusion models (LDMs) that apply diffusion to latent space. An example of LDMs is the Stable Diffusion model[202]. After reviewing the brief history of AI, we can now delve into the basics of deep learning models.

1.5.1 Deep Neural Networks

Deep Neural Networks (DNNs) are artificial neural networks that consist of multiple layers. They are designed to model complex patterns in data by processing them through these layers. The input layer of a DNN receives input samples and passes them to the next layer. Hidden layers, located between the input and output layers, process inputs from previous layers using similar operations (illustrated in Figure 1.13). The output layer is the final layer. These inputs are multiplied by a weight matrix, summed, and then a nonlinearity function, such as a sigmoid or rectified linear unit (ReLU), is applied[160]. The output layer predicts the target value and compares it to the actual target value using a loss function, such as mean squared error for regression tasks or cross-entropy for classification tasks. After calculating the loss, the network's performance can be improved by adjusting the weights using an algorithm called backpropagation. Backpropagation calculates the gradient of the loss function with respect to the network's weights and optimizes them using optimization techniques such as stochastic gradient descent (SGD)[160].

1.5.2 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a type of neural network designed specifically for processing data with a grid-like topology, such as images (which can be represented as a 2D grid of pixels)[117]. The input layer receives images with height, width, and channels instead of a 1D vector. The convolution layer is the core building block of CNNs, utilizing the spatial neighborhood. It applies a filter

to the input, with each filter being spatially small but having the same number of channels as the input. The filter slides around the input and produces a feature map from the responses[117]. This operation extracts features from the input image. Activation functions, also known as nonlinearity, are applied in a similar manner to those used in DNNs. In addition to the convolution layer, a CNN architecture may include various types of layers, such as the pooling layer, which reduces spatial size by using max or average operations, the normalization layer, which stabilizes learning by normalizing input statistics (e.g., Batch Normalization)[107], or the dropout layer, which prevents overfitting by randomly deactivating some of the units in the network during the training phase[228]. CNNs are typically utilized for data with 2D structures and neighborhood information. This is because they reduce the number of trainable parameters compared to DNNs, which have dense connections in every layer. Additionally, CNNs are less prone to overfitting due to the reduced effect of the curse of dimensionality[260].

1.5.3 Deep Generative Models

As mentioned above in the brief history of deep learning, data can be generated using deep generative models such as VAEs, GANs, auto-regressive models, and latent diffusion models (LDM). Our studies also involved variations of these models, including Instance-Conditioned GAN (IC-GAN), Very Deep VAE (VDVAE), and Versatile Diffusion (VD) models. This section provides technical details of these models.

Variational Autoencoders (VAEs) are a type of autoencoder that use a prob-

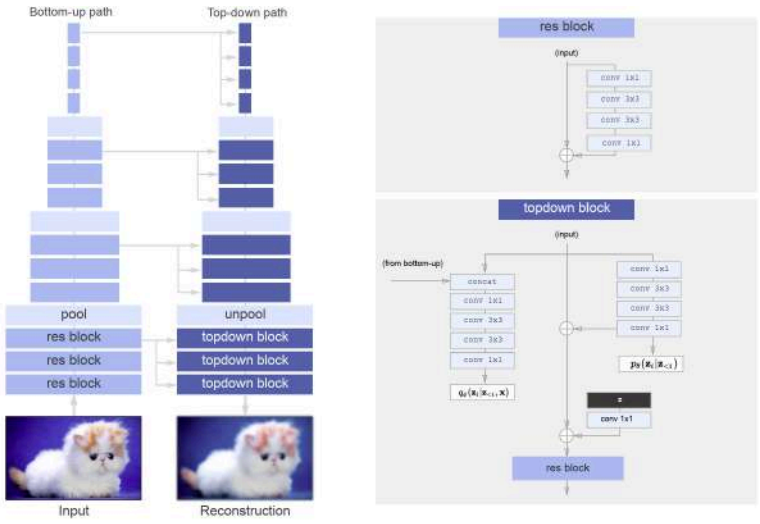

Figure 1.14: Structure of VDVAE. It consists of many hierarchical latent variables used to generate images from latent values (Figure from Child 2020)[32].

abilistic mapping for the latent space[122]. Unlike traditional autoencoders, which map the input directly onto a latent vector, VAEs map the input data into Gaussian distributions with means and variances[68]. The architecture of VAE consists of two main parts: the encoder and the decoder. The encoder part takes an input x and computes the parameters of the latent distribution $q_\varphi(\mathbf{z}|\mathbf{x})$, where φ presents the parameters of the encoder neural network[21]. The decoder component attempts to reconstruct the input data $p_\theta(\mathbf{x}|\mathbf{z})$ using the latent samples z, with θ representing the parameters of the decoder neural network[21]. To compute backpropagation for VAE training, the reparameterization trick is used. This involves sampling from a standard normal distribution and then shifting the samples based on their mean and variance[165]. The loss function for VAEs is the sum of the reconstruction loss, which measures the difference between the predicted and actual samples, and the Kullback-Leibler (KL) divergence between the learned

distribution and the prior distribution, which is typically a normal distribution[21]. Chapter 3 presents a study that utilized a type of VAE known as Very Deep VAE (VDVAE). VDVAE is a model with multiple hierarchical latent variables[33] (shown in Figure 1.14). We explain the details of the model in section 3.2.3.2.

$$\min_G \mathcal{L}_G = \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))] \tag{1.1}$$

$$\max_D \mathcal{L}_D = \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))] \tag{1.2}$$

In 2014, Ian Goodfellow and his colleagues introduced Generative Adversarial Networks (GANs), and they've become increasingly popular[75]. The GAN is a generative model composed of two networks: the generator and the discriminator. The generator produces new data instances from a random noise input, aiming to generate data that is similar to the real data obtained from the dataset and mimic its distribution[22]. The discriminator assesses data instances from both the generator and the dataset, attempting to differentiate between real data from the dataset and fake data generated by the generator. The discriminator adjusts its weights based on its errors in predicting the authenticity of both the real and generated data, while the generator adjusts its weights based on the discriminator's responses[22].After a certain amount of training, the generator's output becomes indistinguishable from real data. The generator G attempts to reduce the probability of the discriminator classifying its outputs as fake, as shown in Equation 1.1[75]. The discriminator D aims to maximize the probability of predicting real and fake

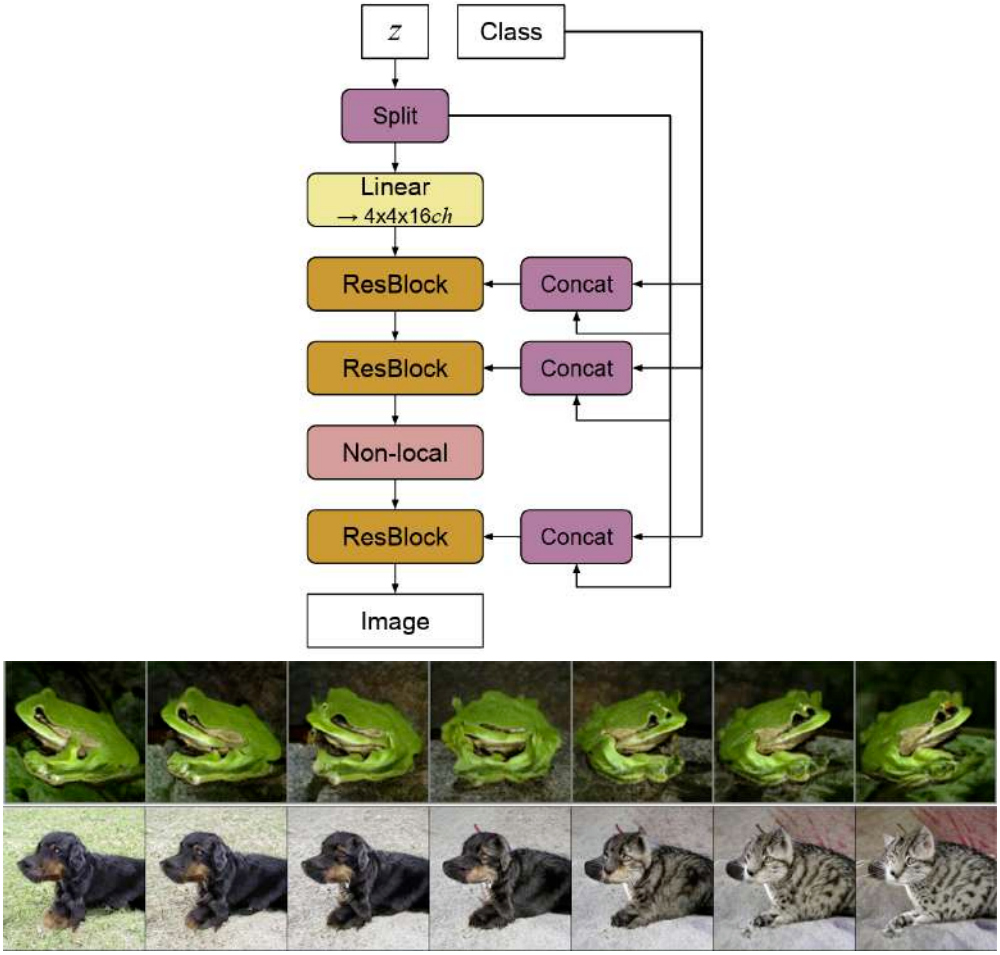

Figure 1.15: Structure of BigGAN model at the top. The model receives latent z and class variable to generate images. Examples of latent interpolation for BigGAN model at the bottom. The first row demonstrates the effect of the z variable on rotation through latent interpolation. The second row shows that the pose remains the same as the image category changes, demonstrated through latent interpolation of the class variable. (Adapted from Brock et al. 2018 , and Voynov and Babenko 2020)[18,247].

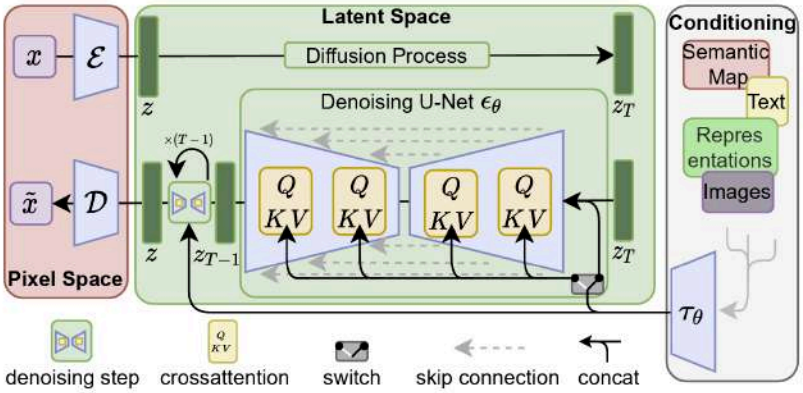

Figure 1.16: Representation of the latent diffusion model. The model applies diffusion iteratively to the data, transforming it into noise. Then, denoising is applied using neural network models to generate images that are similar to the original data. (Figure from Rombach et al. 2022)[202].

data correctly, as shown in Equation 1.2[75]. Traditional GANs generate data only from noise sampled from an unconditional distribution. In contrast, conditional GANs receive additional input, such as class or map, to produce more specific or controlled generation[68]. For instance, GANs trained on the ImageNet dataset are typically trained conditionally and receive category labels as inputs. The BigGAN model was one of the initial models trained in this manner (illustrated in Figure 1.15)[18]. BigGAN was designed to generate high-resolution images. The IC-GAN model, which was used in the first study (Chapter 2), is a successor of BigGAN. In this model, conditioning is achieved through instance vectors instead of category labels[26] (more details in section 2.2.3.1).

Latent Diffusion Models (LDMs) are generative models that iteratively refine data to generate high-fidelity samples from a noise sample[202] (represented in Figure 1.16). LDM generates samples from a dataset by transforming a noise sample from a normal distribution (Gaussian noise). Unlike naive diffusion models,

LDMs encode high-dimensional data, such as images, into a lower-dimensional latent space using an encoder. The diffusion process applied in the latent space consists of two operations: forward diffusion and reverse diffusion or denoising. Forward diffusion is the process of gradually adding noise to the data until only the noise remains[202]. Reverse diffusion is the process of gradually removing noise through denoising, which is an operation learned by the model[202]. The model learns how to denoise by predicting the added noise in each forward diffusion step. When generating new samples, the LDMs apply reverse diffusion by predicting the added noise for a certain number of steps and subtracting it from the latent variable for each step[6]. LDMs have become popular due to their success in generating high-quality samples that match the dataset distribution. This success has been demonstrated in text-to-image generation models such as Stable Diffusion. The Versatile Diffusion model used in the second study (Chapter 3) is a variation of the stable diffusion model with multimodal inputs and pathways, including both images and text[255] (more details in section 3.2.3.3).

1.6 Using Deep Learning for Neuroscience Studies

As stated in section 1.5, deep learning models have their roots in computational neuroscience models, and many neural network models draw inspiration from neuroscience. Despite criticism from some researchers regarding the neuro-inspired connectionist paradigm and suggestions for a more engineering approach[112], or symbolic approach[145], neuroscience and AI research continue to collaborate. Neuroscientific insights are used to establish robust models, while AI models are

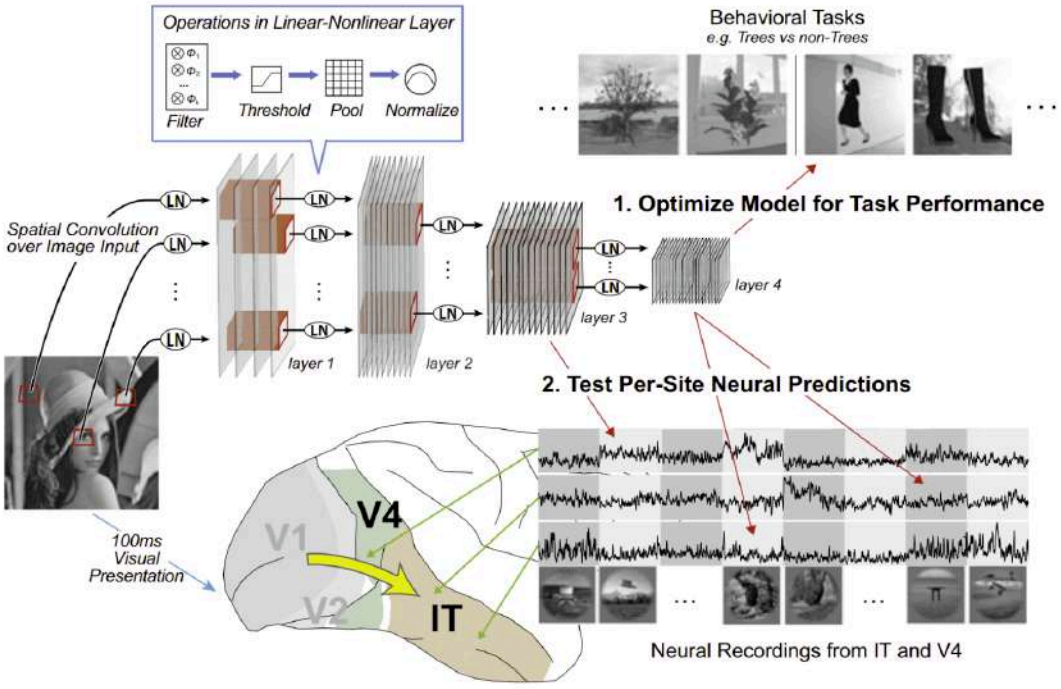

Figure 1.17: Yamins et al. shows higher layers of a CNN model (which was trained on an object recognition task) are capable of predicting neural recording from V4 and IT (Adapted from Yamins et al. 2014)[256].

employed for neural decoding, brain imaging analysis, and behavioral analysis. Following the success of AlexNet in the ImageNet object recognition competition, cognitive computational neuroscience researchers began analyzing the representational similarities and differences between deep learning models and the brain. Mathematical models for low-level visual processing had already been established[103,1], but until the advent of deep learning models, there were no models that adequately captured high-level features. DNNs have demonstrated hierarchical feature learning when trained for specific tasks, such as object recognition. Early analyses have shown that DNNs can predict neural activity in mid-visual areas like V4[256,82] (presented in Figure 1.17). Similar analyses are also conducted for the inferior temporal (IT) cortex, which is recognized for its high-level visual processing[120]. The analysis of visual cortical activity prediction (encoding) is repeated for different types of neuroimaging data, including electrophysiology[20], fMRI[60,252,59], MEG[36] and EEG[76]. Contrary to the prediction of cortical activity, deep neural networks (DNNs) are also utilized to provide feature spaces for decoding certain aspects of visual stimuli, such as category, or even the stimuli themselves, such as visual reconstruction[252,100]. Our research focuses on the use of AI and deep learning in neuroscience, specifically for neural decoding and visual reconstruction.

1.7 Visual Reconstruction using Deep Learning

Researchers started using deep generative models to reconstruct images from fMRI signals after witnessing the representational modeling capabilities of DNNs

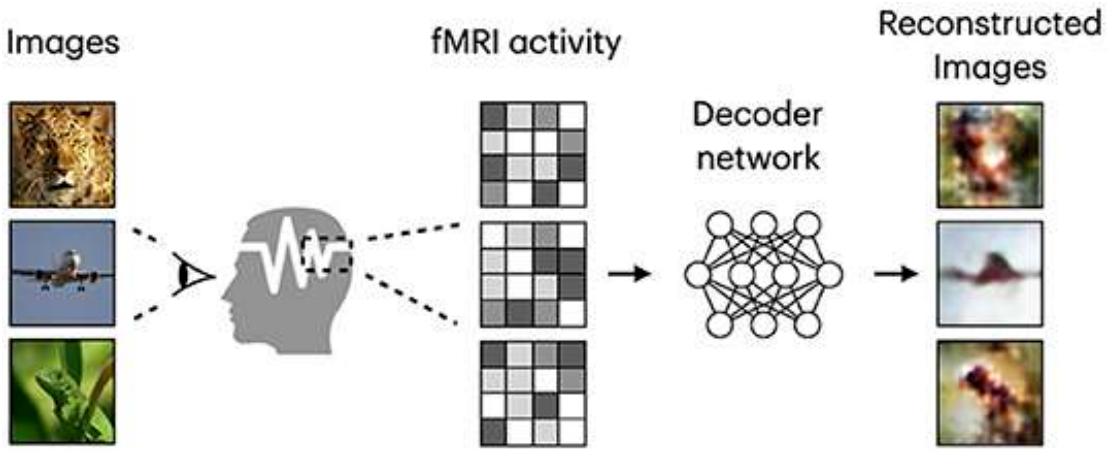

Figure 1.18: General approach for visual reconstruction using deep learning methods. The method involves presenting images to subjects while recording their brain activity via fMRI. A decoder model is then trained using these fMRI-image pairs, and test reconstructions are generated using fMRI activity in the test set (Figure from Rakhimberdina et al. 2021)[188].

for neuroscience. Although the details of approaches may vary depending on the methods used, the general approach for visual reconstruction using deep learning models is shown in Figure 1.18. Du et al. conducted one of the earliest studies in this area. They utilized various deep learning models, such as deep canonically correlated autoencoders, deconvolutional neural networks, and deep generative multiview models, to reconstruct geometric shapes, alphabet letters, handwritten digits, and characters from fMRI activity patterns[57]. Researchers have shifted their focus towards reconstructing more natural images, such as faces, by utilizing deep generative models that have been trained on face images[83,245,42] (presented in Figure 1.19). Subsequently, researchers began collecting fMRI datasets of natural images, which are discussed below. These images are more complex and contain various categories, such as those found in the ImageNet[48] and COCO[138] datasets.

Figure 1.19: Face reconstruction results of VAE-GAN model and PCA (baseline) (Adapted from VanRullen and Reddy 2019)[245].

1.7.1 Natural Image-fMRI Datasets

Publicly available fMRI datasets are essential for advancing neural decoding research. They enable scientists worldwide to collaboratively study the functional organization of the brain and assess the performance of methods through benchmarking. The main studies of the thesis utilized two public fMRI datasets. Chapter 2 presents the study that used the Generic Object Decoding dataset prepared by Kamitani Lab. This dataset includes fMRI recordings of 5 subjects while images from the ImageNet dataset are presented. In Chapter 3 of the thesis, we utilized the Natural Scenes Dataset, which was prepared under the supervision of Kendrick Kay and Thomas Naselaris. The dataset includes fMRI recordings of 8 subjects who were presented with images from the COCO dataset. We used data from the 4 subjects who completed all sessions.

The Generic Object Decoding (GOD) dataset contains 1200 images from 150 object categories in the training set (8 images per category) and 50 images from

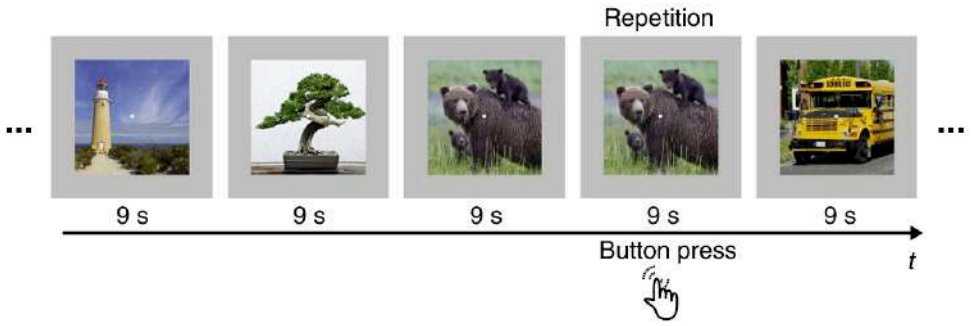

Figure 1.20: Demonstration of Generic Object Decoding dataset, where subjects are instructed to press a button whenever they see the same stimuli consecutively (one-back test) during the presentation experiment (Adapted from Horikawa et al. 2017[100]).

50 object categories in the test set. The categories in the training and test sets are distinct from each other, and all images are retrieved from the ImageNet dataset[100]. During the study, fMRI data were collected from five subjects while they viewed a series of images in a 3 Tesla scanner. Each stimulus image was presented for 9 seconds, flashed at 2 Hz, and displayed at 12x12° of visual angle. A fixation point was located at the center of the images (presented in Figure 1.20). Subjects performed a one-back test by pressing a button when they saw an image twice in a row. This helped them focus their attention. During the testing phase, each image was repeated 35 times, whereas during the training phase, each image was presented only once. In addition to image presentation sessions, GOD also included retinotopy and localizer experiments to extract regions of interest (ROIs). Retinotopic regions, including V1, V2, V3, and V4, were identified for each subject through retinotopy experiments. Additionally, regions such as the Lateral Occipital Complex (LOC), Fusiform Face Area (FFA), and Parahippocampal Place Area (PPA) were extracted for each subject using functional localizer

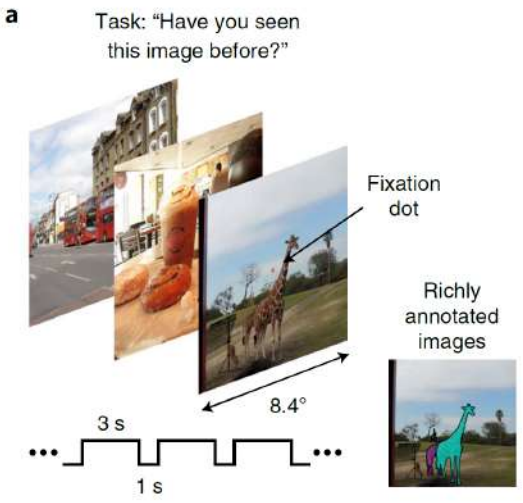

Figure 1.21: Demonstration of stimuli presentation experiment in Natural Scenes Dataset. (Adapted from Allen et al. 2022[2]).

experiments.

The Natural Scenes Dataset (NSD) is a large-scale dataset collected using a 7 Tesla fMRI scanner for 8 subjects over the course of a year, across 30-40 sessions[2]. The training dataset consists of 8,859 images, while the test set has 982 images for each subject from the COCO dataset. Unlike GOD, the training set images are unique to each subject, while the test set images are shared among all subjects. Only four out of eight subjects completed all sessions (sub1, sub2, sub5, and sub7). The participants viewed stimulus images for 3 seconds each, with a 1-second interval between them. The images were displayed at a visual angle of 8.4x8.4° and were preceded by a central fixation point. All images were shown up to three times, resulting in 24,980 training trials and 2,770 fMRI trials for the test set. During the sessions, participants are asked if they have seen the stimulus images before, a task designed to help them maintain their attention (demon-

strated in Figure 1.21). Prior to the planned 40 sessions of image presentation, participants are gathered for population receptive field (pRF) and functional localization (fLoc) experiments. These experiments are used to rank participants in terms of performance and to extract ROIs. pRF experiments define regions in visual areas based on hierarchy (V1, V2, V3, and hV4) and eccentricity ($0°<e<0.5°$, $0.5°<e<1°$, $1°<e<2°$, $2°<e<4°$, and $4°<e$, where "e" represents eccentricity) for each subject. fLoc experiments are conducted to define category-selective areas such as Face-ROI (OFA, FFA-1, FFA-2, mTL-faces and aTL-faces), Word-ROI (OWFA, VWFA-1, VWFA-2, mfs-words and mTL-words), Place-ROI (OPA, PPA and RSC) and Body-ROI (EBA, FBA-1, FBA-2 and mTL-bodies) for each subject. Preprocessing included temporal interpolation for slice time correction and spatial interpolation for head motion and spatial distortion correction. The authors employed a generalized linear model (GLM) to obtain approximate single-trial beta weights. They also incorporated a hemodynamic response function and applied additional procedures such as GLMDenoise and ridge regression. For our study, we used preprocessed fMRI signals masked with NSDGeneral ROI, manually drawn on fsaverage and covering voxels responsive to the NSD experiment in the posterior cortex, with a resolution of 1.8 mm.

1.7.2 Natural Image Reconstruction Models using Deep Learning

The Kamitani Lab prepared two datasets that have become benchmarks for natural image reconstruction: the Generic Object Decoding (GOD) dataset[100] (described above) and the Deep Image Reconstruction (DIR) dataset[217] (shown in

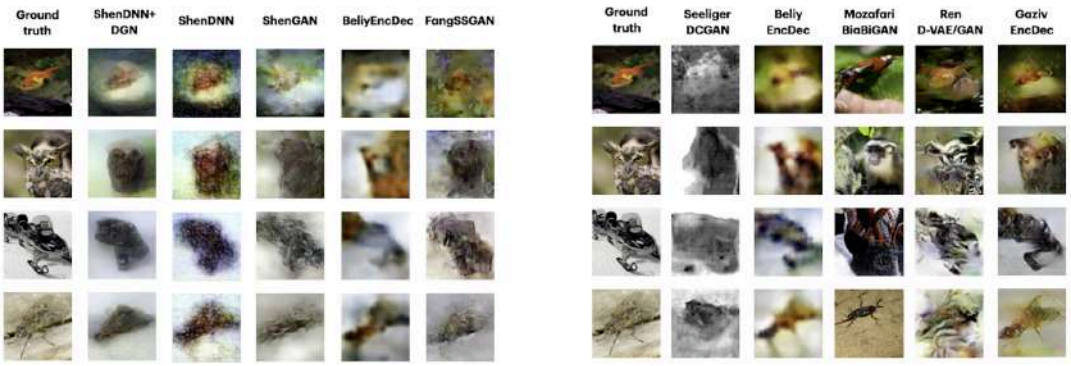

Figure 1.22: Reconstructed images from natural image reconstruction models for the Deep Image Reconstruction (DIR) dataset (left) and the Generic Object Decoding (GOD) dataset (right) prior to our first study, the IC-GAN model. The models for the DIR dataset, on the left, are from Shen et al.[217], Shen et al.[217], Shen et al.[216], Beliy et al.[14], and Fang et al.[64] respectively. The models for the GOD dataset, on the right, are from Seeliger et al.[214], Beliy et al.[14], Mozafari et al.[155], Ren et al.[198], and Gaziv et al.[73] respectively. (Adapted from Rakhimberdina et al. 2021)[188].

Figure 1.23). Both datasets use the same set of images from the ImageNet dataset. However, the number of subjects and fMRI repetitions per image differs between the datasets. Some studies also utilized the Visual Imaging 1 (vim-1) dataset, which contains grayscale natural images with a circular mask[119]. Seeliger et al. used the Deep Convolutional GAN (DCGAN) model to reconstruct images from fMRI patterns for the vim-1 and GOD datasets, as well as a dataset consisting of handwritten characters[214]. St-Yves and Naselaris trained a conditional energy-based GAN (EBGAN) to reconstruct images from the vim-1 dataset[229]. Shen et al. took a different approach and iteratively optimized the reconstructed image with a deep generative network utilizing the decoded VGG-19 features[217]. Later, an adversarial loss component was added to the model, making it capable of end-to-end training[216]. Beliy et al. developed an Encoder-Decoder model to train between stimuli images and fMRI activity. They also employed a self-supervised approach

by stacking the encoder and decoder back-to-back and training the network with an additional cycle consistency loss[262] to mitigate the problem of scarce labeled data[14]. In a follow-up study, Gaziv et al. introduced an additional loss component utilizing perceptual similarity loss[261], which was calculated based on the extracted features from reconstructed and ground truth images[73]. Qiao et al. introduced a Bayesian Visual Reconstruction Model based on BigGAN (GAN-BVRM) to reconstruct images from the vim-1 dataset[183]. Fang et al. developed a new framework in which they decoded shape and semantic representations separately from various visual areas and then combined them using GAN[64]. Ren et al. designed a dual VAE-GAN network framework for visual reconstruction. The framework uses visually-guided cognitive representation and adversarial learning to bridge the domain gap between fMRI signals and visual images. This is achieved by gradually distilling knowledge between encoders[198]. Prior to our IC-GAN study, the aforementioned models were able to capture certain aspects of the stimuli, such as position and layout. However, they were unable to generate images that accurately depict high-level features and appear realistic, with the exception of Mozafari et al.[155] (presented in Figure 1.22).

Mozafari et al. used the BigBiGAN model (an unconditional BigGAN with an encoder) to reconstruct images in the GOD dataset as semantically more similar to groundtruth images, shifting the focus of visual reconstruction to a more semantic-oriented approach[155]. The approach by Mozafari et al. was an inspiration for our Instance-Conditioned GAN (IC-GAN) approach, where we developed a reconstruction framework that produces images that are semantically similar

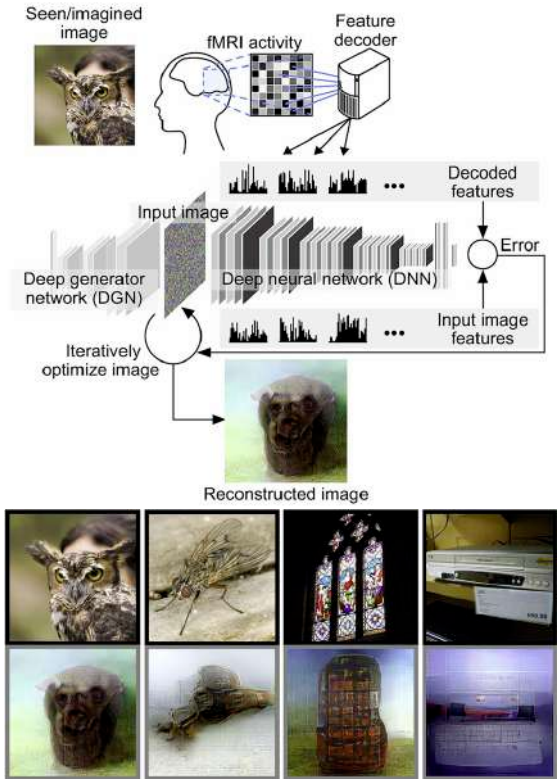

Figure 1.23: The reconstruction framework proposed by Shen et al. The image is optimized by a deep generator network using error signals obtained from decoded features of fMRI activity. The ground truth images and reconstructions of four test samples, which are presented below (Adapted from Shen et al 2019)[217].

to the ground truthwhile retaining some of the low-level structure[170]. These approaches influenced later studies, such as those by Ferrante et al.[65] and Chen et al.[31].

Deep generative models have shown significant improvement in reconstructing both the semantic and shape aspects of images from the GOD and DIR datasets. However, it is important to note that these datasets consist of single-object-centered images, specifically ImageNet images with one category of object at the center. Therefore, the next challenge is to reconstruct complex scenes with multiple objects and actions. The Natural Scenes Dataset (NSD) was curated by Allen et al. from COCO images. During the viewing of these images, fMRI signals were recorded from 8 subjects[2]. After this dataset was made public, researchers studying visual reconstruction began using it as the new benchmark. Lin et al. were the first to use NSD for visual reconstruction. They utilized a framework that adapted an unconditional model (StyleGAN2) for text-to-image generation, called Lafite[137]. They used CLIP text embeddings of image captions instead of extracted image features due to the model's design. Gu et al. were inspired by our IC-GAN paper and employed the IC-GAN model, which allowed for the utilization of image features. They used a surface-based convolutional network to process fMRI signals for reconstructing images of the NSD dataset[80]. Takagi et al. conducted the first study utilizing the Stable Diffusion model. However, similar to Lin et al., they only incorporated text embeddings of the CLIP model for semantic conditioning[234]. After these studies, we developed the two-stage Brain-Diffuser framework, using the VDVAE model for low-level features and initial image recon-

struction, and the Versatile Diffusion model, a multimodal model that uses both image and text features from the CLIP model, for high-level features and final image reconstruction. Brain-Diffuser outperformed all previous models in both high-level and low-level metrics[171].

1.8 Outline of the Thesis:

The thesis presents two methods for reconstructing and combining high-level (semantic) and low-level (shape and layout) features from fMRI patterns to achieve high-fidelity reconstructions using deep generative models.

In Chapter 2, we propose a natural image reconstruction framework using an Instance-Conditioned GAN model that performs accurate semantic reconstruction and preserves low-level details from fMRI patterns on the Generic Object Decoding dataset.

In Chapter 3, we introduce the 'Brain-Diffuser' framework, a two-stage scene reconstruction method that uses latent diffusion models to reconstruct high-complexity images from fMRI signals on the Natural Scenes Dataset.

In Chapter 4, we first extend our discussion for the studies in Chapters 2 and 3. Later, practical applications and ethical implications of our study and neural decoding research in general are then discussed, followed by a summary and closing thoughts.

Chapter 2

Reconstruction of perceived images from fMRI patterns and semantic brain exploration using instance-conditioned GANs

This chapter proposes a framework that utilizes the IC-GAN model for perceived image reconstruction. Reconstruction results and qualitative metrics for the Generic Object Decoding dataset are presented. The semantic information encoded in several brain regions provided with the dataset is also analyzed.

2.1 Prologue to the main article :

In this section, I will briefly introduce the story behind the project. During my bachelors studies at Istanbul Technical University, I explored studies related to deep learning and neuroscience. I was particularly intrigued by the studies on

reconstructing stimuli from brain signals. The first studies I encountered were 'Deep Image Reconstruction from Human Brain Activity' by Kamitani Lab, 'Reconstructing Faces from fMRI Patterns Using Deep Generative Neural Networks' by Rufin VanRullen and Leila Reddy, and 'Generative Adversarial Networks for Reconstructing Natural Images from Brain Activity' by Van Gerven Lab. All of these studies were remarkable and demonstrated the possibility of reconstructing images with high fidelity using deep learning models. In the last semester of my bachelor's degree, I worked on visual image reconstruction from fMRI signals as my graduation project. Due to the limited expertise in fMRI processing in our Artificial Intelligence and Computer Vision Lab, I kept the project simple. I worked on the "BRAINS" dataset published by Schoenmakers et al.[211] and devised simple convolutional autoencoder and GAN models to reconstruct handwritten letters. The project was successful, but it did not satisfy my desire to create better reconstruction frameworks for natural images. At the end of the semester, I did not consider applying to universities abroad for a master's degree. Instead, I continued my studies in computer vision and deep learning at Istanbul Technical University to improve my skills in AI and deep learning. During the final year of my master's program, I developed a keen interest in natural image reconstruction. Pursuing this field required me to study abroad. I expressed my enthusiasm for the topic and my desire to work with Rufin during my PhD in an email. Rufin responded promptly with a positive reply. My PhD journey began during the COVID-19 pandemic, which forced me to work remotely for my studies. During my first year, I designed and experimented with various models for natural

image reconstruction. Despite obtaining many promising results, I was unable to find a suitable candidate for a state-of-the-art model. In my second year, I was finally able to go to Toulouse and begin working in the lab. While there, I had the opportunity to meet with many students and researchers, which greatly motivated my studies. During my exploration and experimentation with various models, I came across the Instance-Conditioned GAN model. Upon reviewing its generation framework and results, I recognized its potential as a state-of-the-art candidate for natural image reconstruction. When we began this study, visual reconstruction studies typically focused on reconstructing shape and layout information resulting in silhouettes. However, Mozafari et al.'s BigBiGAN approach, which still had limitations, was an exception. With the IC-GAN approach, we wanted to integrate semantic information and low-level information to generate realistic images, as opposed to non-realistic silhouettes. The following article presents the results of these studies.

2.2 Main article :

2.2.1 Abstract

Reconstructing perceived natural images from fMRI signals is one of the most engaging topics of neural decoding research. Prior studies had success in reconstructing either the low-level image features or the semantic/high-level aspects, but rarely both. In this study, we utilized an Instance-Conditioned GAN (IC-GAN) model to reconstruct images from fMRI patterns with both accurate semantic attributes and preserved low-level details. The IC-GAN model takes as

input a 119-dim noise vector and a 2048-dim instance feature vector extracted from a target image via a self-supervised learning model (SwAV ResNet-50); these instance features act as a conditioning for IC-GAN image generation, while the noise vector introduces variability between samples. We trained ridge regression models to predict instance features, noise vectors, and dense vectors (the output of the first dense layer of the IC-GAN generator) of stimuli from corresponding fMRI patterns. Then, we used the IC-GAN generator to reconstruct novel test images based on these fMRI-predicted variables. The generated images presented state-of-the-art results in terms of capturing the semantic attributes of the original test images while remaining relatively faithful to low-level image details. Finally, we use the learned regression model and the IC-GAN generator to systematically explore and visualize the semantic features that maximally drive each of several regions-of-interest in the human brain.

2.2.2 Introduction

Understanding the brain and cognition has always been one of the fundamental research areas of science. One of the ways researchers approach this task is by establishing neural encoding and decoding methods. New ways to decode information from brain signals have emerged with recent developments in modeling and computation.

In vision research, many studies have used statistical methods and machine learning to decode specific information like position[237] or orientation[113,91], to classify image categories[90,40], to retrieve the closest images from a candidate set[119],

or even to reconstruct images with low-complexity like basic shapes and structures[154].

With the emergence of deep learning, and in particular advanced deep generative models, reconstructing more complex images like handwritten digits[211], faces[245], and natural scenes[217] has become possible. These deep generative models include variational auto-encoders (VAEs), generative adversarial networks (GANs), and many variants and hybrids of both. Although many studies have used these models, they typically managed to reconstruct either low-level or high-level features of the images, but rarely both at the same time.

Here, we propose a method to reconstruct natural images from fMRI activation patterns with both accurate semantic attributes and relatively preserved low-level details, using an Instance-Conditioned GAN (IC-GAN) – a recent generative model[26] inspired by the success of self-supervised feature learning[110]. In our framework, we first extract latent representations for a set of training images (see Figure 2.1): high-level attributes of the images, called "instance features" in IC-GAN, are computed with a single forward pass through the SwAV ResNet-50 model; lower-level aspects of the image (e.g. reflecting the size, position or orientation of an object, details of the background, etc.) are obtained via a two-stage optimization of the IC-GAN "noise" and "dense" latent vectors (inspired by the method of Pividori et al.[176]). Next, we train three ridge regression models to predict these latent image representations from the corresponding fMRI patterns, recorded while human subjects viewed the same training images (Figure 2.2, Step 1). Finally, for each image in the test set, we predict the instance feature, noise

vector, and dense vector from fMRI data (using the previously learned regression models), and then reconstruct an estimate of the image using IC-GAN's generator (Figure 2.2, Step 2 and 3). The code of this paper can be found in the official GitHub repository*.

Our method establishes a new state-of-the-art performance for capturing the semantic attributes of the images, while preserving a reasonable amount of low-level details. We present both qualitative and quantitative results, and a comparison with previous methods to support our claims. We also take advantage of our brain-based image reconstruction system to explore and visualize the semantic image attributes encoded in various brain regions-of-interest (ROIs), and discuss how these findings align with neuroscientific evidence.

2.2.3 Materials and Methods

2.2.3.1 Instance-Conditioned GAN

We utilized an Instance-Conditioned GAN (IC-GAN) model, pretrained for natural image generation on the ImageNet dataset[48]. IC-GAN can be considered as a generic framework rather than a single model, because it can be applied to different GAN backbones, e.g. StyleGAN[118] or BigGAN[18]. In the usual conditional GAN setting[150], class labels are provided along with noise vectors sampled from a normal distribution to generate images. Images belonging to that specific class are labeled as "real", and generated images from the generator are labeled as "fake". Both the generator and discriminator are trained with these images and labels in

*https://github.com/ozcelikfu/IC-GAN_fMRI_Reconstruction

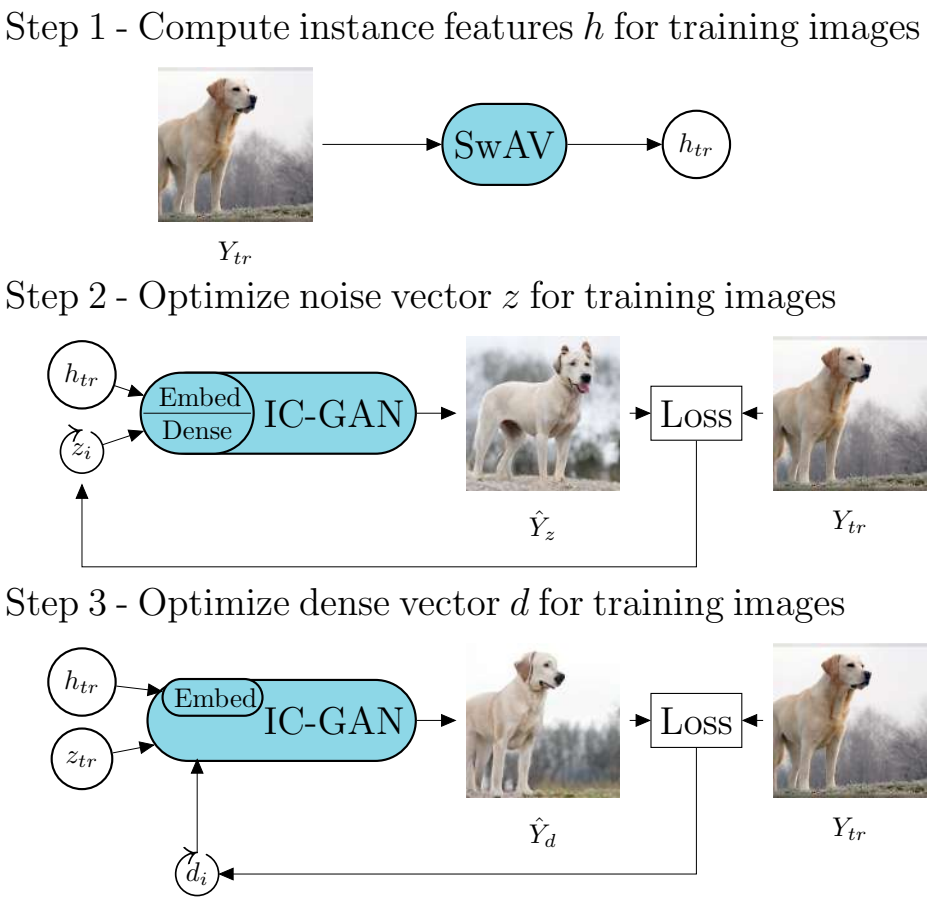

Figure 2.1: Extraction of the latent variables (h_{tr}, z_{tr} and d_{tr}) for each training image (Y_{tr}). Step 1: Instance features of training images (h_{tr}) are extracted using SwAV ResNet-50. This 2048-dim instance feature vector (h_{tr}) captures the semantic attributes of the image. Step 2: In addition to the instance feature vector, the IC-GAN also requires a noise vector (z_i) as input, which encodes lower-level properties of the image (e.g., pose, orientation, background etc.). While providing h_{tr} obtained from Step 1 to the IC-GAN's generator, we optimize the noise vector (z_i) to generate the closest image ($\hat{Y}_z$) to the groundtruth image (Y_{tr}). The resulting optimized noise vector is z_{tr}. Step 3: To further improve image reconstruction so as to better match the more detailed spatial structure of the training image, we apply another optimization stage, in which we optimize the dense layer vectors of IC-GAN itself. To achieve, this, we pass the first 17 dimensions of z_{tr} to the dense layer of the IC-GAN's generator and obtain initial dense vectors (d_0). While keeping both h_{tr} and the remaining 102 dimensions of z_{tr} fixed, we optimize the dense vector d_i to generate the closest image ($\hat{Y}_d$) to the groundtruth image (Y_{tr}). d_{tr} is the resulting optimized dense vector.

an adversarial learning framework. In the instance-conditional setting, instead of giving a class label, instance features that capture the semantic attributes of a given image are extracted (via a pre-trained feature extractor) and provided to the generator as conditioning, alongside a sampled noise vector. For training, IC-GAN selects k images in the neighborhood of the conditioning image (according to the feature extractor); these images are labeled as real, while generated images are considered as fake images to train both the generator and discriminator.

For the instance feature extraction, IC-GAN models use the SwAV (Swapping Assignments between Views) architecture[24] with a ResNet-50[92] backbone. SwAV is a self-supervised learning model which means that it does not require hand-crafted labels from humans. Similar to contrastive learning methods[110], SwAV minimizes the distance in feature space between representations of two transformed images (coming from the same original image).

It is possible to train the IC-GAN framework with different feature extractors, as long as they provide rich feature representations. However, using features from self-supervised learning models (e.g., SwAV) is better suited to the problem of neural decoding and natural image reconstruction. Indeed, many recent studies show that representations gathered from self-supervised learning models present more similarity to brain representations than other learning methods[125,263].

The specific IC-GAN model we used here relies on a BigGAN[18] architecture with 7 layers. It generates $256 \times 256 \times 3$ images from a 2048-dim (dimensional) instance feature vector extracted from SwAV ResNet-50 and a 119-dim noise vector sampled from a normal distribution. The 2048-dim instance features are given to

an embedding layer and thus reduced to 512-dim embedded vectors. The 119-dim noise vector, which encodes lower-level properties of the image (e.g. pose, size, orientation of the object), is split into seven hierarchical levels, each with 17 dimensions. The first 17-dim level is directly given to the first dense layer of the IC-GAN generator. The remaining six hierarchical levels are concatenated with the embedded instance vector to be fed to the generator in each of the six BigGAN residual blocks.

Overall, the purpose of IC-GAN is to generate, from one conditioning image, new and diverse image instances that share semantic attributes (as captured by SwAV instance features), but differ in low-level properties (e.g. object position, size, orientation, background details). The diversity of low-level properties is determined by randomly sampled "noise" vectors (and by the "dense" vectors directly derived from them). However, for the purpose of fMRI-based image reconstruction, both high-level and low-level properties must be specified. Therefore, rather than randomly sampling noise vectors, we computed a specific noise vector (and the associated dense vector) for each training image in the dataset, as detailed below.

2.2.3.2 Extracting Latent Variables from Training Stimuli

We illustrate the computation of latent variables in Figure 2.1. We first extracted a 2048-dim instance feature vector for each training image in our dataset (see dataset details below) by presenting it to a SwAV ResNet-50 feature extractor. We then provided these instance features to the IC-GAN generator, and

optimized the 119-dim noise vector for the same image using the covariance matrix adaptation evolution strategy (CMA-ES)[85]. We used this method because we empirically observed that global optimization strategies worked better than local optimization strategies (like gradient-based methods) for the noise vector. The loss function for this optimization was the distance between the generated image and the original training image in Layer-4 of SwAV ResNet-50; this representation level, hierarchically lower than the instance feature level, encodes more spatially structured information.

Finally, to further match the more detailed spatial structure of the original image, we applied one more optimization stage. Inspired by the two-stage inversion method of Pividori et al.[176], we provided the first 17 dimensions of the previously optimized noise vector to the first dense layer of the IC-GAN, resulting in a $1536 \times 4 \times 4$-dim dense vector. While the instance features and the remaining 102 dimensions of the noise vector were kept fixed, we optimized these dense vectors with the RMSProp optimizer. For this second-stage optimization, the previous loss (SwAV ResNet-50 Layer-4 feature distance) was combined with a Learned Perceptual Image Patch Similarity (LPIPS)[261] loss gathered from a pretrained VGG16 model[220] and a pixel (MSE) loss from 64×64 resized images.

2.2.3.3 Generic Object Decoding Dataset

In this study, we used previously published fMRI recordings of five human subjects presented with images from the ImageNet dataset[100]. The dataset contains training and testing image perception sessions where subjects looked at 1200

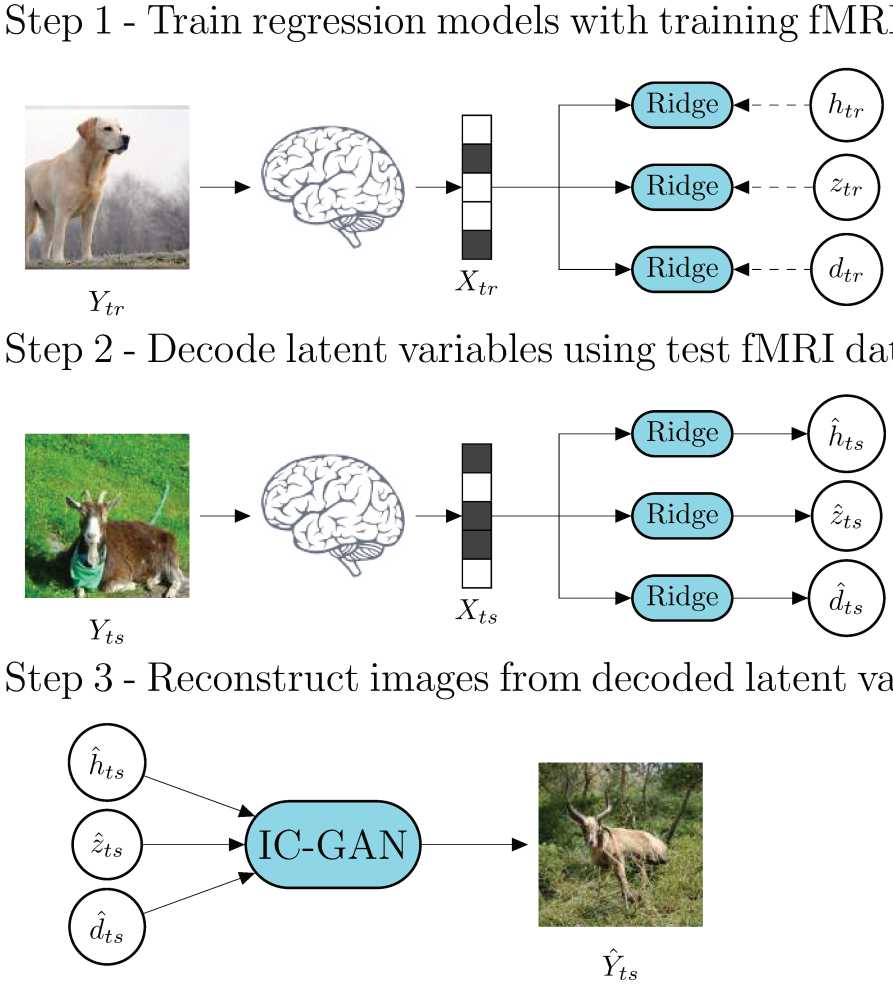

Figure 2.2: Decoding latent variables from fMRI patterns and reconstructing images from decoded variables. Step 1: Having obtained the instance features (h_{tr}), noise vectors (z_{tr}) and dense vectors (d_{tr}) of training images (Y_{tr}) as described in Figure 2.1, we train three ridge regression models to map fMRI patterns of the training set (X_{tr}) to these latent variables. Step 2: Using these trained regression models, we decode latent variables of the test set ($\hat{h}_{ts}$,$\hat{z}_{ts}$,$\hat{d}_{ts}$) from test fMRI patterns (X_{ts}). Step 3: We pass the decoded latent variables to the IC-GAN Generator to obtain reconstructed images ($\hat{Y}_{ts}$)

training samples drawn from 150 categories (8 samples each) and 50 testing samples chosen from 50 categories (1 sample each), respectively. Training and testing categories were chosen independently and were non-overlapping. Each training image was presented only once, while testing images were repeated 35 times during the whole experiment. All fMRI runs followed a similar design: fixation (33s), 50 image presentations (9s per image flashing at 2Hz), fixation (6s). Moreover, subjects were also asked to perform a one-back task by pressing a button whenever the same image was presented two times in a row (five such events occurred per run).

The fMRI data were pre-processed for each subject by three-dimensional motion correction followed by coregistration to the high-resolution anatomical image. Then, the brain representation of each image was calculated by averaging the percent signal change values of each voxel over the 9-s presentation window. Additionally, the dataset provides functional regions of interest (ROIs) that cover the entire visual cortex, including V1-V4, the fusiform face area (FFA), parahippocampal place area (PPA), and lateral occipital complex (LOC). The pre-processed data is available to download at brainliner.jp[†].

2.2.3.4 fMRI Decoding and Image Reconstruction

Details of fMRI decoding and image reconstruction are depicted in Figure 2.2. The procedure involves two separate stages for training and testing the brain decoding system of each subject.

[†] http://brainliner.jp/data/brainliner/Generic_Object_Decoding

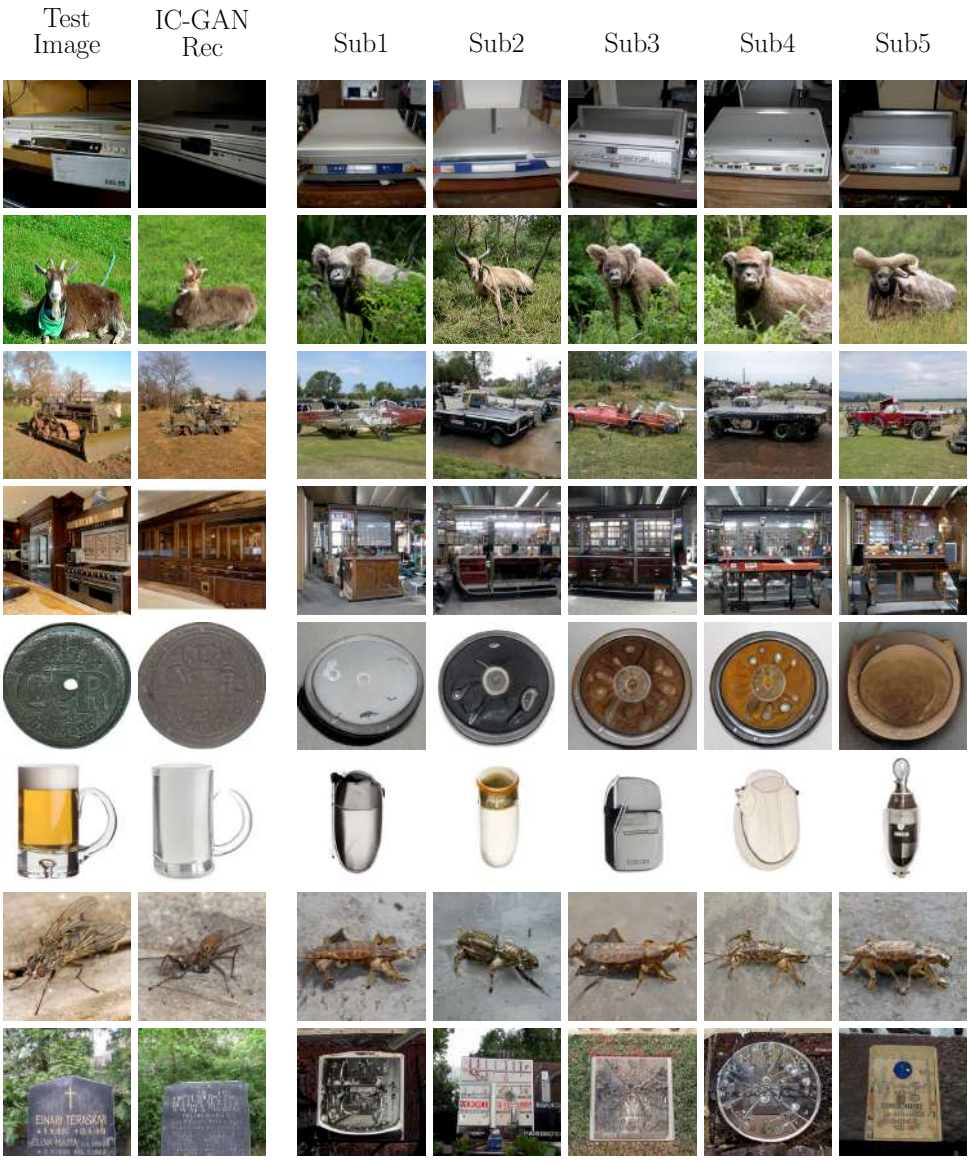

Figure 2.3: fMRI Reconstructions by the IC-GAN model for all subjects. The first column is the groundtruth test image, whereas the second column is the reconstructed image by IC-GAN using extracted latent variables. The following five columns demonstrate the equivalent reconstructions using fMRI-decoded latent variables for each subject. fMRI reconstructions are generally consistent with the groundtruth images in terms of semantic attributes, while they preserve the low-level details to a certain degree.

First, we trained three separate ridge regression models to predict the latent variables (instance features; noise vectors; dense vectors) for each of the 1200 training images based on the corresponding fMRI patterns. Since both the fMRI data and the latent variables are high-dimensional, we applied L_2 regularization on the regression weights during training.

At test time, we averaged the 35 repetitions of fMRI signals corresponding to each test stimulus. Next, we used the previously trained regression models to predict the instance features, noise vectors, and dense vectors from these averaged fMRI signals. Finally, we used these predicted latent variables to generate image reconstructions using the IC-GAN generator.

2.2.4 Results and Analyses

2.2.4.1 Image Reconstruction Results

Examples of image reconstructions produced by our method are displayed in Figure 2.3. First of all, it is important to examine IC-GAN reconstructions (second column) based on the optimized "ground-truth" latent vectors (derived as detailed in Figure 2.1): we can see that IC-GAN can successfully reconstruct the semantic attributes of the test images; however, it often misses some visual details, like parts of the vehicle (third row), liquid in the glass (fifth row), or the precise text in the gravestone (eighth row). These reconstructions help us understand how the IC-GAN generator would behave if we perfectly decoded latent variables from fMRI patterns, i.e. they serve as an upper bound on the expected reconstruction quality.

When we inspect actual fMRI reconstructions for the five subjects (third to seventh columns), our first observation is that reconstructions look like natural images. Furthermore, they are consistent across subjects. Again, these reconstructions capture some of the semantic attributes, while also missing specific aspects of the test images. For example, the system generates images of horned animals for the goat image (second row), but their species are not clearly identifiable. For the token image (fifth row), round objects are reconstructed, but not with the right texture. For the gravestone image (last row), similar square-shaped objects with text and symbols are generated, but most of them would not qualify as a gravestone. Overall, our method appears to reconstruct semantic attributes with slight but significant variations in details.

How does it compare to previously proposed methods? In Figure 2.4, we present image reconstructions using alternative methods proposed in five other studies, together with our results for comparison[‡]. From these reconstructions, we can see that many methods capture low-level details rather than high-level ones; as a result, many of the reconstructions do not look natural. A notable exception is the study of Mozafari et al., based on the BigBiGAN architecture[155], in which reconstructions often capture high-level properties and are more naturalistic. Even this method, however, does not correctly reconstruct semantic details for some of the images; furthermore, it misses many of the low-level details. Among the other studies, Ren et al.[198] succeed in reconstructing colors and textures better than other methods, while Gaziv et al.[73] give sharper object edges.

[‡]We selected these seven images because it was the only common set of reconstruction exemplars presented across all of the considered studies.

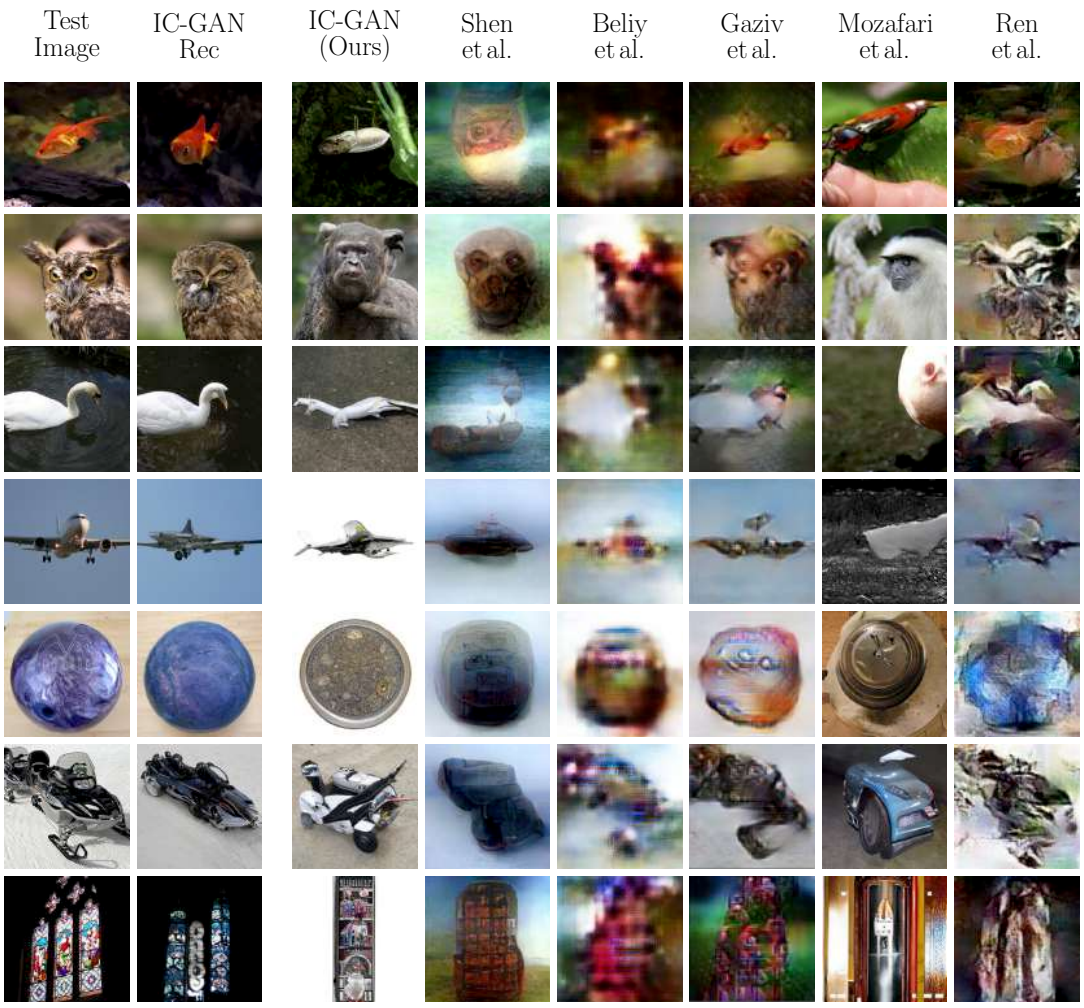

Figure 2.4: Comparison of fMRI reconstructions for several methods. The first column is the groundtruth image, the second column is the reconstructed image with IC-GAN using extracted latent variables. Columns three to eight present fMRI reconstructions from IC-GAN (Ours), Shen et al.[217], Beliy et al.[14], Gaziv et al.[73], Mozafari et al.[155], and Ren et al.[198], respectively. fMRI reconstructions by the IC-GAN method demonstrate more naturalistic-looking images with accurate semantic attributes, while preserving some low-level details (e.g. object position, size or orientation).

Our method generates realistic-looking image reconstructions with appropriate semantic features, while preserving the low-level aspects to a certain degree.

Table 2.1: Quantitative comparison of image reconstructions. For each measure, the best value is in bold. (For Pix-Comp/SSIM, higher is better; for Inception/CLIP distance, lower is better)

Method	Similarity Measure			
	Low-Level		High-Level	
	Pix-Comp ↑	SSIM ↑	Inception ↓	CLIP ↓
Shen et al.[217]	79.7%	0.582	0.829	0.358
Beliy et al.[14]	85.3%	0.597	0.865	0.424
Gaziv et al.[73]	**91.5%**	**0.601**	0.841	0.387
Ren et al.[198]	87.3%	0.588	0.847	0.383
Mozafari et al.[155]	54.3%	0.450	0.818	0.352
IC-GAN (Random)	64.1%	0.467	0.761	0.328
IC-GAN (Noise)	66.5%	0.489	0.744	**0.320**
IC-GAN (Dense)	67.2%	0.491	**0.742**	0.330

These qualitative observations are supported by the quantitative comparison of methods in Table 2.1, according to both low-level measures of image quality (Pix-Comp, SSIM) and higher-level "semantic" measures (Inception or CLIP distance). Pix-Comp is a 2-way comparison of pixel-wise correlation measures computed over the whole test set. We used the results reported by authors in their respective papers, except for Gaziv et al.[73], who did not report Pix-Comp: we re-computed it over the reconstructed images provided in their supplementary material. All other metrics (SSIM[250], Inception-V3[233] distance, and CLIP ViT-B/32[185] distance), were computed over the seven common image reconstructions presented in Figure 2.4. Our own results are presented for three different versions of IC-GAN decoding, using different combinations of the three brain regression models in Figure 2.2, to evaluate the effects of each regressor on per-

formance. First, the IC-GAN (Random) version uses brain-decoded instance features together with randomly sampled noise vectors from a normal distribution. Second, the IC-GAN (Noise) version combines brain-decoded instance features with brain-decoded noise vectors, without using the brain-decoded dense vectors (instead, the output of the first dense layer is used directly). Finally, IC-GAN (Dense) is the complete framework described in Figure 2.2, which uses all the brain-decoded latent variables (thus overriding the dense vector with its brain-decoded version). The table indicates that most other methods yield better results than IC-GAN on the low-level measures (Pix-Comp, SSIM), except for Mozafari et al[155]; like ours, that study was aimed at matching higher-level "semantic" aspects of the input images. Importantly, IC-GAN outperforms the Mozafari et al method for both low-level measures. For the high-level measures (Inception and CLIP Distances), IC-GAN demonstrates state-of-the-art performance, surpassing all methods–including Mozafari et al.–by a significant margin.

The comparison of the 3 versions of our IC-GAN method reveals that the inclusion of both the brain-decoded noise vector (IC-GAN Noise) and the brain-decoded dense vector (IC-GAN Dense) helps improve the model's ability to capture low-level details. Still, the full model remains inferior to many previous methods in this respect. Regarding high-level semantic attributes, while the full method IC-GAN (Dense) is superior to IC-GAN (Noise) for the Inception distance, the opposite is true for the CLIP distance. This could be because Inception features include more spatially structured information than CLIP features; indeed, the function of dense vectors in our method is precisely to capture the image spatial

structure that is less explicitly encoded in the noise vectors.

2.2.4.2 Semantic Analysis of Visual Encoding in Brain ROIs

Our brain decoding model, relying on the latent space(s) of the IC-GAN network, can reconstruct the high-level content of perceived images better than all prior methods, while retaining more low-level details than at least some of these methods. From a neuroscience viewpoint, can this brain decoding model also help us understand the neural coding of visual information in the brain? Here, we use our model to explore and directly visualize the types of information that are preferentially represented in various brain regions-of-interest (ROIs).

The fMRI dataset counts seven distinct ROIs across visual cortex for each subject–in hierarchical order: V1, V2, V3, V4, LOC (Lateral Occipital Complex), FFA (Fusiform Face Area) and PPA (Parahippocampal Place Area). First, we ask whether each region carries more information about high-level latent features–as captured by the model's instance features–or about low-level properties–as captured by the model's dense vector (note that similar results, not shown here, were obtained for the noise vector instead of the dense vector). To answer this question, for each brain voxel we compared the L_1 norm of the model's ridge regression weights for the instance features vs. dense vectors (Figure 2.5). As expected, lower brain regions (V1-V3) were more informative about the dense vector, while higher brain regions (V4, LOC, FFA, PPA) carried more information about instance features.

Next, we use our brain decoding model to visualize the "optimal" stimulus

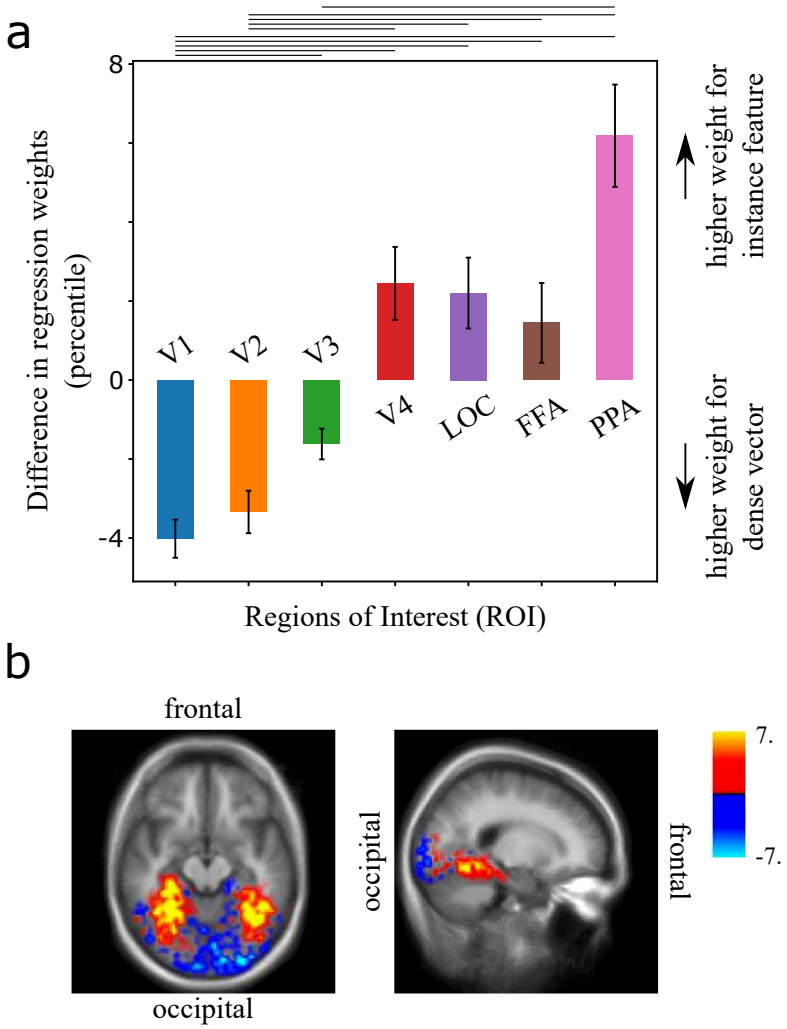

Figure 2.5: Mapping of instance features vs. dense vectors over brain regions. (a) Difference between the percentiles of the regression weights (L_1 norm) for the instance features vs. the dense vector, averaged over voxels in each ROI. Positive values indicate relatively higher weight for instance features compared to the dense vector, and vice versa. Error bars represent standard error of the mean across 5 subjects. Horizontal bars at the top indicate statistical significance of the comparison between ROIs at the two endpoints, with Welch's t-test ($p < 0.008$, Bonferroni correction for six multiple comparisons) (b) Voxel-by-voxel maps (left: axial; right: sagittal) of the difference between the percentiles of the regression weights (L_1 norm) for the instance features (red) vs. the dense vector (blue), averaged over the 5 subjects. Dense vector weights are higher in early visual cortex (occipital regions), while instance feature weights are larger in higher visual cortex (temporal regions).

Figure 2.6: Generated images from synthetic fMRI patterns constructed by activating all voxels in a specific brain region-of-interest (ROI), and none outside of the ROI. The rows represent various brain regions: V1, V2, V3, V4, LOC, FFA, and PPA. The first column is generated after averaging the brain-predicted latent variables for all five subjects. The following columns are for individual subjects.

for each brain region. Instead of using fMRI patterns recorded from subjects, we synthesized seven patterns, one for each ROI, with a value of 1 for all voxels inside the ROI and 0 outside. We provided these synthetic patterns to the three trained ridge regression models to obtain predicted latent variables (as described in Figure 2.2). To mitigate the scaling problem, we normalized instance features to have unit norms. We then passed the predicted latent variables through the IC-GAN generator to generate images.

Previously, Gu et al.[81] synthesized optimal images for different ROIs using a BigGAN generator and a feature extractor. They iteratively optimized the latent variables of the generator in such a way that predicted fMRI patterns (obtained via the feature extractor) maximized activation in a specific ROI. In contrast, our method involves a single pass through our image reconstruction pipeline, and does not require iterative optimization of the latent variables. Figure 2.6 presents the generated images from each subject (second to sixth columns), together with reconstructions using averaged latent variables across all five subjects (first column).

In lower visual cortex (V1-V2), basic textures (foliage, trees, stones) are produced rather than (or in addition to) identifiable objects. The textures emphasize the periphery of the visual field, in line with the fact that V1-V2 have small receptive fields that can be positioned at high visual eccentricity. For V3 and V4, the generated textures present more regularity than V1 and V2, and we begin to see visuals close to objects with multiple parts, including text-like symbols, notably in V4. LOC is known for its selectivity to object shapes; when maxi-

mizing this region's response, IC-GAN generates complete objects at the center of the image, rather than extended textures. At this stage, the visual periphery appears empty or blurry, in contrast with the crisp peripheral textures produced for V1-V2. In FFA, a high-level region known for its selectivity to face images, IC-GAN generates human and animal faces. The presence of animal faces is not unexpected, since the ImageNet dataset (on which IC-GAN was trained) contains many more animal images than human images. Some previous experimental and computational work[81,16] also suggests that fusiform regions may show a preferential response to animals, and particularly dogs. Nonetheless, the model still generates human face images for two of the subjects. The last ROI from the higher visual cortex is PPA, known for its selectivity to environmental scenes like indoor and outdoor places. IC-GAN also generates indoor and outdoor places when the voxels of this region are activated. Some of the images have an object in the center of the scene; this might be caused by the training of the IC-GAN model on ImageNet–an object-centered dataset. It is worth noting that PPA-optimized images produce more details in the visual periphery than FFA-optimized images; this is compatible with the known difference in preferential eccentricity between the two regions[135]. Overall, the outcomes of this analysis are consistent with findings from the neuroscience literature, indicating that our IC-GAN-based model learned to appropriately decode visual feature selectivity in the brain. Most importantly, the method allows us to *directly visualize* this selectivity, rather than inferring it from extended experiments.

2.2.5 Discussion

In this paper, we presented a framework for natural image reconstruction from fMRI patterns using the IC-GAN model, pretrained on ImageNet. First, we extracted instance features, noise vectors, and dense vectors from training images, and trained ridge regression models from fMRI patterns to these latent variables. With these regression models, we decoded latent variables from the test fMRI patterns, and finally reconstructed images with the IC-GAN generator.

Many previous studies implemented fMRI reconstruction frameworks with deep generative models. However, these models were able to reconstruct either low-level or high-level features of the images. Our method demonstrated state-of-the-art performance on reconstructing semantic (high-level) attributes of the images, both qualitatively and quantitatively, while generating naturalistic-looking images. Meanwhile, compared to other semantically oriented models (e.g. Mozafari et al.[155], an approach based on BigBiGAN), it was able to maintain more low-level details. Furthermore, we could use our fMRI-based image reconstruction model to visualize images decoded from synthetic fMRI patterns, designed to maximize activations in specific brain ROIs. The results of this analysis were aligned with the existing neuroscience literature, opening a range of possibilities for future brain exploration and visualization techniques.

We acknowledge that there is still room for improving our model, especially in terms of better reproducing low-level details. This may be achieved in future work by improving our optimization of the noise and dense vectors, or by pairing IC-GAN with other generative networks more focused on low-level image properties.

2.2.6 Acknowledgements

This work was funded by AI-REPS grant ANR-18-CE37-0007-01 and ANITI (Artificial and Natural Intelligence Toulouse Institute) grant ANR-19-PI3A-0004.

2.3 Epilogue to the main article:

In this chapter, we used the IC-GAN model to create a framework for natural image reconstruction. The framework demonstrated state-of-the-art performance both quantitatively and qualitatively, particularly for high-level attributes. We also developed a semantic analysis method to explore the semantic information in specific brain regions using visualization methods.

After completing the article, we submitted it to the International Joint Conference on Neural Networks 2022. It was accepted as an oral presentation, and I presented it in Padua, Italy at the IEEE World Congress on Computational Intelligence 2022 on behalf of our team.

Chapter 3

BrainDiffuser: Natural scene reconstruction from fMRI signals using generative latent diffusion

This chapter presents the second study of the thesis. The study establishes a two-stage framework, involving Very Deep VAE and Versatile Diffusion models, to reconstruct natural scenes from fMRI signals. As in the first study, the semantic information in certain regions of the brain involved in the visual task is analyzed.

3.1 Prologue to the main article :

Following the success of our IC-GAN reconstruction paper, it has gained recognition among researchers in this field and has received numerous citations in a short amount of time. During the submission process of that paper, the Natural Scenes Dataset (NSD) was publicly released. This dataset's vast amount of images and fMRI trial samples has drawn the attention of those studying neural

decoding. The dataset used in this study was more complex, involving multi-subject complex scenes, compared to the single-object-oriented images used in the first study. Initial reconstruction experiments were conducted on the dataset using the IC-GAN approach before any reconstruction papers were published on NSD. Although the reconstructions were not very similar to the original images, they demonstrated that NSD was suitable not only for neural decoding analyses but also for scene reconstruction. Therefore, in my third year, when I returned to Toulouse from vacation, I began working on image reconstruction using NSD. Meanwhile, text-to-image generation models and latent diffusion models were becoming increasingly popular due to their impressive results. It was anticipated that these models would be useful for reconstructing complex scenes. The first paper on NSD reconstruction was 'Mind Reader: Reconstructing complex images from brain activities', which utilized a Style-GAN backbone for text-to-image generation. During our initial experiments with the reconstruction of NSD scenes using Stable Diffusion models, we encountered the preprint 'High-resolution image reconstruction with latent diffusion models from human brain activity.' These studies demonstrated that generative models utilizing text modalities produced semantically satisfactory results for scene reconstruction. However, we concluded that a model utilizing semantic information from both images and text captions was necessary. In the following weeks, Xu et al. released the Versatile Diffusion model[255] and we conducted our initial experiments with it. We found that the results were semantically superior and more realistic than those of previous models. We believed that our layout information could be improved by utilizing a model

that effectively incorporates low-level information. Therefore, we employed the Very Deep VAE architecture, which we had previously found to perform well in low-level reconstruction, as a prior stage to Versatile Diffusion. After obtaining state-of-the-art results for both semantic and low-level reconstruction, we also conducted semantic ROI analyses similar to those in the IC-GAN paper. The following article presents the results of these studies.

3.2 Main article :

3.2.1 Abstract

In neural decoding research, one of the most intriguing topics is the reconstruction of perceived natural images based on fMRI signals. Previous studies have succeeded in re-creating different aspects of the visuals, such as low-level properties (shape, texture, layout) or high-level features (category of objects, descriptive semantics of scenes) but have typically failed to reconstruct these properties together for complex scene images. Generative AI has recently made a leap forward with latent diffusion models capable of generating high-complexity images. Here, we investigate how to take advantage of this innovative technology for brain decoding. We present a two-stage scene reconstruction framework called "Brain-Diffuser". In the first stage, starting from fMRI signals, we reconstruct images that capture low-level properties and overall layout using a VDVAE (Very Deep Variational Autoencoder) model. In the second stage, we use the image-to-image framework of a latent diffusion model (Versatile Diffusion) conditioned on predicted multimodal (text and visual) features, to generate final reconstructed images. On the publicly

available Natural Scenes Dataset benchmark, our method outperforms previous models both qualitatively and quantitatively. When applied to synthetic fMRI patterns generated from individual ROI (region-of-interest) masks, our trained model creates compelling "ROI-optimal" scenes consistent with neuroscientific knowledge. Thus, the proposed methodology can have an impact on both applied (e.g. brain-computer interface) and fundamental neuroscience.

3.2.2 Introduction

Establishing neural encoding and decoding techniques is one way for researchers to discover how the brain and cognition work. Recent developments in modeling and computation have opened up new ways of decoding information from brain signals. Numerous studies in the field of vision research have employed statistical techniques and machine learning to decode specific information from fMRI (functional Magnetic Resonance Imaging) neural activity, such as position[237] or orientation[113,91], to predict categories of images[90,40], to match exemplar images from a candidate set[119], and to reconstruct images with low levels of complexity, such as simple shapes and structures[154].

In recent years, following the success in the development of deep learning models, many studies utilized deep generative models to reconstruct entire images. These deep generative models included Variational Autoencoders (VAE), Generative Adversarial Networks (GAN), and recently Latent Diffusion Models (LDM). Most of these studies used existing deep generative models, pretrained on large-scale data, and then learned a mapping (with simple regression or more

advanced neural network architectures) to reconstruct the corresponding latent variables from the brain signals. This general method was used to reconstruct images with different levels of complexity such as faces[245,42], single-object-centered images[217], and more complex scenes[2,137].

Most of the earlier works on natural scene reconstruction studied either the Generic Object Decoding[100] or the Deep Image Reconstruction[217] datasets curated by the Kamitani Lab. These datasets consist of 1200 training and 50 testing images from the ImageNet[48] dataset and they differ in the number of fMRI repetitions for training and testing images. One of the pioneer studies in this area is by Shen et al.[217] who optimized input images using a deep generator network with a loss function provided by fMRI-decoded CNN features. Beliy et al.[14] utilized supervised training with {fMRI, stimulus} pairs, alongside an additional consistency loss for unsupervised training with test fMRI data and additional image data. Building on this, Gaziv et al.[73] further improved the method by incorporating a perceptual loss on reconstructed images, resulting in sharper reconstructions. Mozafari et al.[155] introduced a reconstruction model based on BigBiGAN that focused on semantics. Ren et al.[198] devised a dual VAE-GAN model with a three-stage learning strategy that incorporates adversarial learning and knowledge distillation. Ozcelik et al.[170] employed the Instance-Conditioned GAN model to generate reconstructions focused on accurate semantics (by extracting semantic information with the SwAV model) and pose information (with latent optimization). Chen et al.[31] utilized a sparse masked brain modeling on large-scale fMRI data and then trained a double-conditioned diffusion model for visual decoding.

Recently, Allen et al. curated another dataset for visual encoding and decoding studies called Natural Scenes Dataset[2]. For this dataset, 8 subjects viewed thousands of images from the COCO[138] dataset. COCO images contain multiple objects and they are more complex in nature compared to ImageNet images. Because of the number, diversity, and complexity of images included, the NSD dataset—although very recent—is becoming the de facto benchmark for fMRI-based natural scene reconstruction. Thus, it is the dataset that we chose for the present work. There are already three studies that reconstructed images from this dataset, and we can use them as baselines against which to compare our model's performance. The first one is by Lin et al.[137], who utilized the Lafite framework that adapts the StyleGAN2 model for text-to-image generation. Takagi et al.[234] devised a method based on Stable Diffusion, using captions for the semantic information and latent variables from images for the low-level information. Gu et al.[80] improved upon Ozcelik et al.'s[170] IC-GAN framework, by establishing a surface-based convolutional network to process fMRI data instead of using vectorized data in the regression models; they also trained an encoder network to predict pose information, instead of performing latent optimization.

The above studies have fostered advances in reconstructing images with high fidelity, especially in the case of object-centered images (i.e., ImageNet images from the Kamitani dataset). Yet, reconstructing scenes with multiple objects and complex semantic descriptions (i.e., COCO images from the NSD dataset) remains a challenge. Given the remarkable recent success of latent diffusion models[202] in generative AI applications such as text-to-image generation[202,189,159,209,255], we rea-

soned that brain decoding studies could also take advantage of such models. Thus, we propose here a visual reconstruction framework called "Brain-Diffuser", relying on the powerful generation capabilities of Versatile Diffusion[255], a model conditioned on both vision and language representations acquired from the pretrained CLIP[185] model.

Our framework consists of two stages. The first stage, illustrated in Figure 3.1, generates a low-level reconstruction of images (akin to an "initial guess") using a Very Deep Variational Autoencoder (VDVAE)[33]. We generate these reconstructions by training a regression model to associate fMRI signals to the corresponding latent variables of VDVAE for the same training images. In the second stage, illustrated in Figure 3.2, we train two additional regression models: one from fMRI patterns to CLIP-Vision features (extracted by feeding the corresponding images to the CLIP model); and the other one from fMRI patterns to CLIP-Text features (collected by providing to the CLIP model the captions of the corresponding images). Finally, we use the multimodal dual-guidance as well as the image-to-image abilities of the pretrained Versatile Diffusion (VD) model to generate the final reconstructions for test images. Using our trained regression models, for each test fMRI pattern we obtain an "initial guess" image (stage 1, VDVAE reconstruction) used by VD's image-to-image pipeline, as well as predicted CLIP-Vision and CLIP-Text feature vectors (stage 2), jointly used for conditioning VD's diffusion process. We used VDVAE, CLIP, and Versatile Diffusion with their pretrained weights, and did not apply any finetuning. We only trained regression models that transform fMRI patterns to latent variables of the models.

We demonstrate below that the resulting scene images reconstructed by the Brain-Diffuser model are highly naturalistic and retain the overall layout and semantic information of the groundtruth images while showing only minor variations in finer details. Compared to earlier models that exhibited proficiency in capturing certain features of groundtruth images, Brain-Diffuser demonstrates qualitatively and quantitatively superior performance in terms of both high-level and low-level metrics, thus establishing itself as state-of-the-art.

3.2.3 Materials and Methods

3.2.3.1 Dataset

We used the publicly available Natural Scenes Dataset (NSD), a large-scale 7T fMRI dataset[2]. The NSD was collected from 8 subjects viewing images from the COCO[138] dataset. Each image was viewed for 3 seconds, while subjects were engaged in a continuous recognition task (reporting whether they had seen the image at any previous point in the experiment). For our study, we used the 4 subjects (sub1, sub2, sub5, sub7) who completed all trials. The training set that we used thus contained 8859 images and 24980 fMRI trials (up to 3 repetitions for each image), and 982 images and 2770 fMRI trials for the test set. We averaged fMRI trials for the images that had multiple repetitions. We also used the corresponding captions from the COCO dataset. Test images are common for all subjects, while training images are different. We used the provided single-trial beta weights, obtained using generalized linear models with fitted hemodynamic response functions and additional GLMDenoise and ridge regression procedures

('betas_fithrf_GLMdenoise_RR'). We masked preprocessed fMRI signals using the provided NSDGeneral ROI (Region-of-Interest) mask in 1.8 mm resolution. The ROI consists of [15724, 14278, 13039, 12682] voxels for the 4 subjects respectively, and includes many visual areas from the early visual cortex to higher visual areas. For further details on this dataset and the corresponding fMRI preprocessing steps, we refer the reader to the initial paper describing the Natural Scenes Dataset[2].

3.2.3.2 Low-Level Reconstruction of Images using VDVAE (first stage)

A Variational Auto-Encoder (VAE)[122] is a generative model trained to capture an input distribution (such as an image dataset) via a low-dimensional latent space, constrained to follow a particular prior distribution (e.g. Gaussian). When the input dataset takes on a more complex distribution, training a Variational Autoencoder (VAE) can be challenging. Indeed, prior work has found that datasets consisting of natural scene images require many latent variables with complex distributions for which a simple VAE would not suffice; this is why the Very Deep Variational Autoencoder (VDVAE) was introduced[33]. The VDVAE is a hierarchical VAE model, with several layers of conditionally dependent latent variables, each layer adding different details from coarse to fine when transitioning from top to bottom. The hierarchical dependence can be seen in equations (3.1) and (3.2), where z indicates latent representations, x is the input variable, q_φ represents the approximate posterior distribution that is learned when training the encoder, and

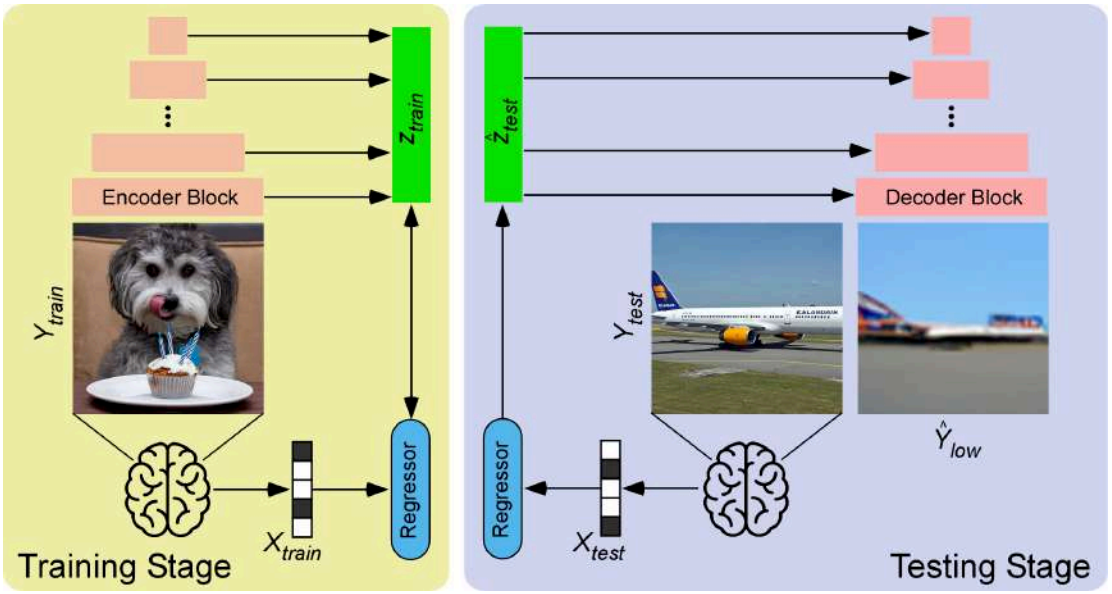

Figure 3.1: Reconstruction of Images via VDVAE (first stage). Training Stage (left). Latent variables (z_{train}) are extracted and concatenated for the first 31 layers of the hierarchy by passing training images (Y_{train}) into the pretrained VDVAE Encoder. A ridge regression model (Regressor) is trained between fMRI patterns (X_{train}) and corresponding latent variables (z_{train}). Testing Stage (right). Test fMRI data (X_{test}) are passed through the trained Regressor to obtain predicted latent variables ($\hat{z}_{test}$). These predicted latent variables are fed to the pretrained VDVAE Decoder to get the low-level reconstruction ($\hat{Y}_{low}$) of the test images (Y_{test}), which will serve as a sort of "initial guess" for the second stage. Note that all VDVAE layers (encoder and decoder blocks) are pretrained and frozen, only the brain-to-latent regression layer (blue box) is trained.

p_θ represents the prior distribution that is learned when training the decoder. The latent variable z_0 is at the top of the hierarchy with the smallest dimension (low resolution, with coarse details) and z_N is at the bottom of the hierarchy with the largest dimension (high resolution, with fine details). Equation (3.1) shows that the latent variables at the bottom of the hierarchy are dependent on those who are at the top (and on the input x). When there is no input (x), it is still possible to generate samples using the prior distribution described in equation (3.2). This hierarchical structure helps the VDVAE learn sufficiently expressive latent

variables to represent complex distributions like natural scene images.

$$q_\varphi(z \mid x) = q_\varphi(z_0 \mid x)\, q_\varphi(z_1 \mid z_0, x) \ldots q_\varphi(z_N \mid z_{<N}, x) \tag{3.1}$$

$$p_\theta(z) = p_\theta(z_0)\, p_\theta(z_1 \mid z_0) \ldots p_\theta(z_N \mid z_{<N}) \tag{3.2}$$

For our study, we used the model provided in[33], trained on a 64×64 resolution ImageNet dataset, and consisting of 75 layers; we only utilized the latent variables from the first 31 layers for the sake of size in regression, since we observed that adding further layers did not make much difference in the reconstruction results (at test time, the latent variables from the remaining layers are sampled according to the prior distribution given in equation (2)).

In the training stage, we fed images to the encoder part of the VDVAE to extract latent variables for each training image (as described in Figure 3.1). We concatenated the latent variables from the 31 layers, which resulted in 91168-dim vectors. Then, we trained a ridge regression model between fMRI training patterns and these concatenated vectors. In the testing stage, we provided test fMRI patterns to the trained regression model and thus predicted latent values for each test image. Then, we fed those latent values to the decoder part of the VDVAE and obtained reconstructed images (64×64 pixels) from the VDVAE. These low-level reconstructions served as an "initial guess" for the diffusion model (second stage).

3.2.3.3 Final Reconstruction of Images using Versatile Diffusion (second stage)

Although the VDVAE was helpful to reconstruct the layout of the image, it is not sufficient for the high-level features, nor does it produce fully naturalistic pictures. For that, we use the Versatile Diffusion[255] model in the second stage of our reconstruction framework. Versatile Diffusion is a recently proposed latent diffusion model (LDM)[202].

LDMs have become highly popular after their success in high-resolution text-to-image generation. In order to train an LDM, first an autoencoder (with encoder $E(.)$ and decoder $D(.)$) is trained on a large-scale image dataset to learn a compressed representation of images x_0, i.e. a latent space $z_0 = E(x_0)$. Then, the forward diffusion process is applied to these latent variables z_0 by adding Gaussian noise in successive timesteps (described in equation (3.3), where t represents the timestep, $\bar{\alpha}_t$ indicates a coefficient derived from the standard deviation of the Gaussian noise, and ε represents the Gaussian noise). The reverse diffusion process is learned via a neural network (Denoising U-Net in the original paper) to predict and remove noise from the noisy latent so as to retrieve the original latent variables. This is done by minimizing the loss function in equation (3.4), where ε is the true Gaussian noise, $\varepsilon_\theta(.)$ represents the neural network being trained to predict the noise, z_t is the latent variable, t is the timestep, and $\tau_\theta(y)$ is the conditioning input for the U-Net. After the reverse diffusion process, the denoised latent variables are passed through the trained decoder $D(.)$ to generate the images. The critical part of this process is that it is possible to condition this reverse

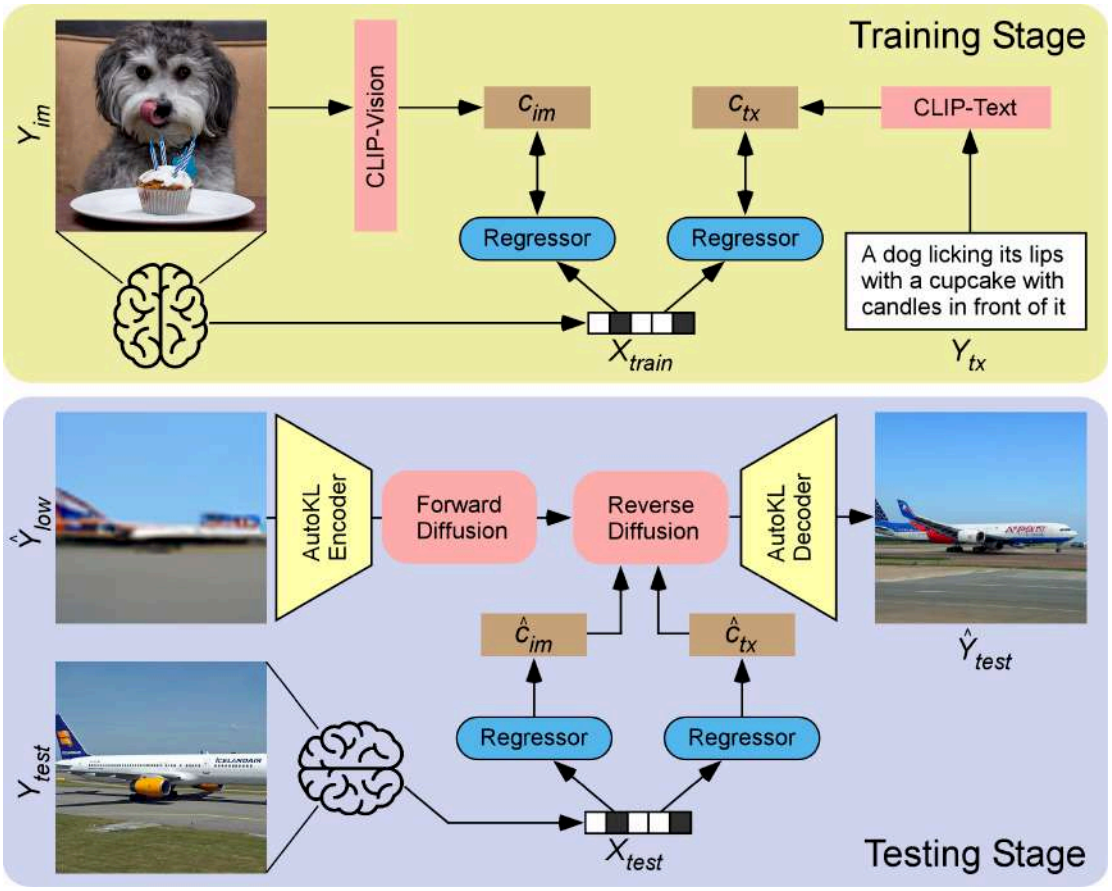

Figure 3.2: Final Reconstruction of Images via Versatile Diffusion (second stage). Training Stage: CLIP-Vision features (c_{im}) are extracted by feeding training images (Y_{im}) to the pretrained CLIP model. CLIP-Text features (c_{tx}) are extracted by providing the corresponding captions (Y_{tx}) to the pretrained CLIP Model. Two different ridge regression models (Regressors) are trained to learn the mapping between these features and fMRI patterns (X_{train}). Testing Stage: Predicted CLIP-Vision ($\hat{c}_{im}$) and CLIP-Text ($\hat{c}_{tx}$) features are computed by giving test fMRI patterns (X_{test}) to the trained regression models. In the image-to-image pipeline of the latent diffusion model, VDVAE reconstructions of test images (the "initial guess" $\hat{Y}_{low}$ from the first stage) are passed through the AutoKL Encoder of the pretrained Versatile Diffusion model, and the obtained latent vectors undergo 37 steps of the forward diffusion process (noise addition). The resulting noisy latent vectors are used to initialize the reverse diffusion process, which is also guided by predicted CLIP-Vision ($\hat{c}_{im}$) and CLIP-Text ($\hat{c}_{tx}$) features jointly in a dual-guided framework. At last, the resulting denoised latent vector is passed through the AutoKL Decoder to generate the final reconstructed image ($\hat{Y}_{test}$). Note that all CLIP (vision and text encoders) and Versatile Diffusion layers (AutoKL encoder and decoder, forward and reverse diffusion blocks) are pretrained and frozen, only the brain-to-latent regression layers (blue boxes) are trained.

diffusion process on different representations (e.g text captions, images, semantic maps). This conditioning process is done by merging conditions ($\tau_\theta(y)$) in the cross-attention block of the Denoising U-Net.

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t}\varepsilon \tag{3.3}$$

$$L_{\text{LDM}} = \mathbb{E}_{t, z_0, \varepsilon, y}\left[\left\|\varepsilon - \varepsilon_\theta\left(z_t, t, \tau_\theta(y)\right)\right\|^2\right] \tag{3.4}$$

The Versatile Diffusion model (see Figure 3.2) is a latent diffusion model with different pathways which allow us to condition the generation process on both text and image features to guide the reverse diffusion process. It is possible to provide CLIP-Vision, CLIP-Text, or both features as conditions in the reverse diffusion process. It is also possible to initialize the reverse diffusion with latent variables obtained from a particular image, rather than from a purely random distribution–this is the image-to-image pipeline that we will use to take advantage of our "initial guess" image from stage 1. The Versatile Diffusion model that we utilized in our framework was trained on the Laion2B-en[212] dataset with 512×512 resolution images and corresponding captions. CLIP (Contrastive Language-Image Pre-training)[185] is a multimodal model designed to assist in different tasks that involve natural language processing and computer vision. It is trained in a contrastive learning approach, where features gathered from images vs. text captions are projected onto separate latent spaces of identical dimensions: CLIP-V refers to the latent space for images and CLIP-T for captions. Similarity scores (e.g. cosine similarity) of the latent space projections for matching images and

captions are optimized throughout training. CLIP is widely used as a feature extractor, due to its high representational capabilities. The CLIP network used in Versatile Diffusion is based on the transformer architecture (ViT-L/14) and pretrained on a large-scale contrastive task.

In stage 2, we thus train two regression models, the first one between fMRI patterns and CLIP-Vision features (with 257×768-dim extracted from the corresponding images where the first vector with 768-dim represents the category-related embedding and the remaining 256 embeddings represent the patches acquired from the images) and the second one between fMRI patterns and CLIP-Text features (77×768-dim extracted from the COCO captions associated with the corresponding images where the 77 embeddings correspond to the number of tokens given to the model as inputs). At testing time, we use the image-to-image pipeline of the latent diffusion model. First, we encode the image reconstructed with the VDVAE model (stage 1) with the AutoKL Encoder (after upsampling the image from 64×64 to 512×512) and add noise to the latent vector for 37 steps of forward diffusion (corresponding to 75% of the 50 steps of full diffusion, which is a commonly used value in the image-to-image pipeline of LDMs.). In this image-to-image pipeline, it is necessary to first add some amount of noise to the latent values using forward diffusion, since LDMs generate images via denoising using reverse diffusion (without noise on the image, the reverse diffusion step would end up with no change). Then, we feed this noisy latent as initialization to the diffusion model and denoise it for 37 steps while conditioning with the predicted CLIP-Vision and CLIP-Text features (stage 2). In every step of reverse diffusion,

we use CLIP-Vision and CLIP-Text jointly in the double-guided diffusion pipeline of Versatile Diffusion, where the cross-attention matrices for both conditions are mixed through linear interpolation (with CLIP-Vision having a relative strength of 0.6 and CLIP-Text of 0.4). The diffusion result is passed through the AutoKL Decoder to produce our final 512×512 pixel reconstruction.

3.2.3.4 Code availability

The code for our project, including scripts to train regression models, pre-trained weights, and scripts to produce reconstructions for test images and for ROI-based synthetic patterns, is publicly available at github.com/ozcelikfu/brain-diffuser.

3.2.4 Results and Analyses

3.2.4.1 Image reconstruction examples

We present examples of reconstructions from our model in Figure 3.3. While we present the results of each individual subject in different columns, we also added results gathered by averaging the latent variables predicted by all subjects. In general, we see that reconstructed images capture most of the layout and semantics of the groundtruth images, while there remain differences in pixel-level details. For instance, looking specifically at the first four images on the left, we see that the reconstructed pose (3D orientation) of the plane (first image) is correct for every subject although there are some differences in the details of the plane and also in the texture of the background. Nonetheless, the fact that a commercial

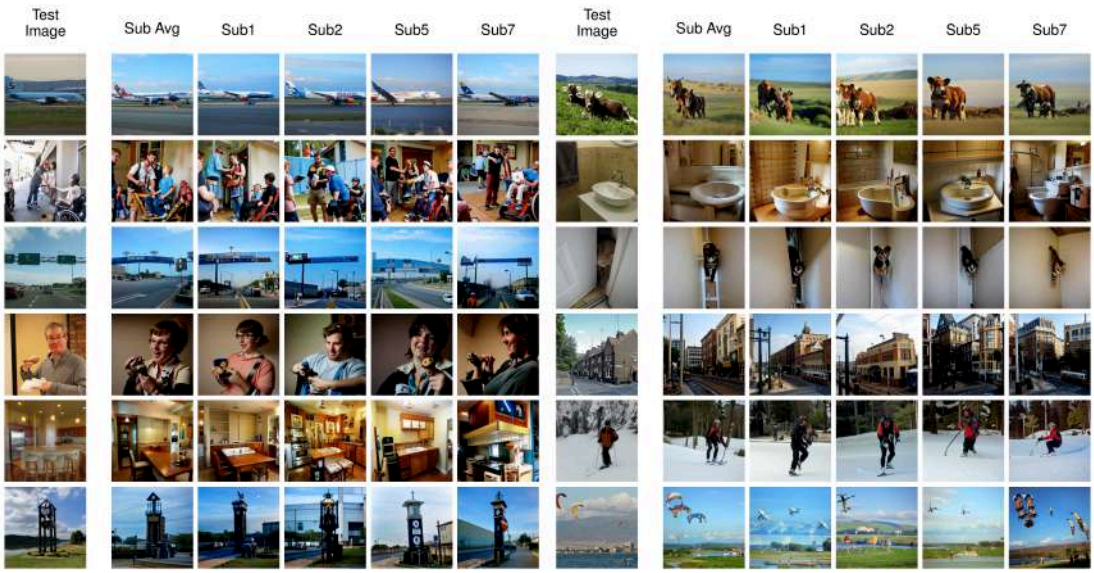

Figure 3.3: Examples of fMRI Reconstructions from our Brain-Diffuser model. The first column is the groundtruth image (Test Image). The second column is generated by averaging the predicted latent variables over all 4 subjects seeing the same picture (Sub Avg). The remaining columns are for each individual subject (Sub1, Sub2, Sub5, Sub7)

plane on a runway, facing to the right on a blue sky background was reconstructed in all instances is not a trivial feat. For the second example, all reconstructed images display a group of people, although layouts tend to differ. Still, a person in a wheelchair is visible in the bottom right corner for three of the four subjects. For the third image, the model reconstructed a highway with road signs correctly, although the orientation of the road is different for some of the subjects, and the details of the signs are not entirely matched. On the fourth sample, all reconstructed images show a single person facing left and holding an object in their hand, as in the groundtruth image. The person's details (gender, age, clothing) are different across subjects, e.g. with glasses only reconstructed for subject 1 and in the average across subjects. Reconstructed image contrast also differs from

the ground truth. Similar conclusions can be generalized to most images of the test set: while never passing for a picture-perfect copy, with visible differences in especially color and contrast (due to inherent limitations of the Versatile Diffusion model in this respect), the reconstructed images are always naturalistic (that is, as much as diffusion models can generate) and plausible alternate renditions of the ground truth. Some of the remaining errors and differences may be caused by inherent limitations of the LDMs instead of unsuccessful predictions made by the fMRI-latent mapping model, as it is known that (current) diffusion models can generate unrealistic images in some occasions (e.g. unusual numbers of eyes on faces, fingers in hands.)

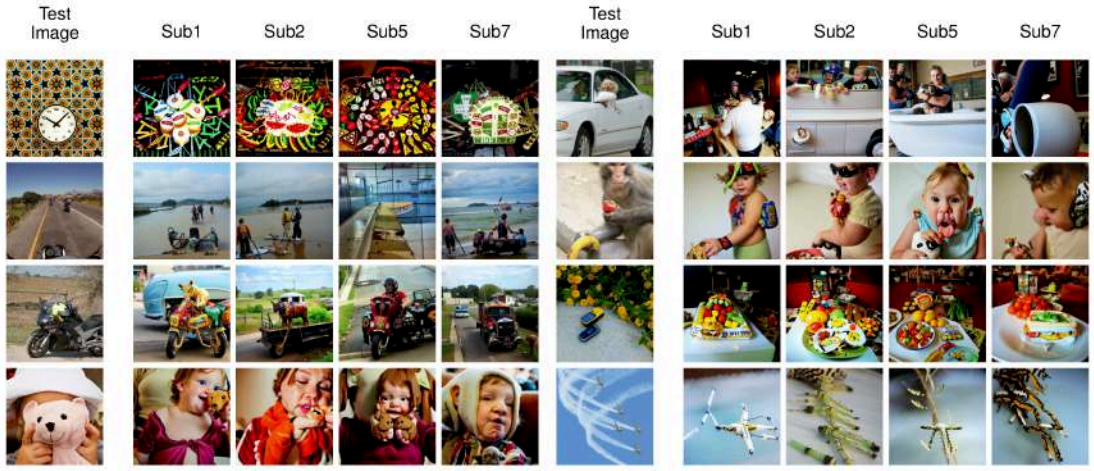

Figure 3.4: Failure cases of fMRI Reconstructions from our Brain-Diffuser model. The first column is the groundtruth image (Test Image). The remaining columns are for each individual subject (Sub1, Sub2, Sub5, Sub7)

We also present some examples of reconstruction failures from our model in Figure 3.4. In these examples, we see that our model can fail due to different reasons. In the first example, although Brain-diffuser reconstructs oval objects

around the center, the complex texture of the background seems to interfere with the object, which is not reconstructed as a clock. For the second example, the reconstructions show sea in the background, although there is no sea in the ground-truth image. On the fourth sample, the teddy bear occluding the kid's face seems to confuse the model, as it generates human faces in the reconstructions. For the sixth example, Brain-Diffuser reconstructs a kid instead of a monkey. These examples highlight the fact that Brain-Diffuser can fail on occasion, due to diverse reasons like complex stimuli, object occlusions, or confusing one object with another.

3.2.4.2 Comparison with state of the art

How do these findings compare to the state of the art? We contrast the qualitative results of our model with three other existing models in Figures 3.5 and 3.6. Lin et al.[137] was the first study that used the NSD dataset for reconstruction. They are similar to our model in terms of utilizing both image and text features as conditions, but they used a StyleGAN2 model instead of an LDM. Takagi et al.[234] is the only other study (in addition to ours) to use a latent diffusion model for reconstructing images from the NSD dataset. Finally, Gu et al.[80] used an Instance-Conditioned GAN model trained on ImageNet. In Figure 3.5, we compared our results with previous studies for the same set of images as in Figure 3.3. Since Lin et al. used a different train-test split for their model, we used a replication of their model on the same train-test split as ours. Takagi et al. and Gu et al. shared generated images from their models with us for comparison. From these

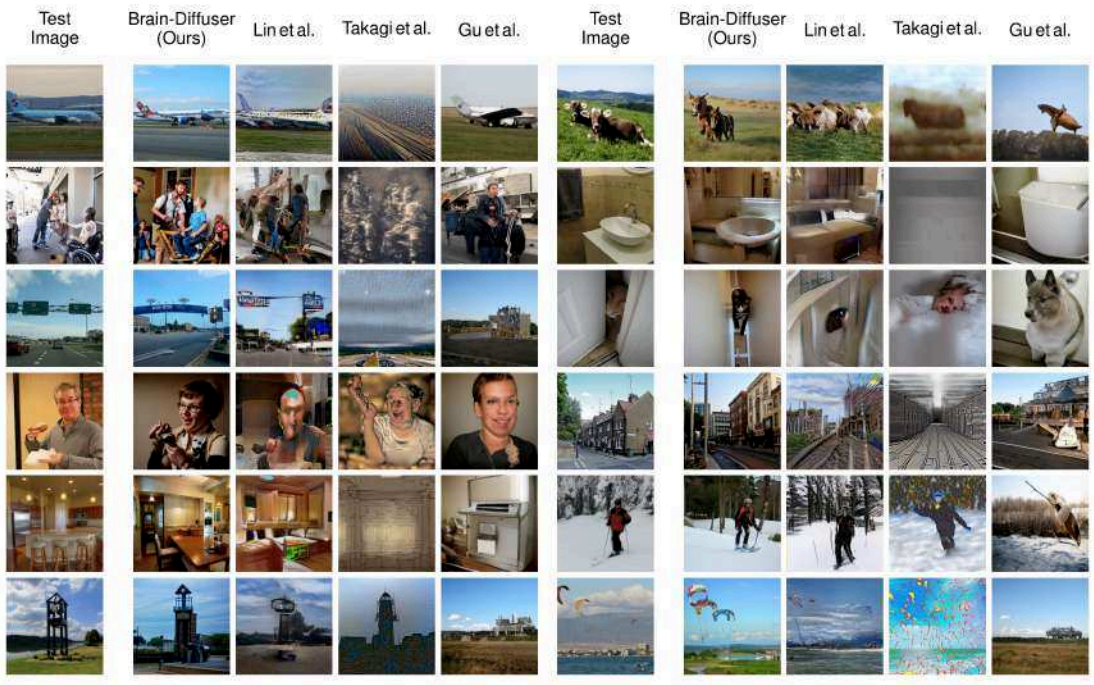

Figure 3.5: Comparison of fMRI Reconstructions for different models on a common set of test images. The first column is the groundtruth image (Test Image). The second column shows reconstructions of our method (Brain-Diffuser). The third column reconstructions are generated by replicating Lin et al.'s method using the code and instructions given by the authors. The fourth and fifth columns are reconstruction results from Takagi et al. and Gu et al. respectively, which were shared by the original authors.

reconstructions, we can see that all methods capture high-level information to a degree, but not all of them are equally good at utilizing this information for image reconstruction. For instance, in the first image with a plane, Brain-Diffuser reconstructed a plane image that looks more similar to the ground-truth image and has a more realistic structure compared to Lin et al. and Gu et al. (and there is no recognizable plane in the reconstruction of Takagi et al.). In the fourth image with a man with glasses, the face is barely recognizable for Lin et al., the reconstruction by Gu et al. does not contain arms or glasses, while Takagi et al.

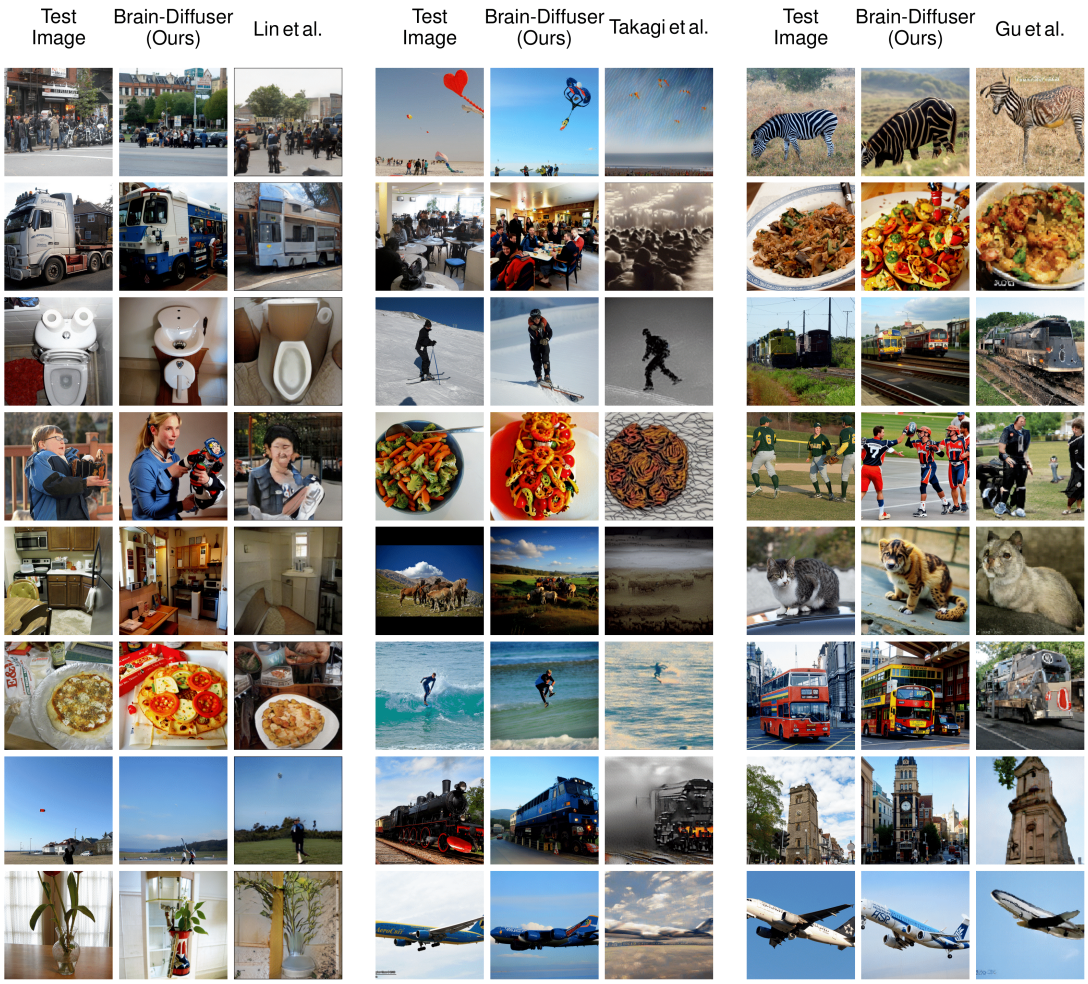

Figure 3.6: Comparison of fMRI Reconstructions for different models on images presented in the papers of the previous methods. Since the presented test images in all methods were different, we did comparisons separately for each model. On the left (first 3 columns), we present the comparison of our model with Lin et al. together with groundtruth test images. On the center (columns 4-6), we present the comparison of our model with Takagi et al. together with groundtruth test images. On the right (last 3 columns), we present the comparison of our model with Gu et al. together with groundtruth test images.

reconstruct an unnatural rendition of a face and arms; in contrast, Brain-Diffuser exhibits a more natural-looking reconstruction and also manages to reconstruct the glasses.

In Figure 2.4, we compare our method against the same three baselines, but using image reconstructions that were reported by the original authors in their papers, which might be more representative of each method's performance. Although Lin et al. seems to be performing better than the other two prior models, in some instances the quality of their reconstructions still lags behind ours. For instance, in the second image, the details of the truck are better represented in our model, while for the third image, the shape of the toilet is better represented in Lin et al. In the fourth image, the color of the clothes is presented more accurately in our model, as well as the fact that the person is holding an item; the person's face also looks more realistic compared to Lin et al. On the other hand, the color and location of the pizza in the sixth image appear more aligned with the ground-truth image for Lin et al. Although Takagi et al. generates easily recognizable silhouettes, they do not seem to perform as well as our model in any qualitative aspect including low-level details, semantics, or naturalness. Finally, when we compare our results to Gu et al., we can see that, although both appear good at reconstructing images with similar semantics, structural aspects are less well represented in their reconstructed images (e.g. unrealistic warped shapes for the train, bus, and building). In contrast, the shape and texture details of our model are more realistic. Since their model has a BigGAN backbone, with few parameters to encode the entire layout of the image (including the object's class, its pose, size, and location), and since it is trained on a single-object-centric dataset (ImageNet), the model seems to be limited in reconstructing complex scenes with multiple objects. On the other hand, since LDMs include a spatially organized

map of features, it is more convenient for them to represent multiple objects; as an example, we see one train in the third image although there are two trains in the groundtruth image, and in the reconstructed image from our model.

Table 3.1: Quantitative Analysis of fMRI Reconstructions. For each measure, the best value is in bold. (For PixCorr, SSIM, AlexNet(2), AlexNet(5), Inception and CLIP metrics, higher is better. For EffNet-B and SwAV distances, lower is better. This is indicated by the arrow pointing up or down, respectively)

Method	Quantitative Measures			
	Low-Level			
	PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑
Lin et al.[137]	–	–	–	–
Takagi et al.[234]	–	–	83.0%	83.0%
Gu et al.[80]	.150	.325	–	–
Brain-Diffuser (Ours)	.254	.356	94.2%	96.2%
	High-Level			
	Inception ↑	CLIP ↑	EffNet-B ↓	SwAV ↓
Lin et al.[137]	78.2%	–	–	–
Takagi et al.[234]	76.0%	77.0%	–	–
Gu et al.[80]	–	–	.862	.465
Brain-Diffuser (Ours)	87.2%	91.5%	.775	.423

3.2.4.3 Quantitative Results

To make the comparison with other models more quantitative, we present the results of 8 different image quality metrics in Table 3.1. PixCorr is the pixel-level correlation of reconstructed and groundtruth images. SSIM[250] is the structural similarity index metric. AlexNet(2) and AlexNet(5) are the 2-way comparisons of the second and fifth layers of AlexNet[129], respectively. Inception is the 2-way comparison of the last pooling layer of InceptionV3[233]. CLIP is the 2-way comparison of the output layer of the CLIP-Vision[185] model. EffNet-B and SwAV are distance metrics gathered from EfficientNet-B1[235] and SwAV-ResNet50[24] models,

respectively. The first four can be considered as low-level metrics, while the last four reflect higher-level properties. For PixCorr and SSIM metrics, we downsampled generated images from 512×512 resolution to 425×425 resolution (i.e. the resolution of groundtruth images in NSD dataset). For the rest of the measures, generated images are preprocessed according to the input properties of each network. Note that not all measures are available for each previous model (depending on what they chose to report). However, each model has at least one point of comparison with ours. Our quantitative comparisons with Takagi et al. and Gu et al. are made according to the exact same test set, i.e., the 982 images that are common for all 4 subjects. Lin et al., on the other hand, reported their results on only Subject 1 and with a custom train-test set split. However, when measuring our model's image quality on the same train-test split as Lin et al, we observed nearly identical results (Inception Score of 87.0%, compared to 78.2% for Lin et al). Our model is the best-performing model by a decent margin for all of the quantitative metrics. Overall, these results show that our model can be considered state-of-the-art for both low-level and high-level quantitative measures.

3.2.4.4 Ablation Studies

In order to reveal the contribution of each component of Brain-Diffuser, we performed an ablation study (with fMRI data of Sub1), and report both quantitative (Table 3.2) and qualitative (Figure 3.7) results. The quantitative results are given in Table 3.2. Our first ablation (Only-VDVAE) considers the results from stage-1 reconstruction only (Figure 3.1) without stage-2 reconstruction (Figure 3.2). This

Table 3.2: Quantitative comparisons of test fMRI reconstructions of Sub1 with various ablations of the full model. For each measure, the best value is in bold. (For PixCorr, SSIM, AlexNet(2), AlexNet(5), Inception, and CLIP metrics, higher is better. For EffNet-B and SwAV distances, lower is better. This is indicated by the arrow pointing up or down, respectively)

Method	Quantitative Measures			
	Low-Level			
	PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑
Only-VDVAE	.358	.437	97.7%	97.6%
Brain-Diffuser w/o VDVAE	.143	.302	85.6%	93.0%
Brain-Diffuser w/o CLIP-Text	.279	.333	95.6%	97.0%
Brain-Diffuser w/o CLIP-Vision	.327	.433	93.9%	94.1%
Brain-Diffuser	.305	.367	96.7%	97.4%
	High-Level			
	Inception ↑	CLIP ↑	EffNet-B ↓	SwAV ↓
Only-VDVAE	77.0%	71.1%	.906	.581
Brain-Diffuser w/o VDVAE	87.3%	92.6%	.775	.414
Brain-Diffuser w/o CLIP-Text	87.9%	91.2%	.796	.436
Brain-Diffuser w/o CLIP-Vision	84.7%	84.5%	.821	.509
Brain-Diffuser	87.8%	92.5%	.768	.415

Only-VDVAE model provides the best results for all low-level measures, but the worst (by a large margin) for all high-level measures. This pattern of results is expected since the VDVAE reconstruction relies on low-level information without a contribution of semantic information from stage-2. By contrast, Brain-Diffuser without the VDVAE component (i.e., stage-2 reconstruction but with random initialization of the autoKL latent vector) performs worst on low-level measures (by a large margin), while it is among the best in high-level measures. This is also reasonable since this ablated model generates the reconstructions from only high-level features obtained from CLIP-Text and CLIP-Vision models and does not have much information about low-level information such as layout. Together, these results indicate that the VDVAE "initial guess" (stage-1) is necessary but

not sufficient for optimal reconstruction. This is evident in the results from the full Brain-Diffuser model (last row in Table 3.2), where the contribution from VDVAE (stage 1) brings strong improvements in low-level measures, with near-optimal high-level features. In another ablation, we evaluate Brain-Diffuser without CLIP-Text. Compared to the full model, there is a sizeable decrement in both low-level and high-level measures, except Inception. While the contribution of CLIP-Text to the reconstruction of high-level semantic features is expected, its improvement of low-level measures is more surprising but could be explained by semantic information related to the image layout itself, such as the number or orientation of objects (see examples in Figure 3.7). Finally, Brain-Diffuser without CLIP-Vision, surprisingly, retains high performance on the low-level PixCorr and SSIM measures (lower than Only-VDVAE, but higher than the full model); we assume that this could be due to insufficient diffusion steps (as discussed further below), preventing the reconstruction from deviating from the VDVAE initial guess. For all other measures (including low-level AlexNet measures), removing CLIP-Vision guidance severely impairs the performance of Brain-Diffuser. Overall, when jointly considering low-level and high-level measures, these quantitative results show that the full Brain-Diffuser model is better than any other variation or ablation.

We also present qualitative results in Figure 3.7 with the same set of images presented in Figure 3.3 of the main manuscript. Reconstructions from the Only-VDVAE model (i.e., stage-1 without stage-2) match the low-level details (e.g. shapes, layouts) of the groundtruth images, but they look like vague silhouettes

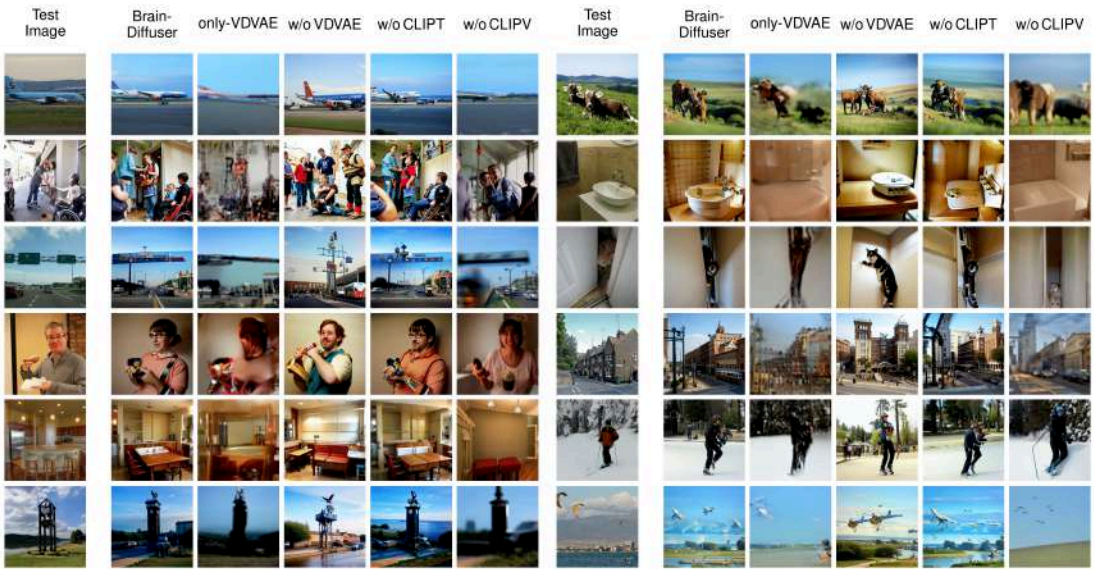

Figure 3.7: Examples of fMRI test reconstructions from Sub1 with various ablations of the full model. The first column is the groundtruth image (Test Image). The second column shows reconstructions from the full Brain-Diffuser model with all of its components. The third column is for reconstructions of the Only-VDVAE model. The remaining columns are for Brain-Diffuser with one of its components excluded, in order: without VDVAE, without CLIP-Text, and without CLIP-Vision.

rather than natural images. In contrast, Brain-Diffuser without VDVAE generates images that match high-level properties (semantics) of groundtruth images but lack positional information about the objects and their layout. This is particularly clear for the fourth image in the right part of the figure, where the layout of the street and buildings is properly captured by VDVAE (and thus, also by the full model), but is lacking in the VDVAE ablation. The images generated by Brain-Diffuser without CLIP-Text appear very close to those from the full model but with some notable differences. One example is the ski image (Row 5 on the right part of the figure), where the full model generates a single person (as in the groundtruth) while the model without CLIP-Text generates two people. An-

other example is the plane image (Row 1 on the left part of the figure) where the model without CLIP-Text does not produce an image with the correctly positioned plane. Finally, reconstructions from Brain-Diffuser without CLIP-Vision appear quite blurry, and somehow in between the Only-VDVAE and the full model reconstructions. This could be an indication that forward and reverse diffusion steps were not sufficient for this model. Still, increasing the number of diffusion steps may not be a good solution since that would cause the model to lose low-level information provided by VDVAE. Overall, these qualitative examples corroborate the quantitative findings in Table 3.2 and make it clear that the Brain-Diffuser model represents the optimal compromise for both low-level details and high-level semantic features.

3.2.4.5 Which brain regions are used?

In order to understand the relationship between brain regions and the various components of our model (VDVAE, CLIP-Vision, CLIP-Text), we performed a region-of-interest (ROI) analysis of the regression weights. We used 4 visual ROIs derived from population receptive field (pRF) experiments, and 4 ROIs derived from functional localization (fLoc) experiments. All experiments were provided along with the NSD dataset by the original authors. These ROIs are as follows (region names following the terminology adopted in Allen et al.[2]): V1 is the concatenation of V1 ventral (V1v) and V1 dorsal (V1d), and similarly for V2 and V3; V4 is the human V4 (hV4); the Face-ROI consists of the union of OFA, FFA-1, FFA-2, mTL-faces, and aTL-faces; Word-ROI consists of OWFA, VWFA-

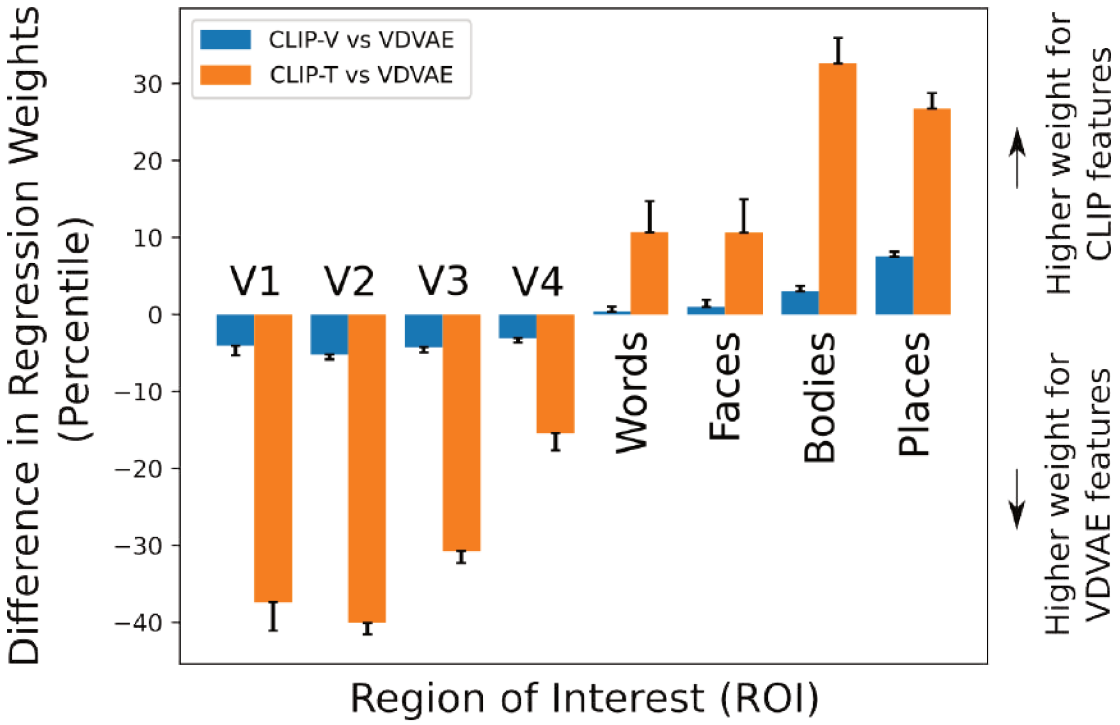

Figure 3.8: Difference between the percentiles of the regression weights (L_1 norm) for the CLIP features (CLIP-V and CLIP-T) vs. the VDVAE features, averaged over voxels in each ROI and normalized by the average percentile of VDVAE features for the same ROI. Positive values indicate relatively higher regression weight for CLIP features compared to the VDVAE features, and vice versa. Error bars represent the standard error of the mean across 4 subjects.

1, VWFA-2, mfs-words, and mTL-words; Place-ROI consists of OPA, PPA, and RSC; Body-ROI consists of EBA, FBA-1, FBA-2, and mTL-bodies. For each voxel in these regions, we computed the strength of the regression weights (L_1 norm) for the CLIP features (CLIP-V and CLIP-T) and the VDVAE features, expressed as a percentile. Because the absolute regression weights can be affected by the number of voxels in each region, as well as the overall activity level and the noise level, we report our results as a *difference* in regression weights between CLIP features and VDVAE features. The results in Figure 3.8 show that early regions (V1-V4)

are more informative about the VDVAE features, while category-specific higher brain regions (Words, Faces, Bodies, Places) carry more information about CLIP features. Another important observation is that the differences between CLIP-V and VDVAE are in the same direction, but much weaker than the differences between CLIP-T and VDVAE. This may indicate that although the Versatile Diffusion model uses CLIP-V features for high-level guidance, these features still contain more information about low-level properties than CLIP-T features.

3.2.4.6 ROI-optimal stimuli

Beyond brain decoding, we show here that our method can also be used to help understand the functional properties of specific regions-of-interest (ROIs) in the brain. Although we know from early studies in the neuroscience literature[103,79,174,71,244,116,62,90,167] what sort of visual properties would best activate neurons in each brain region, there are only a few studies[10,192,81,170,156] which attempted to directly visualize an "optimal" stimulus for a given brain region. Our method can easily be adapted for this purpose. We define "ROI-optimal" as images that would activate a certain ROI maximally while not activating other ROIs (or just activating them minimally). We analyzed the same 8 ROIs (V1, V2, V3, V4, Face-ROI, Word-ROI, Place-ROI, and Body-ROI) that we discussed in the previous section. We used the intersection of these regions with NSDGeneral (the one we used for training our decoding system), each time creating a synthetic fMRI pattern where the ROI was active (signal set to 1) and the rest of the brain inactive (signal set to 0). From this synthetic pattern, our system

Figure 3.9: Images reconstructed from synthetic fMRI patterns created by activating specific regions-of-interests (ROIs). The first 4 rows present individual subjects: Sub1, Sub2, Sub5, and Sub7. The last row is generated by averaging the latent vectors predicted from all 4 subjects. The columns present ROIs: First four are ROIs from the visual cortex (V1-V4) gathered by population receptive field experiments, and the last four are ROIs that are specified with functional localization experiments (Face-ROI, Word-ROI, Place-ROI, Body-ROI). Since our synthetic fMRI patterns produce distribution shifts in the latent variables, which in turn can affect the contrast of the reconstructed images, histogram stretching and equalization are applied on color histograms of generated images for visualization purposes.

could then generate predicted latent variables, and directly reconstruct an equivalent visual scene, corresponding to the "ROI-optimal" image. Surprisingly, this simple and deterministic approach, inspired by the analysis in Ozcelik et al[170], still gives plausible results. Since the synthetic fMRI patterns can be considered out-of-distribution (because there are no similar patterns in the training set), we re-normalized the generated latent variables to give them a similar euclidean norm to the training samples. This procedure helped the diffusion model to generate

meaningful images that are shown in Figure 3.9.

Upon inspecting the generated "ROI-optimal" images for visual ROIs, we see that V1 produces high-contrast scenes with very detailed textures extending to the visual periphery, such as trees and foliage in a park with numerous small human or animal figures. V2 is similar (especially for Sub1 and Sub5, which also display humans in a luxuriant garden environment), but with slightly broader elements and less peripheral detail (e.g. trays filled with various foods in Sub2, Sub7, and in the subject-average). Continuing along the same trend, V3 and V4 produce larger objects compared to the earlier regions, with repeating patterns and global structure. V4 especially generates colorful, high-contrast objects resembling toys on a bright background.

The ROI-optimal images for functionally defined high-level ROIs are even easier to interpret, as they tend to coincide with each region's known category preference. For instance, the model generated multiple face images for the Face-ROI, including humans and sometimes even animal faces (e.g. dogs in Sub5 and in the subject-average). For the Word-ROI, the model generated characters and pseudo-words on objects or signs (except for Sub7). Architecturally plausible indoor scene layouts were produced for the Place-ROI. Finally, for the Body-ROI, the reconstructed images show both human and animal body parts like arms and legs engaged in active behavior like sports or running.

In another exploratory experiment aiming to understand the effects of combining ROIs, we repeated the analysis of Figure 3.9, this time using combinations of activations for different ROIs (Figure 3.10). In the first column where we activated

Figure 3.10: Images reconstructed from synthetic fMRI patterns created by activating combinations of different regions-of-interests (ROIs). The first 4 rows present individual subjects: Sub1, Sub2, Sub5, and Sub7. The last row is generated by averaging the latent vectors predicted from all 4 subjects. The columns present different combinations of ROIs: The first column is where all four regions in the visual cortex are activated at once (V1, V2, V3, and V4). The remaining columns are combinations of activations of these visual ROIs with one of the functional ROIs: Face-ROI, Word-ROI, Place-ROI, and Body-ROI, respectively.

all the low-level visual regions together (V1-V4), we see scrambled and regular patterns in different parts of the images as well as some identifiable objects for some of the subjects—but there are no apparent objects that are commonly identifiable across all subjects. In the next columns, we combined activations across all low-level visual regions (V1-V4) and one of the functionally defined high-level ROIs. In the second column where we combined the visual regions (V1-V4) with Face-ROI, we see human and animal faces in all images, although some scrambled high-contrast patterns also continue to exist in different parts of images. In the third column where we combined the visual regions with Word-ROI, letter-like patterns or pseudo-words can be seen in the upper part of the image for subject 1 and middle and lower regions for subject 5, but they are less visible than in the analysis of Figure 3.9. In the fourth column where we combined the visual regions with Place-ROI, the model generates architectural interior and exterior parts, and the scrambled patterns cease to exist for these images. Finally, in the fifth and last column where we combined the visual regions with Body-ROI, we see vaguely identifiable human and animal body parts like arms and legs. This proof-of-concept experiment reveals what happens when we combine the activations of different regions instead of activating one ROI in isolation. Although there are visual differences between generated images from Figure 3.10 and Figure 3.9, we continue to observe similar semantic relationships between the functional ROIs and the corresponding images.

While these results mainly confirm decades of converging knowledge from the neuroscience literature on neuronal selectivity in the ventral visual pathway, this

Figure 3.11: Images reconstructed from synthetic fMRI patterns created by activating regions-of-interests (ROIs) in the visual cortex with different eccentricities. The first 4 rows present individual subjects: Sub1, Sub2, Sub5, and Sub7. The last row is generated by averaging the latent vectors predicted from all 4 subjects. The columns present concentric regions with increasing eccentricity coverage (0°<e<0.5°, 0.5°<e<1°, 1°<e<2°, 2°<e<4°, and 4°<e, where "e" stands for eccentricity). Histogram stretching and equalization is applied for visualization purposes.

method allowed us to directly visualize functional properties in vivid detail and high-resolution images. Furthermore, the technique introduced here could easily

be extended to study retinotopic or eccentricity-based cortical organization. As a proof of concept, we also applied our ROI analysis to visual regions defined by different eccentricity preferences. Similar to hierarchical regions in the visual cortex (V1, V2, V3, and hV4) these eccentricity-based regions ($0°<e<0.5°$, $0.5°<e<1°$, $1°<e<2°$, $2°<e<4°$ and $4°<e$, where e stands for "eccentricity") were also extracted by population receptive field (pRF) experiments. These regions thus reflect the eccentricity preference of the retinotopic cortex, where degrees close to 0° indicate central vision (closer to the fovea) and higher degrees indicate peripheral vision. The corresponding results are shown in Figure 3.11. It is difficult to see a clear pattern for eccentricities between 0° and 0.5° ($0°<e<0.5°$), as the corresponding portion of the image might be too small to be considered meaningful for the model. A noticeable aspect, however, is that all images for that ROI have detailed and high-contrast objects in the center (though there are also objects in the periphery). For eccentricities between 0.5° and 1° ($0.5°<e<1°$), and between 1° and 2° ($1°<e<2°$), we begin to see larger objects (e.g. humans, animals, blobs) at the center of the images. When we reach eccentricities between 2 and 4° ($2°<e<4°$) and beyond ($4°<e$), we start to see these objects (or animals, humans, and blobs) move towards the periphery, while the center of the images is mostly empty. These results highlight two important findings: first, the latent representations used by Brain-Diffuser (combining latent features from VDVAE, CLIP-Vision, and CLIP-Text) are precise enough to convey information about the spatial localization of objects in the image; second, we see that Brain-Diffuser managed to learn the eccentricity-based retinotopic organization of the cortex

from these representations.

3.2.5 Discussion

In this study, we designed a two-stage framework (Brain-Diffuser) that reconstructs images from fMRI patterns using generative models based on latent diffusion. In the first stage, we used the VDVAE model to generate "initial guess" reconstructions focusing on low-level details. Then in the second stage, we used the image-to-image pipeline of the Versatile Diffusion model, starting from this initial guess, to generate final reconstructions via diffusion, guided by both predicted CLIP-Vision and CLIP-Text features. As we relied on pre-trained and publicly available models for image generation (VDVAE, Versatile Diffusion) and multimodal feature extraction (CLIP), our method only required training ridge regression models from multivoxel brain activity to the relevant model latent spaces (Figures 3.1 and 3.2).

We analyzed the results both qualitatively (Figure 3.3) and quantitatively (Table 3.1) We observed that reconstructed scene images generated by Brain-Diffuser, although not perfectly identical to groundtruth images, preserve most of the layout and semantic information. They also appear more naturalistic compared to reconstructions from earlier studies (Figure 2.4). When evaluated quantitatively, we saw that Brain-Diffuser outperforms previous models in both high-level and low-level metrics. After advancing the state-of-the-art in image generation applications[202,189,159,209,255], it appears that latent diffusion models can also be used to improve the state-of-the-art in fMRI-based image reconstruction.

Although latent diffusion models are very recent[202], we noted at least two competing studies that used LDMs for fMRI-based image reconstruction. Chen et al.[31] proposed MinD-Vis, a method based on an LDM conditioned on image category labels (rather than text captions) to reconstruct images from the Kamitani dataset. As mentioned above, this is a less challenging, single-object-centered dataset; thus, their results cannot be directly compared with ours, obtained using the richer and more complex NSD dataset. Takagi et al.[234], on the other hand, used NSD and were thus included in our quantitative comparisons. There are multiple possible reasons why our model performed better than theirs, on both low-level and high-level metrics. Beyond the use of distinct pretrained LDMs (Stable Diffusion[202] for Takagi et al. vs. Versatile Diffusion[255] in our study), our framework contains several improvements such as the use of VDVAE reconstructions for low-level details (Figure 3.1) and the dual conditioning on CLIP-Vision and CLIP-Text features (Figure 3.2), which together resulted in better qualitative and quantitative results.

There are several ways in which this work may be pursued in the future. First of all, it will be important to test and validate our method on other image-fMRI datasets. As deep generative models will likely continue to improve at a breakneck pace, it is probable that there will soon come models better suited for complex scene reconstruction from fMRI signals. Of course, among a pool of many generative models, it may not be a trivial task to select the most appropriate ones and to experiment on them, and adapt them for brain decoding and image reconstruction. If future generative models reach a ceiling in their ability to linearly

explain brain activity, we may need to look for better alternatives than just doing ridge regression between fMRI patterns and latent variables. These alternatives (non-linear regressions, deep hierarchical networks), however, may require larger training datasets to learn the correspondence between fMRI patterns and visual features than ridge regression. It is possible that our reconstructions would benefit from including larger brain regions (or even the whole-brain) in our analysis. However, this is not guaranteed due to the possibility of overfitting in the presence of high-dimensional inputs. Also, expanding the region of fMRI inputs would dramatically raise the computational cost (in time and/or memory) of the training process of regression models. This is why the NSDGeneral ROI appears as an ideal compromise used in most studies (including ours). Using a common ROI also facilitates comparisons between studies. In the future, we may also see more accurate movie reconstruction studies that process temporal patterns together with spatial ones on movie-fMRI datasets[249,131]. Besides improving the reconstruction quality, future work could also design novel experiments and analyses on the NSD dataset using generative models. For instance, in this study, we have shown that we can use generative models to reveal the "optimal" stimulus for anatomically, functionally, or retinotopically-defined ROIs, by analyzing the reconstructions of synthetic fMRI patterns created from the corresponding ROI masks. This approach could easily be extended to probe less well-known regions of the visual cortex, to help settle theoretical arguments about distinct sub-regions of (e.g.) the face processing network, or to render images for arbitrary combinations of ROIs (e.g., what image would optimally activate V1, V4, and the face-ROI, but not V2 or the Body-

ROI; see also Figure 3.10). Important advances in this direction were made using an iterative optimization method by Gu et al (2022)[81] Directly passing synthetic fMRI patterns to the image reconstruction pipeline, however, is computationally advantageous, which may prove important when there are numerous combinations of sub-regions to be tried. Similar "virtual experiments" in this framework could help us address outstanding questions in neuroscience, and understand the organization of sensory and semantic knowledge in the brain.

3.2.6 Acknowledgements

This work was funded by the Agence Nationale de la Recherche ANR grants AI-REPS ANR-18-CE37-0007-01 and ANITI ANR-19-PI3A-0004. We thank Yu Takagi and Zijin Gu for sharing their test results with us and also we thank Alex Nguyen, Paul Scotti, and MedARC team for helping acquire the replicated results of the Mind-Reader study.

3.3 Epilogue to the main article:

In this chapter, we presented a two-stage reconstruction framework for reconstructing stimuli images in the Natural Scenes Dataset from fMRI patterns. In the first stage, we used the Very Deep VAE model for layout reconstruction, and in the second stage, we utilized the Versatile Diffusion model for the final reconstruction. The model's reconstruction results match the ground-truth images at both the low-level and semantic level while exhibiting realistic views. The model performed at a state-of-the-art level both qualitatively and quantitatively.

We conducted experiments to analyze the information captured by the model in terms of ROIs using regression weight percentile analysis and ROI-optimal stimuli visualization.

This work was accepted for publication in the Nature Scientific Reports journal after multiple revision cycles and was published in 2023.

Chapter 4

Discussions

This thesis presents frameworks for improving the performance of natural image and scene reconstruction by rethinking the problem and experimenting to find suitable generative models. The studies demonstrate better results in semantic and layout aspects of image reconstruction.

The implications of this development are significant. Before discussing those, let us elaborate on the main chapters' discussions.

4.1 Extended Discussion on Chapter 2

The work presented in Chapter 2 aimed at implementing a natural image reconstruction framework for the Generic Object Decoding dataset (fMRI recordings for single-object oriented images from the ImageNet dataset) using the Instance-Conditioned Generative Adversarial Network (IC-GAN) model. Prior studies focused on reconstructing low-level image features from fMRI data (except Mozafari et al.[155]). Our focus shifted to achieving a better trade-off between high-level and low-level features of reconstructed images, with the goal of increasing realism. We

demonstrated this achievement by presenting both qualitative and quantitative results. Our study concluded that it is possible to establish better frameworks for maintaining a well-balanced reconstruction of semantic and layout properties. These results influenced many studies that followed ours, such as Ferrante et al.[65], Chen et al.[31], Gu et al.[80], and Chen et al.[29].

We retained the regression models that had decent performance in decoding semantic and low-level features of brain signals. This allowed us to explore the information contained in these signals and the voxels in different brain regions at a finer level. Our findings are consistent with the neuroscience literature, which we briefly mentioned in Chapter 1, regarding the visual feature selectivity of ROIs.

4.1.1 Representational Power of Self-Supervised Learning and Prototype-Level Contrastive Learning

We should also emphasize on the representation power of the latent space of IC-GAN model. The success of the IC-GAN reconstruction model in Generic Object Decoding is not coincidental, particularly for high-level attributes. IC-GAN utilizes Swapping Assignments Between Views (SwAV)[24] model for instance conditioning, which mainly affects the semantic properties. The output of the SwAV model (referred to as instance features in Chapter 2) provides a powerful representation due to its training process. Unlike most contrastive learning methods, SwAV discriminates between clusters of images instead of individual images and learns prototype vectors assigned to these clusters.

The advantages of prototype-level contrastive learning approach are well ex-

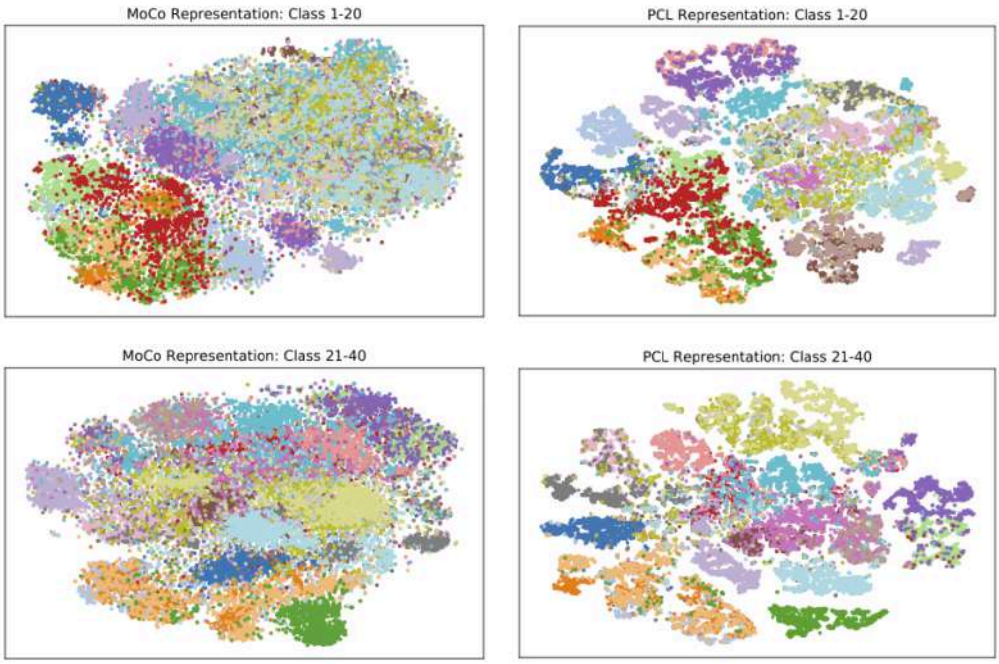

Figure 4.1: Comparison of T-SNE visualization of the unspervised learned representation embedding spaces of an instance-level contrastive learning model (MoCo) and prototype-level contrastive learning model (PCL) over 40 ImageNet categories (Figure from Li et al. 2020)[136].

Figure 4.2: Dataset of images of inanimate objects that were used as stimuli for recording fMRI activity in Konkle and Alvarez's study[125](Adapted from Konkle and Alvarez 2022)[125].

plained in a concurrent study to SwAV called Prototypical Contrastive Learning (PCL)[136]. Instance-level contrastive learning considers two samples as a negative pair only if they are from different images. This approach affects the representational properties of the embedding space, as some samples are pushed apart despite having similar semantic information. In prototype-level contrastive learning, samples with shared semantics converge around the prototype vectors, creating clusters for those shared semantic attributes. (An example is showed in Figure 4.1). Thus, prototype-level contrastive learning approaches outperform instance-based methods in terms of encoding semantic structures.

Recent studies conducted by Talia Konkle's lab have provided valuable insights into the relationship between the brain's representations and contrastive learning methods. One study showed that contrastive learning models, including SwAV ResNet-50, performed similarly to category-supervised models (ResNet-50 and AlexNet trained on ImageNet) in predicting fMRI activity for 72 inanimate object

images[125]. This result supports the idea that contrastive learning models may have more representational power than category-supervised learning counterparts. It is important to note that the evaluation on predicting fMRI activity for only 72 samples with simple object stimuli, without including the background as shown in Figure 4.2, may not be representative for a task such as reconstructing visual features of ImageNet images. In a sequential study, it was demonstrated that self-supervised learning models, such as SimCLR[30] (a contrastive learning model) and Barlow Twins[257] (a different self-supervised learning method showing similar results as SwAV in many tasks), performed slightly better (but with statistical significance) than category-supervised models in predicting brain activity for the Natural Scenes Dataset[126].

In another study, it was observed that category selectivity for faces, bodies, scenes, and visually presented words is present in the higher layers of self-supervised learning models. This phenomenon has also been observed in the brain in several studies, as discussed in Chapter 1. Lesioning these category-selective units causes deficits in predicting the relevant categories[181] (depicted in Figure 4.3). Although these results do not clearly prove the superiority of self-supervised learning models or contrastive learning models over category-supervised learning models in terms of having a more brain-like structure, they demonstrate why semantically accurate reconstructions of IC-GAN are not a coincidence. However, it may be possible to find a more suitable representational space using different models and training methods in the future. It is important to note that the ImageNet dataset is not representative of human-level categorization in terms of

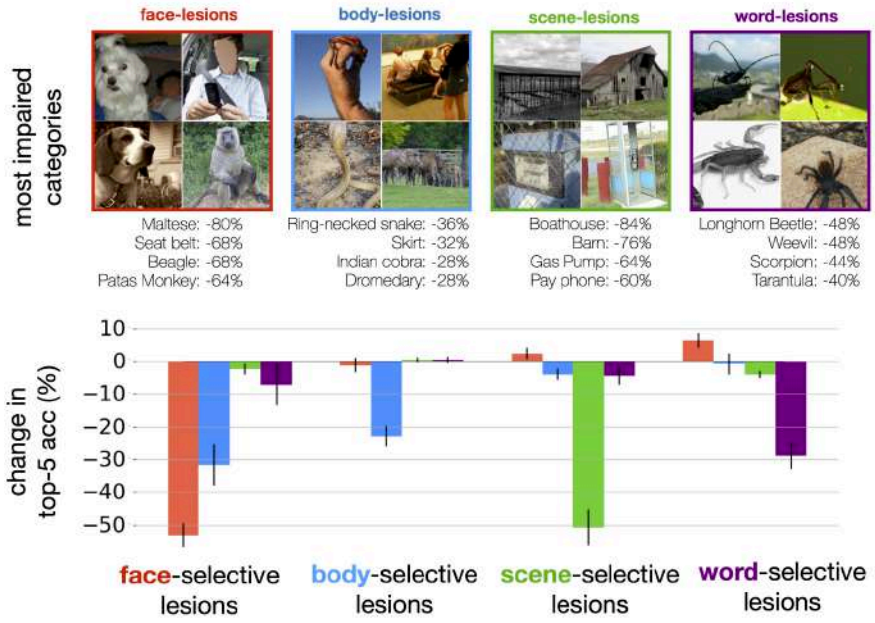

Figure 4.3: The effects of lesioning the units of a Convolutional Neural Network that was trained on a self-supervised learning regime. The figure presents examples of selective lesions in four main categories: face, body, scene, and words. The change in average accuracy for the ImageNet validation set's top-5 accuracy is presented below. (Adapted from Prince et al. 2023)[181].

Figure 4.4: Image reconstruction results of Shen et al.[217] on Generic Object Decoding Dataset for 5 subjects (left) and Deep Image Reconstruction Dataset for 3 subjects (right) (Adapted from Supplementary of Shen et al. 2019)[217].

granularity, as it includes 90 breeds of dogs as class categories among 1000 classes.

4.1.2 Confusion of Generic Object Decoding and Deep Image Reconstruction Datasets

During our studies on IC-GAN decoding, we observed a confusion made by the visual reconstruction studies. Some studies, such as Beliy et al.[14], Ren et al.[198], and Mozafari et al.[155], compared their model's results on the GOD dataset to the results of Shen et al.[217] on the DIR dataset. Both datasets had the same training and test stimuli, except for the artificial shapes and letters included in DIR. However, the studies differed in the number of subjects and fMRI repetitions. GOD has 5 subjects, while DIR has 3 subjects. In terms of fMRI repetitions, GOD has 1 repetition for training images and 35 repetitions for test images, whereas DIR has 5 repetitions for training images and 24 repetitions for test images. Although some

may not consider these differences vital, it is not standard practice to compare reconstruction results of these two datasets. The results appear similar in quality (as shown in Figure 4.4) but it is unclear how this will affect quantitative comparisons. Therefore, we believe that researchers working on visual reconstruction of GOD and DIR datasets should be aware of this issue.

4.2 Extended Discussion on Chapter 3

In Chapter 3, our objective was to develop a natural image reconstruction method that achieves high-fidelity reconstructions in terms of both low-level and high-level features, as we did in Chapter 2. However, this task was more challenging due to the complexity of the dataset, which consisted of natural scenes with multiple objects. To address this issue, we devised a two-stage scene reconstruction framework. The initial stage of the model, which utilized the Very Deep Variational Autoencoder, focused on reconstructing the low-level features and layouts of the stimuli. The second stage of the model, which utilized Versatile Diffusion, generated the final reconstruction by refining the initial reconstructions obtained from VDVAE using textual and visual features extracted with the CLIP model. Our goal was accomplished by demonstrating superior reconstruction quality compared to previous studies, both qualitatively and quantitatively. These findings have also influenced subsequent studies, such as Scotti et al.[213], Kneeland et al.[123], Liu et al.[140], Meng et al.[149], Xia et al.[254], Benchetrit et al.[15], and Sun et al.[231].

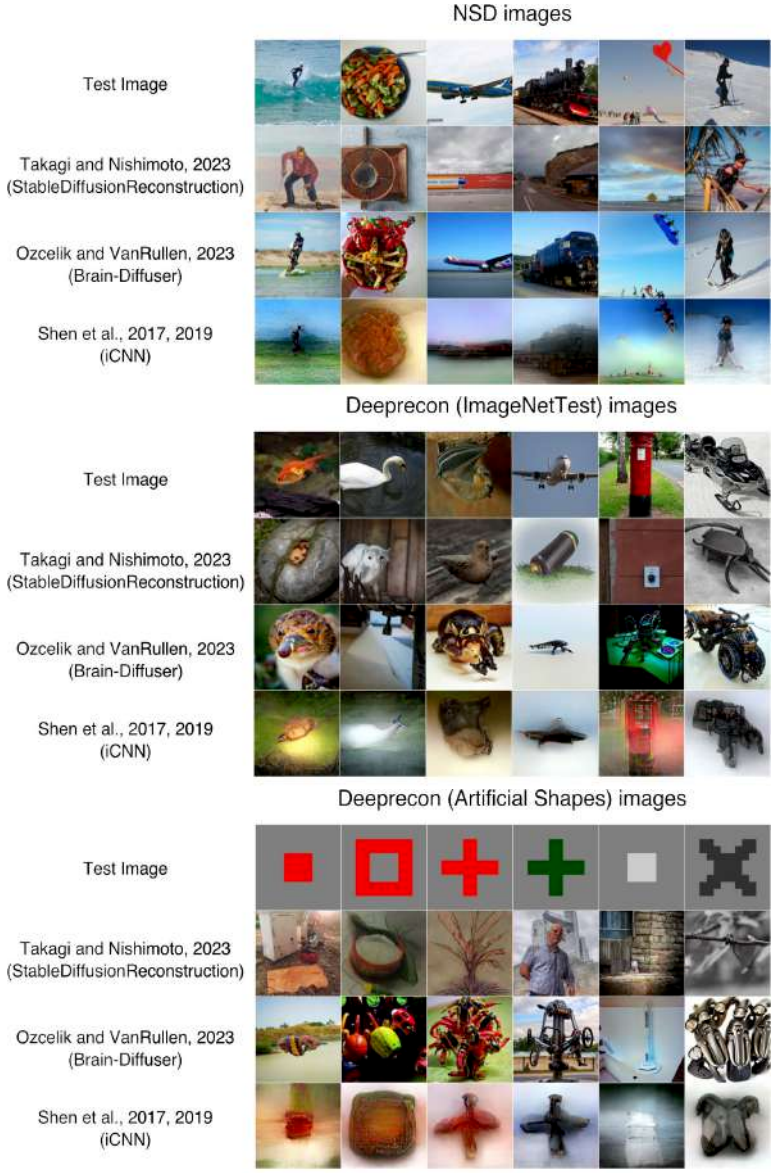

Figure 4.5: Image reconstruction results of Takagi et al.[234], Ozcelik et al.[171], and Shen et al.[217] on Natural Scenes Dataset (top), Deep Image Reconstruction Dataset (middle) and Artificial Shapes Dataset (bottom) (Adapted from Shirakawa et al. 2023)[218].

4.2.1 Response To Shirakawa's Criticisms on NSD Reconstruction

Ken Shirakawa, a PhD student from Kamitani Lab, criticized two studies on visual reconstruction of Natural Scenes Dataset,[218], including Takagi et al.[234] and our Brain-Diffuser[171]. The main criticism is that while these methods have demonstrated high performance on NSD, they are not as competitive on the Deep Image Reconstruction (DIR) dataset, indicating that they may not be generalizable to other visual reconstruction datasets, in contrast to Shen et al.[217]. Meanwhile, the authors criticize not only the reconstruction methods but also the structure of NSD. They argue that NSD lacks semantic diversity and that the categories between the train and test sets are not distinct, unlike the DIR dataset, which has no overlapping categories between the train and test sets.

It is evident that Shen et al.[217] continue to perform poorly across different datasets, while Takagi et al.[234] and Ozcelik et al.'s[171] performance is not as strong on other datasets as it is on NSD (presented in Figure 4.5). However, although it generalizes to all datasets, Shen et al.'s method captures different properties compared to ours. Their reconstructions resemble silhouettes and mainly utilize retinotopic information from early visual regions to capture low-level attributes, while our method aims to combine both high-level and low-level attributes for a full reconstruction. Diffusion-based models may underperform for several reasons. One reason is the semantic differences between the train and test images, which can affect models that capture semantic attributes but not those that do not (e.g., Shen et al.). It is important to note that semantic distinctness between train and test sets is not a necessary requirement for devising a neural decoding and visual

reconstruction task. Referring to this lack of distinctness as a 'semantic leak' may not be entirely accurate as it does not necessarily have a negative impact on the task. It is acceptable to have similar categories for both the train and test sets if the goal is to analyze semantic information using neural decoding and visual reconstruction methods. However, this may not be the only reason, as the IC-GAN framework performed well on the Generic Object Decoding dataset, which has the same category structures as DIR except for Artificial Shapes, despite having distinct categories for the train and test sets. Latent diffusion models are better at reconstructing complex scenes than previous models because of their ability to represent space in the latent dimension. The reconstruction performance and ability to capture semantic information from the ground truth also depend on condition features, such as CLIP features. Observing that IC-GAN performs well on the GOD dataset while Brain-Diffuser does not perform as well on DIR suggests that the difference may be caused by the relationship between the condition features and the datasets. The current use of CLIP may not be ideal for representing DIR images due to the high-dimensional and sparse representation space of CLIP, which may not be able to learn decent representations from a small training set. The dimensionality of IC-GAN SWaV features (2048-dim) is significantly different from that of non-projected CLIP features (257x768-dim), which may explain why IC-GAN performs well on the GOD dataset. We could suggest to experiment with a latent diffusion model trained with SWaV features to determine its effectiveness in reconstructing DIR and GOD datasets. It is also worth noting that Brain-Diffuser does not suffer from overfitting or memorization problems. This is

evident from ROI analyses, which show that it learns regression weights that are semantically meaningful.

4.3 General Discussion

This section will cover general topics related to the thesis. Chapters 2 and 3 present two main studies that use ridge regression for fMRI decoding and binary masks for generating ROI-preferred stimuli. Both of the studies focus on reconstructing both low-level and high-level attributes of the stimuli. The following discussion describes the motivations for the chosen methods and alternatives.

4.3.1 Using Ridge Regression for fMRI Decoding

In both studies, ridge regression was used, which is linear regression with L2 regularization. This approach is justified for several reasons. Firstly, the predictivity analysis[256] and representational similarity analysis[120,128], along with other studies exploring the correspondences between brain and deep neural network representations, briefly mentioned in section 1.6, suggest that the representations in the brain and deep neural networks are similar to each other due to hierarchical processing in both modalities. It is important to note that this similarity is not due to chance, but rather a result of the hierarchical structure of both systems. Therefore, linear models can accurately capture these relationships without the need for additional complexity introduced by nonlinearity when translating from one to the other. Also, there are practical reasons for using Ridge regression over non-linear regression. It is simpler and easier to interpret. When analyzing the

relationship between DNN features and brain activity, simplicity is valuable. The informative value of the features in terms of brain ROIs was analyzed using percentile analyses, which was possible due to the linear regression model. Another practical advantage of ridge regression is its ease of optimization compared to non-linear alternatives. Full-batch optimization was performed in both of our studies, which is theoretically helpful in finding a global minimum in the loss landscape of linear regression. Better optimization also means that it is less computationally demanding and requires fewer resources than non-linear methods. Ridge regression models are known to be less prone to overfitting, particularly when the amount of available data is limited. This is often the case in fMRI studies due to the cost and complexity of data collection. Finally, in our initial experiments, we did not observe a significant increase in performance when using non-linear regression models compared to linear models, despite encountering difficulties in optimization. Therefore, it is reasonable to choose linear regression as the simpler explanation, following the principle of Occam's Razor. The ridge regression models developed in both studies successfully captured attributes of the stimuli from the fMRI patterns.

Still, Scotti et al.[213] suggested that additional operations could be performed when translating fMRI patterns into DNN features. They used both non-linear multilayer perceptron (MLP) and diffusion to align fMRI patterns and DNN features. This alignment can be observed in a reduced dimension, as shown in Figure 4.6. Meanwhile, it can be observed that the models with only MLP backbone or MLP with projector do not perform better than Brain-Diffuser, but only com-

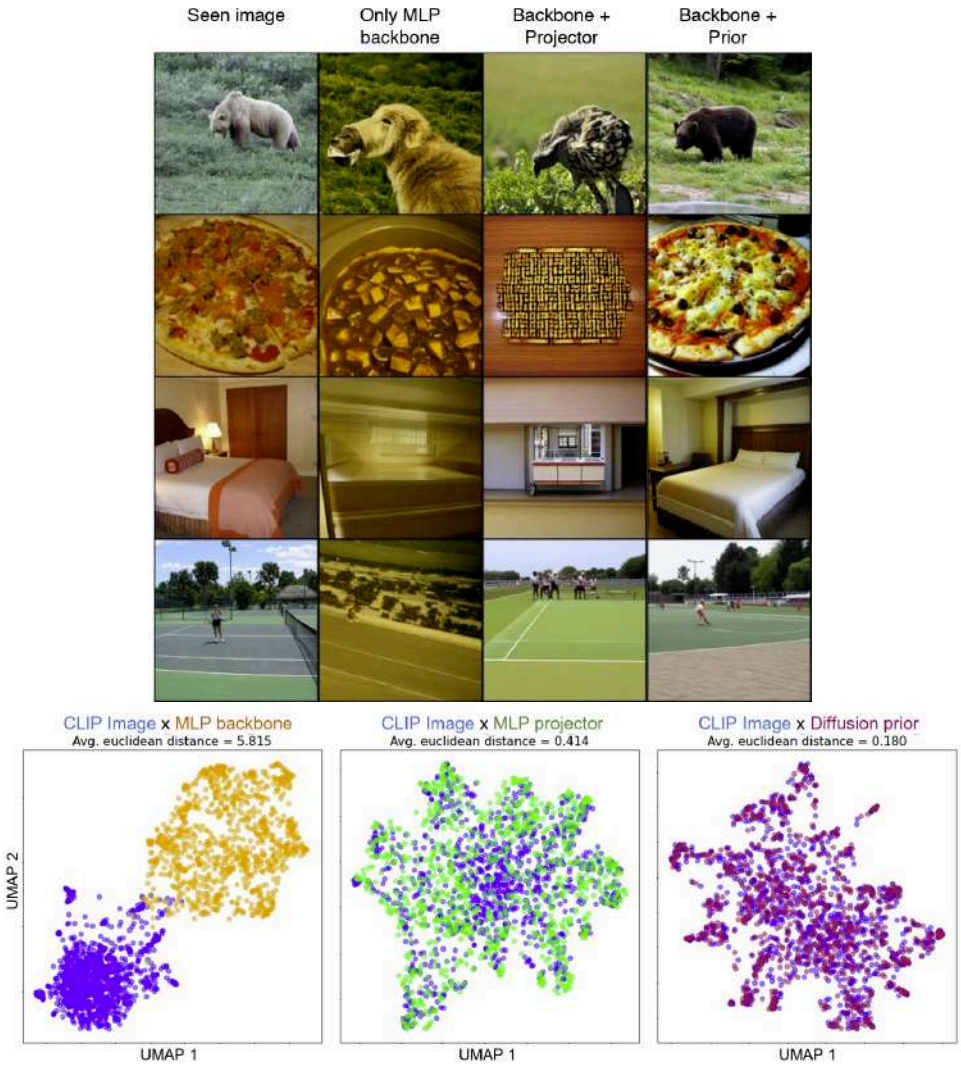

Figure 4.6: Example reconstruction for ablation models from Scotti et al.[213] above, presenting groundtruth images, nonlinear regression (MLP) reconstructions, MLP + projection reconstructions, and MLP + diffusion prior reconstructions respectively. Below, UMAP plots illustrating the increasing alignments between CLIP Image and predicted features, including MLP backbone, MLP projector and Diffusion prior respectively (Adapted from Scotti et al. 2023)[213].

pete with it when the diffusion prior is added. Therefore, the diffusion prior operation may have some effects beyond those of regression models (linear or non-linear). It is unclear whether the issue of disjoint embeddings is unique to CLIP features or if it applies to the features of different DNNs. This problem has only been studied in relation to CLIP features in Ramesh et al., where the diffusion prior was first introduced[189].

4.3.2 Using Binary Masks for Generating ROI-preferred Stimuli

For both of our studies, we made the decision to use binary masks to generate ROI-preferred stimuli in the semantic analysis sections. This method was chosen for several reasons. Firstly, it is easy to implement, as the only requirement is to create synthetic binary fMRI patterns with respect to the ROIs. The rest of the procedure is the same as for image reconstruction, where the fMRI patterns are passed to the reconstruction models. Secondly, the process is a single-pass, making it faster than iterative optimization alternatives like the NeuroGen framework[81] (depicted in Figure 4.7). While fast processing may not be critical for studying large ROIs, it is highly advantageous for studying semantics at a finer scale, such as voxels or clusters of voxels. It is evident that our method is capable of capturing semantic information despite its simplicity.

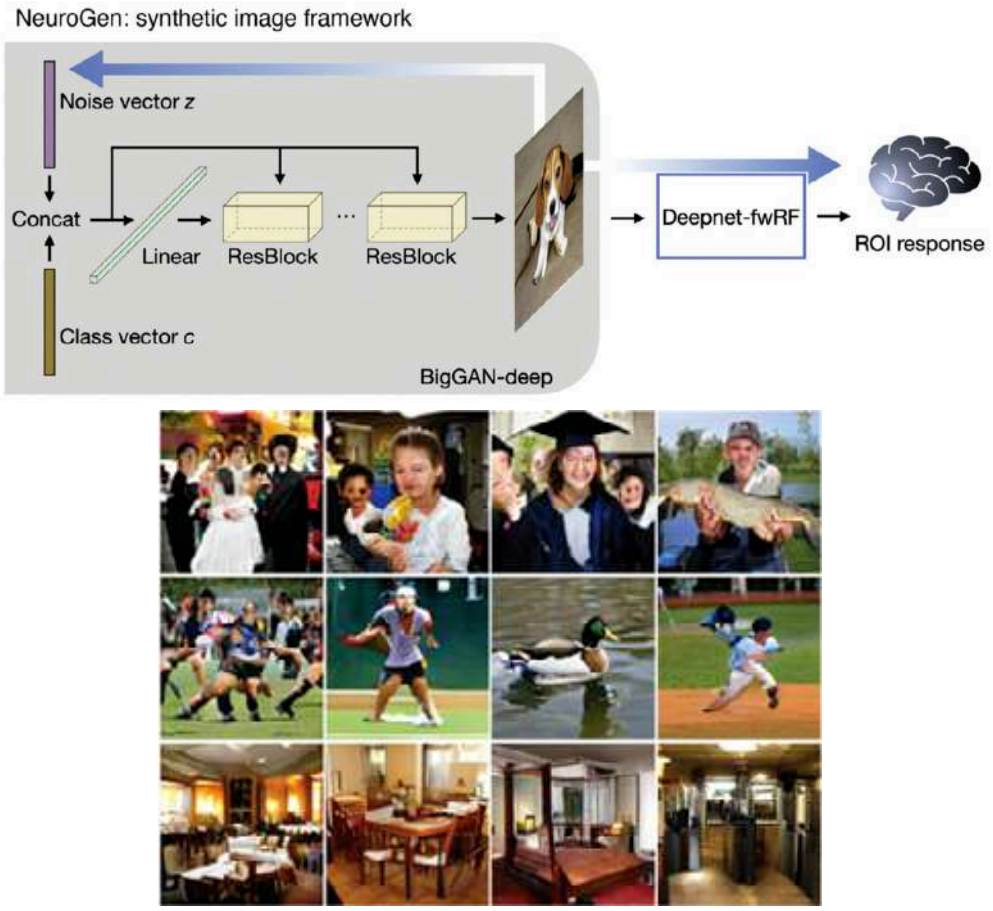

Figure 4.7: Illustration of NeuroGen framework (above). The NeuroGen framework produces images that prioritize regions of interest (ROIs) by optimizing the BigGAN-deep model through iterative processes using loss signals obtained from the Deepnet feature-weighted receptive (fwRF) encoding model. The following synthetic images were generated by NeuroGen, with each row representing a different ROI, namely FFA, EBA, and PPA (Adapted from Gu et al. 2022)[81].

4.3.3 Why Do Semantic Similarity and Natural Appearance Matter for Visual Reconstruction?

In both of our works, we aimed to create reconstructions that resemble the ground truth images in terms of both low-level (layout and position) and high-level (semantic) information, while maintaining a natural appearance. Our motivation for this goal came from our understanding of how visual processing occurs in the brain, as well as our study of the latent spaces of generative models. The brain processes visual features hierarchically, including both low-level and high-level features. Therefore, it is sensible to design frameworks for neural decoding that incorporate information from different hierarchies. A model that solely focuses on low-level features would not utilize all the information available on visual processing in the brain. Natural appearance of the images are also important when we are talking about natural image reconstruction. The natural appearance of reconstructed images is a key attribute when comparing their distribution to that of stimulus images. It is reasonable to assume that reconstructed images, which look like silhouettes of perceived images, would not be in the same distribution as the stimulus images. However, as the images become more natural-looking while retaining layout and semantic attributes, they move closer to the original distribution of the stimuli.

4.4 Comments on Following Studies and Future Directions

In this section, we will discuss some concurrent or subsequent studies that explore topics relevant to ours. Visual scene reconstruction models are still being developed and are becoming more accurate in terms of different properties with new research[213,123,254]. In addition to improving reconstruction performance, neural encoding and decoding models are adopted for different purposes on the Natural Scenes dataset.

In both Chapter 2 and 3, we emphasized the potential of visual reconstruction models, as well as neural encoding and decoding models in general, for neuroscientific exploration. We demonstrated proof-of-concept results with semantic analyses in both studies. Since NSD became publicly available, several studies have been conducted in this area.

Several studies have explored the general computational modeling properties of neural networks using NSD. Wang et al. demonstrated that a multimodal CLIP model (ResNet50 backbone) trained jointly was more effective in predicting responses in high-level visual brain regions compared to ResNet50 trained only with ImageNet images or BERT trained only with text data[248]. This result may explain why LDM models that use CLIP features are effective in reconstructing NSD images. Finzi et al. investigated the emergence of visual streams using various types of deep neural networks with different training objectives and constraints. The computational models used were multi-task-trained models, as well as supervised and self-supervised models trained with local spatial con-

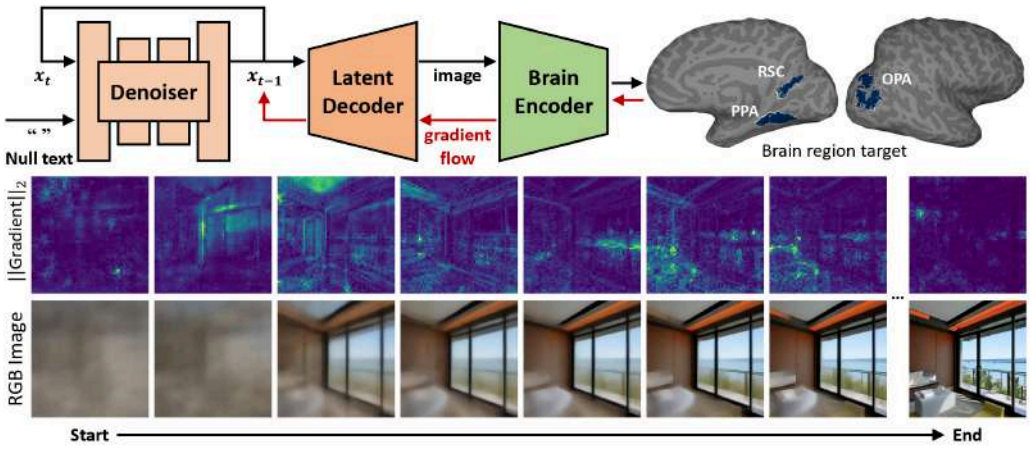

Figure 4.8: BrainDiVE framework generating an image from the fMRI activity of scene-selective regions (RSC, PPA, and OPA) (Figure from Luo et al. 2024)[143].

straints, which are referred to as Topographic Deep Artificial Neural Networks. The study demonstrated that self-supervised TDANN models outperformed multi-task-trained models in capturing the brain's spatial segregation and functional organization[67].

Jain et al.[109] and Khosla et al.[121] used neural network models on the Natural Scenes Dataset to discover category-selective areas for food in the human visual cortex in a hypothesis-free manner. In their initial experiments, these studies used statistical methods instead of neural network models. However, they later conducted further analyses using the features gathered from neural network models, specifically CLIP. Additionally, some studies directly employed neural networks to explore cortex selectivity. Sarch et al. established the Brain Dissection method, drawing inspiration from the works of David Bau, including Network Dissection[11] and GAN Dissection[12]. The authors extracted voxel-specific feature maps from convolutional neural networks to analyze the spatial correspondence between vox-

els and various spatial properties, such as depth, surface normals, curvature, and shading[210]. While Brain Dissection uses encoding approach for their analyses, Brain Diffusion for Visual Exploration (BrainDiVE)[143] and Semantic Captioning using Brain Alignments(BrainSCUBA)[144] studies use decoding methods to explore the selectivity of the regions. BrainDiVE utilizes latent diffusion models to generate images with brain activation guidance, similar to our Brain-Diffuser method (depicted in Figure 4.8). They show the broad category-selective networks found in previous studies, including faces, places, bodies, words, and food. They also provide detailed analyses within specific regions of interest (ROIs), such as separate clusters for indoor and outdoor scenes (although not entirely exclusive).[143]. In contrast to the image generation technique of BrainDiVE, BrainSCUBA uses text generation to create voxel-wise semantic captions that describe the semantic selectivity of the voxels[144]. Although these studies present results on cortical selectivity exploration using fMRI data via neural network decoding, we believe that this area is not fully utilized and that more studies may be conducted in the future.

A potential area of future study is the reconstruction of visual thought using imagery data. Mental imagery is defined as representations of sensory information without a direct external stimulus by Pearson et al[172]. Meanwhile, Kosslyn et al. referred to visual mental imagery as "seeing" in the absence of the appropriate immediate sensory input[127]. Mental imagery is not solely based on introspective reports. In fact, neuroimaging methods allow for the observation of neural representations of mental imagery[157]. Although Chapter 1 presents studies on mental

imagery[195,100,124], applying visual reconstruction models and semantic exploration techniques to imagery data can reveal differences between imagery and visual perception[53]. This approach can also help examine the diversity of vividness between subjects[70], and the role of ROIs in mental imagery in more detail[52]. Therefore, we can develop these decoding techniques to be more robust in decoding imagery data by focusing on this information. Although the Natural Scenes Dataset contains imagery data, this portion of the data has not yet been made publicly available.

4.5 Practical Applications and Ethical Implications

Due to recent advancements in neural decoding research, real-world applications of brain-computer interfaces (BCIs) have become possible. BCIs are frameworks that enable direct communication between the brain and external devices[72]. BCI systems are typically maintained by acquiring signals through either invasive (e.g. intracortical electrodes and ECoG) or non-invasive (e.g. EEG, fNIRS, MEG, and fMRI) neuroimaging techniques. The choice of neuroimaging technique depends on the requirements of the BCI application. There are numerous clinical and non-clinical applications of EEG-based BCI due to its mobility and non-invasiveness. For instance, motor and stroke rehabilitation applications utilize motor imagery and execution tasks for clinical cases[208]. Examples of non-clinical applications of BCI include controlling robots, playing video games, and operating quadcopters, among others[208]. Some clinical applications require invasive methods, such as electrocorticography (ECoG), for more robust signal acquisition. For instance, a study on speech decoding was conducted on a patient with mo-

tor disorders, such as amyotrophic lateral sclerosis (ALS), using high-resolution ECoG neural recording[58]. This study can lead to the development of high-quality neural speech prostheses. Although fMRI is not a commonly used method due to its immobility, expense, and low temporal resolution, it is still utilized for certain clinical cases. For instance, it can be used to obtain responses from patients with disorders of consciousness, such as minimally conscious state, vegetative state, or locked-in syndrome[201,169]. It has been observed that some patients who meet the criteria for being in a vegetative state were able to perform two imagery tasks. During the study, participants exhibited significant activity in the Supplementary Motor Area (SMA) when asked to imagine playing tennis. Similarly, the Parahippocampal Place Area was activated when participants were instructed to imagine walking around their house[169]. As previously mentioned, fMRI-based neural decoding and visual reconstruction are best suited for investigating the fine-grained functional organization and neural coding of brain regions, which is necessary for developing more robust and precise BCI applications.

Although BCI applications can be useful, the ability to decode finer information from brain signals raises ethical questions about potential misuse and violation of mental privacy. Decoding neural processes and thoughts with precision may violate mental privacy, as depicted in dystopian scenarios like George Orwell's "Nineteen Eighty-Four"[206], or in films and TV shows such as "Minority Report" and "Black Mirror". The legal implications of using neuroscientific techniques for tasks such as lie detection, predicting criminal behavior, and rehabilitating criminals are explored in the field of neurolaw[25]. As AI and Neuroscience continue

to develop rapidly, it is important to discuss the ethical concerns surrounding the reliability of neural decoding techniques for critical applications, as well as issues of justice, safety, and privacy. It is also important to raise awareness about potential misuses while avoiding alarmism and conspiracy theories[146].

4.6 Conclusion

This thesis presents two frameworks for visual image reconstruction from fMRI activity using deep generative models. The frameworks focus on different aspects of visual reconstruction compared to their predecessors, such as semantic coherence and realism. They exhibit superior results compared to other methods, both qualitatively and quantitatively. In addition to their ability to reconstruct stimuli, these frameworks have demonstrated their usefulness for neuroscientific exploration through region-of-interest analyses. The contributions of this thesis to the field of neural decoding and visual reconstruction may provide guidance for researchers working in cognitive neuroscience and brain-computer interfaces.

4.7 Closing Thoughts

I would like to conclude with a few remarks. Jack Gallant compared neural decoding to building a dictionary for a foreign language, in this case the language of the brain (quoted in Section 1.4). This analogy is particularly relevant to early studies of neural decoding. However, the use of deep learning models for this task alters the analogy slightly. We see a glimpse of the brain's language by examining fMRI patterns, although this is an indirect method that uses blood oxygenation.

Additionally, a language can be found in the latent variables of deep generative models, which are low-dimensional representations of the data. It is possible to translate between these representations by encoding and decoding using regression models. Both the language of the brain and the language of deep generative models can be alien to the human eye. Fortunately, the generator component of deep generative models transforms latent variables into images that are comprehensible to humans. Why is this wonderful? As humans, our ability to interpret the language of the brain is limited by its complexity. When inspecting fMRI data, humans can easily distinguish basic concepts such as animate versus inanimate or face versus non-face but challenges arise with more specific features. However, deep learning models do not have this limitation. They excel at recognizing patterns when given enough data. Therefore, they offer an opportunity to analyze the language of the brain in greater detail than was previously possible.

Chapter 5

Summary in French

5.0.1 Chapitre 1 : Introduction

L'esprit humain a été un sujet de fascination pour les philosophes et les scientifiques tout au long de l'histoire. Platon et Aristote ont présenté des points de vue divergents sur l'âme et l'esprit, Platon soutenant le dualisme et Aristote suggérant que l'âme est la forme du corps[141]. Cette perspective dualiste a influencé la pensée occidentale, notamment les travaux de Thomas d'Aquin et de Descartes[200]. Bien que le dualisme de Descartes ait été débattu, les progrès de la psychologie et des neurosciences ont entraîné un déclin de la popularité du dualisme[37]. Au 20e siècle, diverses théories sur l'esprit et la conscience ont vu le jour. La théorie psychanalytique de Freud met l'accent sur les processus inconscients[69], tandis que les behavioristes comme Watson et Skinner se concentrent sur le conditionnement sans référence aux états mentaux[251,221]. Le matérialisme, qui explique l'esprit par des processus physiques, a gagné du terrain, avec des partisans anciens comme Démocrite et modernes comme Dennett et les Church-

lands[222,49,191,34]. David Chalmers a critiqué le matérialisme en introduisant les problèmes "faciles" et "difficiles" de la conscience[28]. Le fonctionnalisme, introduit par Hilary Putnam, a mis l'accent sur les fonctions cognitives plutôt que sur la constitution interne et a influencé la recherche sur l'intelligence artificielle[182]. Les neurosciences se sont de plus en plus intégrées aux sciences cognitives, donnant naissance à des théories telles que les corrélats neuronaux de la conscience[41], la théorie de l'espace de travail global[7] et la théorie de l'information intégrée[238]. Les études sur la perception humaine ont été au cœur de ces efforts interdisciplinaires. La vision et la vue peuvent sembler simples, mais notre perception est façonnée par des représentations mentales plutôt que par un reflet direct du monde extérieur. Alhazen a reconnu le rôle de l'inférence inconsciente dans la perception visuelle[215], un concept qui a été défini plus tard par Hermann von Helmholtz pour décrire la manière dont le cerveau interprète les données sensorielles[46]. Les recherches neuroscientifiques menées aux XIXe et XXe siècles ont démontré que notre système nerveux construit des représentations du monde extérieur, ce qui confirme l'idée d'Emmanuel Kant selon laquelle la perception est subjective[115]. Différents aspects de la vision sont affectés par les lésions cérébrales, ce qui entraîne des pathologies telles que l'agnosie visuelle (incapacité à reconnaître des objets), la prosopagnosie (incapacité à reconnaître des visages) et la vision aveugle (réponse à des stimuli visuels sans perception consciente)[207,35]. Un autre trouble, la négligence, se traduit par la méconnaissance d'un côté du champ visuel à la suite d'une lésion cérébrale[199]. Ces cas mettent en évidence le mécanisme de liaison du cerveau, qui intègre différents aspects de la perception visuelle, un sujet qui

fait encore l'objet de discussions entre neuroscientifiques, scientifiques cognitifs et philosophes[66].

La vision commence lorsque la lumière pénètre dans l'œil, traverse la cornée et le cristallin et atteint les photorécepteurs de la rétine (bâtonnets pour la faible luminosité, cônes pour la couleur). Les signaux émis par les photorécepteurs sont transmis au cerveau par le nerf optique, plus précisément au cortex visuel primaire (V1) par l'intermédiaire du chiasma optique et du noyau géniculé latéral (LGN)[55]. Les cellules ganglionnaires de la rétine réagissent à des motifs lumineux spécifiques, en mettant l'accent sur les contrastes. Cette structure est maintenue dans le LGN et V1. V1 contient des cellules simples et complexes qui répondent respectivement à l'orientation des bords et aux motifs abstraits. Le V1 et le LGN préservent l'ordre spatial des stimuli visuels grâce à la cartographie rétinotopique[74]. Le traitement visuel passe par V2, V3 (angles, mouvements) et V4 (couleurs, formes). Le système se divise en deux parties : la voie dorsale ("où/-comment"), qui traite les informations relatives à la localisation et à l'action, et la voie ventrale ("quoi"), cruciale pour la catégorisation des objets[177]. Le traitement visuel implique des régions cérébrales spécialisées au-delà de V1-V4. Le complexe occipital latéral (LOC) reconnaît les objets, les lésions entraînant une agnosie[78]. Le Centre des couleurs (CC) traite les couleurs ; les lésions de cette région entraînent l'achromatopsie[142,9]. L'aire temporale médiane visuelle (MT/V5) détecte les mouvements, les lésions provoquant l'akinétopsie[3,17]. L'aire fusiforme des visages (FFA) identifie les visages et des lésions entraînent une prosopagnosie[116,77]. L'aire parahippocampique des lieux (PPA) encode les dispositions spatiales, ce qui af-

fecte la navigation en cas de lésion[62,94]. L'aire extrastriée du corps (EBA) réagit aux images du corps et aide à la planification de l'action[56]. L'aire visuelle de la forme des mots (VWFA) reconnaît les mots écrits et est liée au traitement du langage[147]. Les études sur les lésions et la neuro-imagerie ont révélé ces fonctions, montrant des représentations détaillées d'objets et de concepts dans le cerveau.

Jusqu'au XXe siècle, il était impossible d'enregistrer l'activité cérébrale, car les chercheurs ne pouvaient étudier que la structure anatomique des cerveaux morts[35]. Le développement des méthodes de neuro-imagerie fonctionnelle a changé la donne, en permettant la détection de signaux provenant de cerveaux vivants. Les méthodes invasives telles que l'enregistrement monocellulaire et l'électrocorticographie (ECoG) mesurent l'activité cérébrale en plaçant des électrodes directement sur le cerveau, ce qui permet d'obtenir une résolution spatiale et temporelle élevée[74]. Toutefois, les méthodes non invasives telles que l'EEG, la MEG, la TEP, l'IRMf et la fNIRS sont préférées en raison de leur risque et de leur coût moindres[19]. L'EEG et le MEG mesurent respectivement l'activité électrique et magnétique du cerveau, tandis que le PET, le fNIRS et l'IRMf mesurent les changements métaboliques[134]. L'IRMf, développée par Seiji Ogawa et ses collègues, est particulièrement remarquable pour sa haute résolution spatiale et sa nature non invasive, ce qui la rend idéale pour les tâches visuelles[164]. Elle mesure l'activité cérébrale par le biais des signaux BOLD (blood-oxygen-level-dependent), qui sont des mesures indirectes de l'activité neuronale basées sur le rapport entre l'hémoglobine oxygénée et l'hémoglobine désoxygénée dans le sang[44]. Bien que la résolution temporelle de l'IRMf soit inférieure à celle de l'EEG, sa résolution spatiale supérieure la rend

précieuse pour les études sur l'activité cérébrale[134].

La découverte par Nancy Kanwisher et ses collègues de régions sélectives dans le lobe temporal pour différents objets a suscité des débats sur la localisation des fonctions cérébrales. Isabel Gauthier a remis en question l'idée que la zone fusiforme du visage (FFA) était une "zone du visage", proposant qu'il s'agisse plutôt d'une "zone d'expertise"[178]. James Haxby a présenté une méthode de décodage, suggérant que les représentations des visages et des objets dans le cortex temporal ventral sont distribuées et se chevauchent[90]. Malgré certaines remises en question des affirmations de Haxby, d'autres études utilisant des techniques telles que les machines à vecteurs de support ont montré que les régions sélectives par catégorie pouvaient différencier les objets non préférés, mais pas parfaitement. Cela a conduit au développement du décodage neuronal, une méthode permettant de prédire des informations sur des stimuli à partir de signaux cérébraux. Les premières études ont utilisé l'analyse multivariée des formes (MVPA) pour décoder les propriétés des stimuli visuels[89]. Les chercheurs ont réussi à reconstruire des formes simples et des images naturelles à partir de signaux IRMf[154], ce qui a conduit à des modèles capables de décoder des images mentales[195]. Le décodage visuel initial s'est avéré prometteur mais a rencontré des limites, ce qui a incité à utiliser des modèles d'apprentissage profond pour obtenir des représentations plus riches et améliorer la qualité de la reconstruction.

Les origines de l'apprentissage profond (DL) et des réseaux neuronaux profonds (DNN) remontent aux premières recherches en neurosciences computationnelles et en IA. Les principales étapes de l'IA connexionniste et de l'apprentissage au-

tomatique comprennent le modèle de neurones McCulloch-Pitts[148], l'apprentissage hébbien[93], les perceptrons[204], la rétropropagation[253], les réseaux Hopfield[99], les machines de Boltzmann[97] et les perceptrons multicouches[205]. Malgré ces progrès, l'IA symbolique a dominé certaines périodes en raison de son succès dans des tâches telles que les échecs et la planification. Cependant, elle s'est heurtée à des difficultés dans des tâches telles que la reconnaissance d'objets, pour lesquelles des modèles d'apprentissage automatique tels que les machines à vecteurs de support et les réseaux neuronaux convolutifs (CNN) ont été utilisés, mais sont restés en deçà des performances humaines. La percée s'est produite en 2012 avec AlexNet, un CNN profond qui a surpassé de manière spectaculaire les modèles précédents dans le défi ImageNet, déclenchant ainsi la révolution de l'apprentissage profond[129]. Les modèles suivants, tels que VGG[220], GoogLeNet[232] et ResNet[92], ont continué à améliorer les performances dans diverses tâches, notamment la classification, la détection et la segmentation d'images. Les modèles génératifs tels que les autoencodeurs variationnels (VAE)[122], les réseaux adversariaux génératifs (GAN)[75] et les modèles autorégressifs ont suivi[242], chacun présentant des forces et des limites uniques. Le développement de modèles Transformers a révolutionné le traitement du langage naturel (NLP) et a ensuite eu un impact sur la vision par ordinateur[246]. Des techniques telles que l'apprentissage contrastif ont permis d'améliorer encore les capacités des modèles[185]. Les innovations récentes comprennent les modèles de diffusion et les modèles de diffusion latente, illustrés par le modèle de diffusion stable, qui permettent une génération d'images efficace et de haute qualité[202]. Cet historique ouvre la voie à l'exploration des principes

fondamentaux des modèles d'apprentissage profond.

Les réseaux neuronaux profonds (RNP) se composent de plusieurs couches qui modélisent des modèles de données complexes. La couche d'entrée reçoit les données, qui sont traitées par les couches cachées à l'aide de matrices de poids et de fonctions de non-linéarité telles que ReLU. La couche de sortie prédit les valeurs cibles et les performances du réseau sont optimisées à l'aide de la rétropropagation et d'algorithmes tels que la descente stochastique de gradient (SGD)[160]. Les réseaux neuronaux convolutifs (CNN), conçus pour les données en grille telles que les images, utilisent des couches de convolution pour extraire les caractéristiques et des couches de mise en commun pour réduire la taille spatiale, en évitant l'ajustement excessif et en gérant les paramètres de manière efficace[117]. Les modèles génératifs profonds, y compris les autoencodeurs variationnels (VAE) et les réseaux adversariaux génératifs (GAN), génèrent des données en apprenant les distributions sous-jacentes[68]. Les VAE utilisent des projections probabilistes pour les espaces latents, tandis que les GAN se composent d'un générateur et d'un discriminateur qui se font concurrence pour produire des données réalistes. Les modèles de diffusion latente (LDM) affinent itérativement les données à partir du bruit pour générer des échantillons de haute fidélité[202]. Ces modèles ont été améliorés par les progrès de l'apprentissage profond, ce qui a conduit à des applications dans la classification d'images, la détection d'objets et la génération de texte à partir d'images. Des modèles récents tels que la diffusion stable et la diffusion polyvalente démontrent l'intégration d'entrées et de voies multimodales, reflétant l'évolution et l'impact de l'apprentissage profond sur diverses tâches d'intelligence

artificielle.

Les modèles d'apprentissage profond, enracinés dans les neurosciences computationnelles, s'inspirent des neurosciences malgré certaines critiques préconisant des approches plus techniques[112] ou symboliques[145]. La collaboration entre les neurosciences et la recherche en IA se poursuit, les connaissances neuroscientifiques contribuant au développement de modèles d'IA robustes et les modèles d'IA étant utilisés pour le décodage neuronal, l'analyse de l'imagerie cérébrale et l'analyse comportementale. À la suite du succès d'AlexNet dans la reconnaissance d'objets, les chercheurs ont commencé à explorer les similitudes entre les modèles d'apprentissage profond et les fonctions cérébrales, en particulier dans l'apprentissage des caractéristiques de haut niveau. Les réseaux neuronaux profonds (RNP) ont montré leur capacité à prédire l'activité neuronale dans les zones visuelles moyennes comme V4 et le cortex temporal inférieur (IT), responsable du traitement visuel de haut niveau[256,82]. Ces analyses ont été réalisées à l'aide de diverses données de neuro-imagerie, notamment l'électrophysiologie[20], l'IRMf[60,252,59], la MEG[36] et l'EEG[76]. Les DNN sont également utilisés pour décoder les caractéristiques des stimuli visuels et reconstruire les images[252,100]. Notre recherche se concentre sur l'exploitation de l'IA et de l'apprentissage profond pour le décodage neuronal et la reconstruction visuelle en neurosciences.

Les chercheurs ont commencé à utiliser des modèles génératifs profonds pour reconstruire des images à partir de signaux IRMf, en tirant parti des capacités de représentation des DNN en neurosciences. Les premières études de Du et al. ont utilisé divers modèles pour reconstruire des formes, des lettres et des chiffres

à partir de modèles d'IRMf[57]. L'attention s'est ensuite portée sur des images plus naturelles, telles que les visages. Des ensembles de données IRMf accessibles au public, comme l'ensemble de données Generic Object Decoding (GOD)[100] et l'ensemble de données Natural Scenes Dataset (NSD)[2], ont facilité cette recherche. L'ensemble de données GOD comprend des images provenant de l'ensemble de données ImageNet, tandis que l'ensemble de données NSD comprend des images provenant de l'ensemble de données COCO, avec des enregistrements de plusieurs sujets. Plusieurs modèles ont été développés pour la reconstruction d'images à partir de signaux IRMf, notamment DCGAN[214], EBGAN[229] et BigBiGAN[155]. Nous avons développé le décodage IC-GAN (Instance-Conditioned GAN) (chapitre 2) et Brain-Diffuser (chapitre 3) pour capturer les caractéristiques sémantiques et de bas niveau des images dans ces ensembles de données.

5.0.2 Chapitre 2 : Décodage GAN conditionné par instance

La reconstruction d'images naturelles perçues à partir de signaux IRMf est l'un des sujets les plus intéressants de la recherche sur le décodage neuronal. Les études antérieures ont réussi à reconstruire soit les caractéristiques de bas niveau de l'image, soit les aspects sémantiques/de haut niveau, mais rarement les deux. Dans le deuxième chapitre, nous avons utilisé un modèle GAN conditionné par instance (IC-GAN) pour reconstruire des images à partir de modèles IRMf avec des attributs sémantiques précis et des détails de bas niveau préservés. Le modèle IC-GAN prend en entrée un vecteur de bruit de 119 dim et un vecteur de caractéristiques d'instance de 2048 dim extrait d'une image cible via un modèle

d'apprentissage auto-supervisé (SwAV ResNet-50) ; ces caractéristiques d'instance agissent comme un conditionnement pour la génération d'images IC-GAN, tandis que le vecteur de bruit introduit de la variabilité entre les échantillons. Nous avons formé des modèles de régression ridge pour prédire les caractéristiques d'instance, les vecteurs de bruit et les vecteurs denses (la sortie de la première couche dense du générateur IC-GAN) des stimuli à partir des modèles IRMf correspondants. Nous avons ensuite utilisé le générateur IC-GAN pour reconstruire de nouvelles images de test basées sur ces variables prédites par l'IRMf. Les images générées ont présenté des résultats de pointe en termes de capture des attributs sémantiques des images de test originales tout en restant relativement fidèles aux détails de bas niveau de l'image. Enfin, nous utilisons le modèle de régression appris et le générateur IC-GAN pour explorer et visualiser systématiquement les caractéristiques sémantiques qui stimulent au maximum chacune des régions d'intérêt du cerveau humain.

Nous avons utilisé un modèle GAN conditionné par instance (IC-GAN)[26], pré-entraîné sur l'ensemble de données ImageNet, pour la génération d'images naturelles[48]. IC-GAN est flexible et peut être appliqué à différents GAN backbones, tels que StyleGAN[118] ou BigGAN[18]. Contrairement aux GAN traditionnels, qui utilisent des étiquettes de classe pour le conditionnement, IC-GAN utilise des caractéristiques d'instance extraites d'un extracteur de caractéristiques pré-entraîné, tel que SwAV avec un backbone ResNet-50[24]. Ces caractéristiques d'instance, ainsi qu'un vecteur de bruit, sont utilisés pour conditionner le générateur. Le modèle IC-GAN que nous avons utilisé repose sur une architecture BigGAN à

7 couches, générant des images de taille 256x256x3 [18]. Pendant l'entraînement, nous avons optimisé le vecteur de bruit pour chaque image en utilisant la stratégie d'évolution de l'adaptation de la matrice de covariance (CMA-ES)[85], dans le but de faire correspondre l'image générée à l'image originale en termes de structure spatiale et d'attributs sémantiques. Pour la reconstruction d'images basée sur l'IRMf, nous avons entraîné des modèles de régression ridge pour prédire les variables latentes (caractéristiques d'instance, vecteurs de bruit et vecteurs denses) à partir des modèles d'IRMf. Ces variables prédites ont ensuite été utilisées pour générer des reconstructions d'images avec le générateur IC-GAN, capturant à la fois les propriétés de haut niveau et de bas niveau des images originales.

Les reconstructions d'images de notre méthode démontrent que IC-GAN peut capturer efficacement les attributs sémantiques des images testées, bien qu'il manque souvent certains détails visuels. Les reconstructions basées sur des vecteurs latents optimisés révèlent le potentiel d'IC-GAN si les variables latentes sont parfaitement décodées à partir des modèles IRMf. Notamment, le BigBiGAN de Mozafari et al. capture mieux les propriétés de haut niveau mais manque encore certains détails[155]. Notre méthode équilibre le réalisme et la précision sémantique, comme le confirment les comparaisons quantitatives. IC-GAN dépasse les autres méthodes dans les mesures sémantiques de haut niveau (Inception et distances CLIP) et correspond aux mesures de bas niveau (Pix-Comp, SSIM) en utilisant des variables latentes décodées par le cerveau. Le modèle IC-GAN complet améliore la capture des détails de bas niveau mais reste légèrement inférieur à certaines méthodes précédentes, tout en conservant une reconstruction sémantique de haut

niveau supérieure.

Notre modèle de décodage cérébral, qui utilise l'espace latent du réseau IC-GAN, permet de reconstruire le contenu de haut niveau des images perçues mieux que les méthodes précédentes, tout en conservant plus de détails de bas niveau que d'autres. Du point de vue des neurosciences, ce modèle nous aide à comprendre le codage neuronal de l'information visuelle. En comparant la norme L_1 des poids de régression ridge pour les caractéristiques d'instance par rapport aux vecteurs denses dans sept régions d'intérêt du cerveau (ROI), nous avons constaté que les régions cérébrales inférieures (V1-V3) étaient plus informatives sur les vecteurs denses, tandis que les régions supérieures (V4, LOC, FFA, PPA) contenaient plus d'informations sur les caractéristiques d'instance.

Nous avons également visualisé le stimulus "optimal" pour chaque région du cerveau en synthétisant des modèles IRMf et en les faisant passer par nos modèles de régression entraînés pour prédire les variables latentes, qui ont ensuite été utilisées pour générer des images avec le générateur IC-GAN. Cette approche diffère des méthodes précédentes en évitant l'optimisation itérative. Les images générées pour le cortex visuel inférieur (V1-V2) présentaient des textures de base, tandis que les régions supérieures (V4, LOC, FFA, PPA) produisaient des images plus structurées, des objets et des visages, s'alignant sur les sélectivités connues. Par exemple, la FFA produisait des visages humains et animaux, et la PPA des scènes détaillées. Ces résultats, conformes à la littérature neuroscientifique, montrent que notre modèle peut visualiser directement la sélectivité des caractéristiques visuelles dans le cerveau, améliorant ainsi notre compréhension du codage neuronal.

Dans le deuxième chapitre, nous avons présenté une méthode pour la reconstruction d'images naturelles à partir de modèles d'IRMf en utilisant le modèle IC-GAN, pré-entraîné sur ImageNet. Nous avons extrait des caractéristiques d'instance, des vecteurs de bruit et des vecteurs denses à partir d'images d'entraînement et nous avons entraîné des modèles de régression de crête pour prédire ces variables latentes à partir de modèles d'IRMf. À l'aide de ces modèles, nous avons décodé des variables latentes à partir de modèles IRMf de test et reconstruit des images à l'aide du générateur IC-GAN. Notre méthode a démontré une performance de pointe dans la reconstruction des attributs sémantiques de haut niveau des images, à la fois qualitativement et quantitativement, tout en conservant plus de détails de bas niveau par rapport à d'autres modèles sémantiquement orientés comme l'approche BigBiGAN de Mozafari et al. En outre, nous avons utilisé notre modèle pour visualiser des images décodées à partir de modèles IRMf synthétiques conçus pour maximiser les activations dans des régions cérébrales spécifiques, avec des résultats cohérents avec la littérature neuroscientifique existante. Cette méthode ouvre de nouvelles possibilités pour l'exploration du cerveau et les techniques de visualisation.

5.0.3 Chapitre 3 : Diffuseur Cérébral (Brain-diffuser)

Les études précédentes ont réussi à recréer différents aspects des images, tels que les propriétés de bas niveau (forme, texture, disposition) ou les caractéristiques de haut niveau (catégorie d'objets, sémantique descriptive des scènes), mais n'ont généralement pas réussi à reconstruire l'ensemble de ces propriétés pour des im-

ages de scènes complexes. L'IA générative a récemment fait un bond en avant avec des modèles de diffusion latente capables de générer des images très complexes. Dans le troisième chapitre, nous étudions comment tirer parti de cette technologie innovante pour le décodage du cerveau. Nous présentons une méthode de reconstruction de scène en deux étapes appelée "Brain-Diffuser". Dans la première étape, à partir des signaux IRMf, nous reconstruisons des images qui capturent les propriétés de bas niveau et la disposition générale à l'aide d'un modèle VDVAE (Very Deep Variational Autoencoder). Dans un deuxième temps, nous utilisons la méthode image à image d'un modèle de diffusion latent (Versatile Diffusion) conditionné par des caractéristiques multimodales prédites (textuelles et visuelles), afin de générer des images reconstruites finales. Sur l'ensemble de données de scènes naturelles (NSD) accessibles au public, notre méthode est plus performante que les modèles précédents, tant sur le plan qualitatif que quantitatif. Lorsqu'il est appliqué à des modèles synthétiques d'IRMf générés à partir de masques ROI (région d'intérêt) individuels, notre modèle entraîné crée des scènes "ROI-optimales" convaincantes, conformes aux connaissances neuroscientifiques. Ainsi, la méthodologie proposée peut avoir un impact sur les neurosciences appliquées (par exemple, l'interface cerveau-ordinateur) et fondamentales.

Nous avons utilisé le Natural Scenes Dataset (NSD), un ensemble de données IRMf 7T à grande échelle[2], impliquant 8 sujets regardant des images COCO[138]. Chaque image a été visionnée pendant 3 secondes et les sujets ont effectué une tâche de reconnaissance. Notre étude s'est concentrée sur les 4 sujets qui ont effectué tous les essais, ce qui a donné 8859 images d'entraînement et 24980 essais

IRMf, et 982 images de test et 2770 essais IRMf. Les signaux IRMf prétraités ont été masqués à l'aide du masque ROI NSDGeneral[2].

Pour la reconstruction des images, nous avons utilisé une méthode comprenant deux étapes. Au cours de la première étape, nous avons utilisé un auto-codeur variationnel (VDVAE)[33] pour extraire des variables latentes de bas niveau et avons entraîné un modèle de régression ridge pour prédire ces variables à partir des signaux d'IRMf. Cette reconstruction initiale a fourni une disposition des images mais manquait de caractéristiques de haut niveau.

Dans un deuxième temps, nous avons utilisé le modèle Versatile Diffusion[255], un modèle de diffusion latent conditionné par les caractéristiques du texte et de l'image. Nous avons formé deux modèles de régression pour prédire les caractéristiques CLIP-Vision et CLIP-Texte à partir des modèles IRMf. En utilisant le pipeline image-image, nous avons encodé la reconstruction initiale de la VAE, ajouté du bruit, et l'avons débruitée en utilisant les caractéristiques CLIP prédites[185]. Ce processus a produit des images finales naturelles à haute résolution, combinant efficacement les caractéristiques de bas niveau et de haut niveau pour une reconstruction précise de l'image à partir des données IRMf.

Les reconstructions de notre modèle capturent la plupart des attributs sémantiques et de disposition des images originales, bien que les détails au niveau des pixels varient. Par exemple, l'avion reconstruit conserve la pose et l'arrière-plan corrects, tandis que les personnes et les objets sont reconnus de manière cohérente, malgré quelques différences dans les détails et les textures. Ces résultats indiquent la capacité de notre modèle à générer des rendus naturels ressemblant aux images

de référence.

Nous mettons également en évidence certains échecs de reconstruction où des stimuli complexes, des occlusions ou des confusions d'objets ont conduit à des imprécisions. Notre modèle est comparé à trois autres modèles (Lin et al.[137], Takagi et al.[234], et Gu et al.[80]) en utilisant l'ensemble de données NSD. Bien que tous les modèles capturent des informations de haut niveau dans une certaine mesure, notre modèle préserve mieux les détails de bas niveau et la sémantique de haut niveau. Par exemple, nos reconstructions de plans et de visages sont plus réalistes et plus détaillées que celles des autres modèles.

D'un point de vue quantitatif, notre modèle est plus performant que d'autres sur les métriques de bas niveau (PixCorr, SSIM) et de haut niveau (Inception, CLIP). Une étude d'ablation révèle que la combinaison de VDVAE (étape 1) avec le modèle Brain-Diffuser complet (étape 2) optimise l'équilibre entre les détails de bas niveau et les caractéristiques sémantiques de haut niveau. La suppression de composants tels que CLIP-Text ou CLIP-Vision réduit les performances, ce qui confirme leur importance dans le processus de reconstruction. Dans l'ensemble, notre modèle Brain-Diffuser offre un meilleur compromis pour une reconstruction d'image détaillée et sémantiquement précise.

Pour comprendre la relation entre les régions cérébrales et les composantes de notre modèle (VDVAE, CLIP-Vision, CLIP-Text), nous avons effectué une analyse des régions d'intérêt (ROI) en utilisant 4 ROI visuelles provenant d'expériences sur les champs récepteurs de la population (pRF) et 4 ROI provenant d'expériences sur la localisation fonctionnelle (fLoc). Nous avons calculé la force des poids

de régression pour les caractéristiques CLIP et VDVAE pour chaque voxel dans ces régions, montrant que les régions visuelles précoces (V1-V4) sont plus informatives sur les caractéristiques VDVAE, tandis que les régions cérébrales supérieures (Mots, Visages, Corps, Lieux) portent plus d'informations sur les caractéristiques CLIP.

Nous avons également utilisé notre modèle pour visualiser les stimuli "optimaux" pour des régions cérébrales spécifiques en générant des schémas IRMf synthétiques avec des zones d'intérêt activées. Cela nous a permis de créer des images qui activent au maximum certaines régions du cerveau. Pour les premières régions visuelles, nous avons observé des scènes très contrastées avec des textures détaillées, tandis que les régions de niveau supérieur ont généré des images spécifiques à une catégorie, telles que des visages pour la ROI Visage et des scènes d'intérieur pour la ROI Lieu. Ces résultats s'alignent sur la littérature neuroscientifique connue, confirmant que notre méthode peut visualiser de manière convaincante les propriétés fonctionnelles des régions cérébrales.

En outre, nous avons exploré l'organisation rétinotopique en analysant les régions visuelles basées sur l'excentricité. Nos résultats montrent que Brain-Diffuser peut transmettre la localisation spatiale des objets dans les images et a appris l'organisation rétinotopique du cortex basée sur l'excentricité, les régions de la vision centrale générant des objets détaillés au centre et les régions de la vision périphérique générant des objets vers les bords. Cela démontre la précision et la robustesse des représentations latentes de notre modèle.

Dans le troisième chapitre, nous avons développé une méthode à deux étapes

(Brain-Diffuser) pour reconstruire des images à partir de modèles IRMf en utilisant des modèles génératifs basés sur la diffusion latente. Dans la première étape, nous avons utilisé le modèle VDVAE pour la reconstruction initiale des détails de bas niveau. Dans un deuxième temps, nous avons utilisé le modèle Versatile Diffusion, en utilisant les caractéristiques CLIP-Vision et CLIP-Text prédites pour affiner les reconstructions initiales. Cette approche a nécessité l'apprentissage de modèles de régression ridge pour cartographier l'activité cérébrale dans les espaces latents pertinents.

Nos résultats, analysés à la fois qualitativement et quantitativement, montrent que Brain-Diffuser préserve la plupart des informations sémantiques et de mise en page des images originales, générant des images plus naturalistes par rapport aux études précédentes. D'un point de vue quantitatif, Brain-Diffuser surpasse les modèles antérieurs dans les métriques de bas niveau et de haut niveau. Notre méthode, qui utilise des modèles de diffusion latente, établit un nouvel état de l'art dans la reconstruction d'images basée sur l'IRMf, offrant des améliorations par rapport à d'autres approches récentes telles que celles de Chen et al. et Takagi et al.

5.0.4 Chapitre 4 Discussion

Les travaux futurs de cette thèse pourraient consister à tester notre méthode sur d'autres ensembles de données d'image-IRM et à explorer de nouveaux modèles génératifs profonds. Au fur et à mesure que les modèles génératifs s'améliorent, la sélection des modèles les plus appropriés pour le décodage du cerveau sera cruciale.

En outre, l'inclusion de régions cérébrales plus vastes ou l'analyse du cerveau entier pourraient améliorer les reconstructions, mais pourraient augmenter les coûts de calcul. Les études futures pourraient également se concentrer sur la reconstruction de films à l'aide de modèles temporels[249,131], ainsi que sur la conception de nouvelles expériences pour sonder des régions cérébrales moins connues ou pour créer des stimuli optimaux pour des zones cérébrales spécifiques. Cette approche pourrait faire progresser notre compréhension de l'organisation sensorielle et sémantique du cerveau, en fournissant de nouvelles informations sur le traitement et la représentation neuronaux.

Les recherches futures pourraient se concentrer sur la reconstruction des pensées visuelles à l'aide des données d'imagerie. L'imagerie mentale, la représentation d'informations sensorielles sans stimuli externes, peut être observée par neuro-imagerie[172]. L'application de modèles de reconstruction visuelle aux données d'imagerie pourrait révéler des différences entre l'imagerie et la perception visuelle, explorer la diversité de la vivacité entre les sujets et examiner le rôle des ROI dans l'imagerie mentale. Bien que le jeu de données NSD contienne des données d'imagerie, elles ne sont pas encore accessibles au public.

Les progrès récents en matière de décodage neuronal ont rendu possibles les applications réelles des interfaces cerveau-ordinateur (ICO)[72]. Ces interfaces permettent une communication directe entre le cerveau et des dispositifs externes en utilisant des techniques de neuroimagerie invasives (électrodes intracorticales, ECoG) ou non invasives (EEG, fNIRS, MEG, IRMf), choisies en fonction des exigences de l'application. Les BCI basés sur l'EEG sont populaires pour leur mo-

bilité et leur caractère non invasif. Ils sont utilisés dans la rééducation motrice et la réadaptation après un accident vasculaire cérébral, dans le contrôle des robots[208], dans les jeux vidéo et dans le pilotage de quadricoptères. Les méthodes invasives comme l'ECoG sont utilisées pour l'acquisition de signaux robustes dans des applications cliniques, telles que les prothèses vocales neurales de haute qualité pour les patients atteints de SLA[58]. L'IRMf, malgré ses limites, est utilisée dans des cas cliniques comme l'évaluation des réponses chez les patients souffrant de troubles de la conscience. Ces techniques permettent d'étudier l'organisation fonctionnelle et le codage neuronal des régions du cerveau, ce qui améliore les applications des ICO.

Toutefois, la capacité de décoder des informations très fines soulève des questions éthiques sur la protection de la vie privée. Un décodage neuronal précis pourrait conduire à des abus, comme le montrent des scénarios dystopiques tels que "Nineteen Eighty-Four" d'Orwell, "Minority Report" et "Black Mirror"[206]. Les implications juridiques dans le domaine du droit neurologique comprennent des questions telles que la détection du mensonge, la prédiction du comportement criminel et la réadaptation des criminels[25]. À mesure que l'IA et les neurosciences progressent, il est essentiel de répondre aux préoccupations éthiques concernant la fiabilité du décodage neuronal pour les applications critiques, et de garantir la justice, la sécurité et la vie privée, tout en évitant l'alarmisme et les théories du complot[146].

Cette thèse présente deux méthodes pour la reconstruction d'images visuelles à partir de l'activité IRMf en utilisant des modèles génératifs profonds. Les méth-

odes se concentrent sur différents aspects de la reconstruction visuelle par rapport à leurs prédécesseurs, tels que la cohérence sémantique et le réalisme. Ils présentent des résultats supérieurs à ceux d'autres méthodes, tant sur le plan qualitatif que quantitatif. En plus de leur capacité à reconstruire les stimuli, ces méthodes ont démontré leur utilité pour l'exploration neuroscientifique par le biais d'analyses de régions d'intérêt. Les contributions de cette thèse au domaine du décodage neuronal et de la reconstruction visuelle peuvent guider les chercheurs travaillant dans le domaine des neurosciences cognitives et des interfaces cerveau-ordinateur.

References

[1] Adelson, E. H. & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Josa a*, 2(2), 284–299.

[2] Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., Nau, M., Caron, B., Pestilli, F., Charest, I., et al. (2022). A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1), 116–126.

[3] Anderson, S. J., Holliday, I. E., Singh, K. D., & Harding, G. F. (1996). Localization and functional analysis of human cortical area v5 using magneto-encephalography. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 263(1369), 423–431.

[4] Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. In *International conference on machine learning* (pp. 214–223).: PMLR.

[5] Armstrong, D. M. (2002). *A materialist theory of the mind*. Routledge.

[6] at Berkeley, M. L. (2022). Cs 198-126: Lecture 12 - diffusion models.

[7] Baars, B. J. (1993). *A cognitive theory of consciousness*. Cambridge University Press.

[8] Barlow, H. B. (1953). Summation and inhibition in the frog's retina. *The Journal of physiology*, 119(1), 69.

[9] Bartels, A. & Zeki, S. (2000). The architecture of the colour centre in the human visual brain: new results and a review. *European Journal of Neuroscience*, 12(1), 172–193.

[10] Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*, 364(6439), eaav9436.

[11] Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6541–6549).

[12] Bau, D., Zhu, J.-Y., Strobelt, H., Zhou, B., Tenenbaum, J. B., Freeman, W. T., & Torralba, A. (2018). Gan dissection: Visualizing and understanding generative adversarial networks. In *International Conference on Learning Representations*.

[13] Bear, M., Connors, B., & Paradiso, M. A. (2020). *Neuroscience: exploring the brain, enhanced edition: exploring the brain*. Jones & Bartlett Learning.

[14] Beliy, R., Gaziv, G., Hoogi, A., Strappini, F., Golan, T., & Irani, M. (2019). From voxels to pixels and back: Self-supervision in natural-image reconstruction from fmri. *Advances in Neural Information Processing Systems*, 32.

[15] Benchetrit, Y., Banville, H., & King, J.-R. (2023). Brain decoding: toward real-time reconstruction of visual perception. *arXiv preprint arXiv:2310.19812*.

[16] Blonder, L. X., Smith, C. D., Davis, C. E., Kesler, M. L., Garrity, T. F., Avison, M. J., Andersen, A. H., et al. (2004). Regional brain response to faces of humans and dogs. *Cognitive Brain Research*, 20(3), 384–394.

[17] Born, R. T. & Bradley, D. C. (2005). Structure and function of visual area mt. *Annu. Rev. Neurosci.*, 28, 157–189.

[18] Brock, A., Donahue, J., & Simonyan, K. (2019). Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*.

[19] Bunge, S. & Kahn, I. (2009). Cognition: An overview of neuroimaging techniques. In L. R. Squire (Ed.), *Encyclopedia of Neuroscience* (pp. 1063–1067). Oxford: Academic Press.

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

[21] Canziani, A. (2020a). Week 8 – practicum: Variational autoencoders.

[22] Canziani, A. (2020b). Week 9 – practicum: (energy-based) generative adversarial networks.

[23] Carlson, T. A., Schrater, P., & He, S. (2003). Patterns of activity in the categorical representations of objects. *Journal of cognitive neuroscience*, 15(5), 704–717.

[24] Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33, 9912–9924.

[25] Caruso, G. D. (2024). Neurolaw. *Elements in Philosophy of Mind*.

[26] Casanova, A., Careil, M., Verbeek, J., Drozdzal, M., & Romero, A. (2021). Instance-conditioned GAN. In A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in Neural Information Processing Systems*.

[27] Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of consciousness studies*, 2(3), 200–219.

[28] Chalmers, D. J. (1997). *The conscious mind: In search of a fundamental theory*. Oxford Paperbacks.

[29] Chen, J., Qi, Y., & Pan, G. (2023a). Rethinking visual reconstruction: Experience-based content completion guided by visual cues. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research* (pp. 4856–4866).: PMLR.

[30] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597–1607).: PMLR.

[31] Chen, Z., Qing, J., Xiang, T., Yue, W. L., & Zhou, J. H. (2023b). Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 22710–22720).

[32] Child, R. (2020). Very deep vaes generalize autoregressive models and can outperform them on images. *arXiv preprint arXiv:2011.10650*.

[33] Child, R. (2021). Very deep {vae}s generalize autoregressive models and can outperform them on images. In *International Conference on Learning Representations*.

[34] Churchland, P. M. (2013). *Matter and consciousness*. MIT press.

[35] Churchland, P. S. (1989). *Neurophilosophy: Toward a unified science of the mind-brain*. MIT press.

[36] Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, 6(1), 27755.

[37] Clarke, D. (2003). *Descartes's Theory of Mind*. Oxford University Press.

[38] Conference, N. (2022). Yuki kamitani : Externalizing and sharing the neuroverse...

[39] Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20, 273–297.

[40] Cox, D. D. & Savoy, R. L. (2003). Functional magnetic resonance imaging (fmri)"brain reading": detecting and classifying distributed patterns of fmri activity in human visual cortex. *Neuroimage*, 19(2), 261–270.

[41] Crick, F. & Koch, C. (2003). A framework for consciousness. *Nature neuroscience*, 6(2), 119–126.

[42] Dado, T., Güçlütürk, Y., Ambrogioni, L., Ras, G., Bosch, S., van Gerven, M., & Güçlü, U. (2022). Hyperrealistic neural decoding for reconstructing faces from fmri activations via the gan latent space. *Scientific reports*, 12(1), 141.

[43] Daneshfard, B., Dalfardi, B., & Mahmoudinezhad, G. (2014). Ibn al-haytham (965–1039 ad), the original portrayal of the modern theory of vision. *Journal of medical biography*.

[44] de Beeck, H. O. & Nakatani, C. (2019). *Introduction to human neuroimaging*. Cambridge University Press.

[45] Decramer, T., Premereur, E., Uytterhoeven, M., Van Paesschen, W., van Loon, J., Janssen, P., & Theys, T. (2019). Single-cell selectivity and functional architecture of human lateral occipital complex. *PLoS biology*, 17(9), e3000280.

[46] Dehaene, S. (2014). *Consciousness and the Brain*. Penguin.

[47] Dehaene, S., Kerszberg, M., & Changeux, J.-P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the national Academy of Sciences*, 95(24), 14529–14534.

[48] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255).: Ieee.

[49] Dennett, D. C. (1993). *Consciousness explained*. Penguin uk.

[50] Descartes, R. (1972). *Treatise of Man: French Text with Translation and Commentary, Trans. Thomas Steele Hall*. Cambridge, Mass.: Newcomb Livraria Press.

[51] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[52] Dijkstra, N. (2024). Uncovering the role of the early visual cortex in visual mental imagery. *Preprints*.

[53] Dijkstra, N., Bosch, S. E., & van Gerven, M. A. (2019). Shared neural mechanisms of visual perception and imagery. *Trends in cognitive sciences*, 23(5), 423–434.

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

[55] Dowling, J. E. (2018). *Understanding the brain: from cells to behavior to cognition*. WW Norton & Company.

[56] Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science*, 293(5539), 2470–2473.

[57] Du, C., Du, C., & He, H. (2017). Sharing deep generative representation for perceived image reconstruction from human brain activity. In *2017 International Joint Conference on Neural Networks (IJCNN)* (pp. 1049–1056).: IEEE.

[58] Duraivel, S., Rahimpour, S., Chiang, C.-H., Trumpis, M., Wang, C., Barth, K., Harward, S. C., Lad, S. P., Friedman, A. H., Southwell, D. G., et al. (2023). High-resolution neural recordings improve the accuracy of speech decoding. *Nature communications*, 14(1), 6938.

[59] Dwivedi, K., Bonner, M. F., Cichy, R. M., & Roig, G. (2021). Unveiling functions of the visual cortex using task-specific deep neural networks. *PLoS computational biology*, 17(8), e1009267.

[60] Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017). Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152, 184–194.

[61] Epstein, R., DeYoe, E. A., Press, D. Z., Rosen, A. C., & Kanwisher, N. (2001). Neuropsychological evidence for a topographical learning mechanism in parahippocampal cortex. *Cognitive neuropsychology*, 18(6), 481–508.

[62] Epstein, R. & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, 392(6676), 598–601.

[63] Esser, P., Rombach, R., & Ommer, B. (2020). Taming transformers for high-resolution image synthesis. 2021 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 12868–12878).

[64] Fang, T., Qi, Y., & Pan, G. (2020). Reconstructing perceptive images from brain activity by shape-semantic gan. *Advances in Neural Information Processing Systems*, 33, 13038–13048.

[65] Ferrante, M., Boccato, T., & Toschi, N. (2022). Semantic brain decoding: from fmri to conceptually similar image reconstruction of visual stimuli. *arXiv preprint arXiv:2212.06726*.

[66] Feser, E. (2006). *Philosophy of mind: A beginner's guide*. Simon and Schuster.

[67] Finzi, D., Margalit, E., Kay, K., Yamins, D. L., & Grill-Spector, K. (2023). A single computational objective drives specialization of streams in visual cortex. *bioRxiv*, (pp. 2023–12).

[68] Foster, D. (2022). *Generative deep learning*. " O'Reilly Media, Inc.".

[69] Freud, S., Strachey, J., Freud, A., Rothgeb, C. L., Richards, A., Strachey, A., & Corporation, S. L. (1953). *The standard edition of the complete psychological works of Sigmund Freud*. London: Hogarth Press London.

[70] Fulford, J., Milton, F., Salas, D., Smith, A., Simler, A., Winlove, C., & Zeman, A. (2018). The neural correlates of visual imagery vividness – an fmri study and literature review. *Cortex*, 105, 26–40. The Eye's Mind - visual imagination, neuroscience and the humanities.

[71] Gallant, J. L., Braun, J., & Van Essen, D. C. (1993). Selectivity for polar, hyperbolic, and cartesian gratings in macaque visual cortex. *Science*, 259(5091), 100–103.

[72] Gao, X., Wang, Y., Chen, X., & Gao, S. (2021). Interface, interaction, and intelligence in generalized brain–computer interfaces. *Trends in cognitive sciences*, 25(8), 671–684.

[73] Gaziv, G., Beliy, R., Granot, N., Hoogi, A., Strappini, F., Golan, T., & Irani, M. (2022). Self-supervised natural image reconstruction and large-scale semantic classification from brain activity. *NeuroImage*, 254, 119121.

[74] Gazzaniga, M., Ivry, R. B., & Mangun, G. R. (2018). *Cognitive neuroscience: fifth international student edition*. WW Norton & Company.

[75] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

[76] Greene, M. R. & Hansen, B. C. (2018). Shared spatiotemporal category representations in biological and artificial deep neural networks. *PLoS computational biology*, 14(7), e1006327.

[77] Grill-Spector, K., Knouf, N., & Kanwisher, N. (2004). The fusiform face area subserves face perception, not generic within-category identification. *Nature neuroscience*, 7(5), 555–562.

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

[79] Gross, C. G., Rocha-Miranda, C. d., & Bender, D. (1972). Visual properties of neurons in inferotemporal cortex of the macaque. *Journal of neurophysiology*, 35(1), 96–111.

[80] Gu, Z., Jamison, K., Kuceyeski, A., & Sabuncu, M. R. (2023). Decoding natural image stimuli from fMRI data with a surface-based convolutional network. In *Medical Imaging with Deep Learning*.

[81] Gu, Z., Jamison, K. W., Khosla, M., Allen, E. J., Wu, Y., St-Yves, G., Naselaris, T., Kay, K., Sabuncu, M. R., & Kuceyeski, A. (2022). Neurogen: activation optimized image synthesis for discovery neuroscience. *NeuroImage*, 247, 118812.

[82] 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.

[83] Güçlütürk, Y., Güçlü, U., Seeliger, K., Bosch, S., van Lier, R., & van Gerven, M. A. (2017). Reconstructing perceived faces from brain activations with deep adversarial neural decoding. *Advances in neural information processing systems*, 30.

[84] Hannula, D. E., Simons, D. J., & Cohen, N. J. (2005). Imaging implicit perception: promise and pitfalls. *Nature Reviews Neuroscience*, 6(3), 247–255.

[85] Hansen, N. & Ostermeier, A. (2001). Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2), 159–195.

[86] Hanson, S. J., Matsuka, T., & Haxby, J. V. (2004). Combinatorial codes in ventral temporal lobe for object recognition: Haxby (2001) revisited: is there a "face" area? *Neuroimage*, 23(1), 156–166.

[87] Hartline, H. K. (1938). The response of single optic nerve fibers of the vertebrate eye to illumination of the retina. *American Journal of Physiology-Legacy Content*, 121(2), 400–415.

[88] Hawking, S. (2010). *The grand design*. Random House Digital, Inc.

[89] Haxby, J. V. (2012). Multivariate pattern analysis of fmri: the early beginnings. *Neuroimage*, 62(2), 852–855.

[90] 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.

[91] Haynes, J.-D. & Rees, G. (2005). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature neuroscience*, 8(5), 686–691.

[92] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

[93] Hebb, D. O. (2005). *The organization of behavior: A neuropsychological theory*. Psychology press.

[94] Henderson, J. M., Zhu, D. C., & Larson, C. L. (2011). Functions of parahippocampal place area and retrosplenial cortex in real-world scene analysis: an fmri study. *Visual Cognition*, 19(7), 910–927.

[95] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., & Lerchner, A. (2016). beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*.

[96] Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, 18(7), 1527–1554.

[97] Hinton, G. E., Sejnowski, T. J., et al. (1986). Learning and relearning in boltzmann machines. *Parallel distributed processing: Explorations in the microstructure of cognition*, 1(282-317), 2.

[98] Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735–1780.

[99] Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8), 2554–2558.

[100] Horikawa, T. & Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. *Nature communications*, 8(1), 1–15.

[101] Horikawa, T., Tamaki, M., Miyawaki, Y., & Kamitani, Y. (2013). Neural decoding of visual imagery during sleep. *Science*, 340(6132), 639–642.

[102] Howard, I. P. (1996). Alhazen's neglected discoveries of visual phenomena. *Perception*, 25(10), 1203–1217.

[103] Hubel, D. H. & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1), 106.

[104] Hubel, D. H. & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1), 215–243.

[105] Husain, M. (2008). Chapter 18 hemispatial neglect. In *Neuropsychology and Behavioral Neurology*, volume 88 of *Handbook of Clinical Neurology* (pp. 359–372). Elsevier.

[106] Ibn al Hayṯam, a.-H. b. a.-H., Witelo, Risnerus, F., (Basel), E., & Nikolaus (II, erven, B. (1572). *Opticae thesavrvs : Alhazeni Arabis libri septem, nunc primùm editi : Eivsdem liber De Crepvscvlis & nubium ascensionibus : Item Vitellonis Thvringopoloni libri X.* Basiliae,: per Episcopios Basiliae,.

[107] Ioffe, S. & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning* (pp. 448–456).: pmlr.

[108] Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1125–1134).

[109] Jain, N., Wang, A., Henderson, M. M., Lin, R., Prince, J. S., Tarr, M. J., & Wehbe, L. (2023). Selectivity for food in human ventral visual cortex. *Communications Biology*, 6(1), 175.

[110] Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D., & Makedon, F. (2021). A survey on contrastive self-supervised learning. *Technologies*, 9(1), 2.

[111] Jolicoeur-Martineau, A. (2018). The relativistic discriminator: a key element missing from standard gan. *arXiv preprint arXiv:1807.00734*.

[112] Jordan, M. I. (2019). Artificial Intelligence—The Revolution Hasn't Happened Yet. *Harvard Data Science Review*, 1(1). https://hdsr.mitpress.mit.edu/pub/wot7mkc1.

[113] Kamitani, Y. & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature neuroscience*, 8(5), 679–685.

[114] Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S., Hudspeth, A. J., Mack, S., et al. (2000). *Principles of neural science*, volume 4. McGraw-hill New York.

[115] Kant, I. (1908). Critique of pure reason. 1781. *Modern Classical Philosophers, Cambridge, MA: Houghton Mifflin*, (pp. 370–456).

[116] Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience*, 17(11), 4302–4311.

[117] Karpathy, A. (2016). Cs231n winter 2016: Lecture 7: Convolutional neural networks.

[118] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8110–8119).

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

[120] Khaligh-Razavi, S.-M. & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11), e1003915.

[121] Khosla, M., Murty, N. A. R., & Kanwisher, N. (2022). A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition. *Current Biology*, 32(19), 4159–4171.

[122] Kingma, D. P. & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

[123] Kneeland, R., Ojeda, J., St-Yves, G., & Naselaris, T. (2023). Brain-optimized inference improves reconstructions of fmri brain activity. *arXiv preprint arXiv:2312.07705*.

[124] Koide-Majima, N., Nishimoto, S., & Majima, K. (2024). Mental image reconstruction from human brain activity: Neural decoding of mental imagery via deep neural network-based bayesian estimation. *Neural Networks*, 170, 349–363.

[125] Konkle, T. & Alvarez, G. A. (2022). A self-supervised domain-general learning framework for human ventral stream representation. *Nature Communications*, 13(1), 1–12.

[126] Konkle, T., Conwell, C., Prince, J. S., & Alvarez, G. A. (2022). What can 5.17 billion regression fits tell us about the representational format of the high-level human visual system? *Journal of Vision*, 22(14), 4422–4422.

[127] Kosslyn, S. M., Behrmann, M., & Jeannerod, M. (1995). The cognitive neuroscience of mental imagery. *Neuropsychologia*.

[128] 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.

[129] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

[130] Kuffler, S. W. (1953). Discharge patterns and functional organization of mammalian retina. *Journal of neurophysiology*, 16(1), 37–68.

[131] Kupershmidt, G., Beliy, R., Gaziv, G., & Irani, M. (2022). A penny for your (visual) thoughts: Self-supervised reconstruction of natural movies from brain activity. *arXiv preprint arXiv:2206.03544*.

[132] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324.

[133] Leibniz, G. W. (1714). *Philosophical Texts*. Oxford University Press. translated by Richard Francs and R.S Woolhouse published in 1998.

[134] Lenartowicz, A. & Poldrack, R. (2010). Brain imaging. In G. F. Koob, M. L. Moal, & R. F. Thompson (Eds.), *Encyclopedia of Behavioral Neuroscience* (pp. 187–193). Oxford: Academic Press.

[135] Levy, I., Hasson, U., Avidan, G., Hendler, T., & Malach, R. (2001). Center–periphery organization of human object areas. *Nature neuroscience*, 4(5), 533–539.

[136] Li, J., Zhou, P., Xiong, C., & Hoi, S. (2020). Prototypical contrastive learning of unsupervised representations. In *International Conference on Learning Representations*.

[137] Lin, S., Sprague, T. C., & Singh, A. (2022). Mind reader: Reconstructing complex images from brain activities. In A. H. Oh, A. Agarwal, D. Belgrave, & K. Cho (Eds.), *Advances in Neural Information Processing Systems*.

[138] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13* (pp. 740–755).: Springer.

[139] Lindberg, D. C. (1976). *Theories of Vision from al-Kindi to Kepler*. University of Chicago Press.

[140] Liu, Y., Ma, Y., Zhou, W., Zhu, G., & Zheng, N. (2023). Brainclip: Bridging brain and visual-linguistic representation via clip for generic natural visual stimulus decoding from fmri. *arXiv preprint arXiv:2302.12971*.

[141] Lorenz, H. (2019). 506507Plato on the Soul. In *The Oxford Handbook of Plato*. Oxford University Press.

[142] Lueck, C., Zeki, S., Friston, K., Deiber, M.-P., Cope, P., Cunningham, V. J., Lammertsma, A., Kennard, C., & Frackowiak, R. (1989). The colour centre in the cerebral cortex of man. *Nature*, 340(6232), 386–389.

[143] Luo, A., Henderson, M., Wehbe, L., & Tarr, M. (2024). Brain diffusion for visual exploration: Cortical discovery using large scale generative models. *Advances in Neural Information Processing Systems*, 36.

[144] Luo, A., Henderson, M. M., Tarr, M. J., & Wehbe, L. (2023). Brainscuba: Fine-grained natural language captions of visual cortex selectivity. In *The Twelfth International Conference on Learning Representations*.

[145] Marcus, G. (2018). Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*.

[146] May, J. (2023). *Neuroethics: Agency in the Age of Brain Science*. Oxford University Press.

[147] McCandliss, B. D., Cohen, L., & Dehaene, S. (2003). The visual word form area: expertise for reading in the fusiform gyrus. *Trends in cognitive sciences*, 7(7), 293–299.

[148] McCulloch, W. S. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5, 115–133.

[149] Meng, L. & Yang, C. (2023). Dual-guided brain diffusion model: Natural image reconstruction from human visual stimulus fmri. *Bioengineering*, 10(10), 1117.

[150] Mirza, M. & Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

[151] MITCBMM (2018a). 2.10 - magnetoencephalography (meg).

[152] MITCBMM (2018b). 2.9 - event-related potentials (erps).

[153] MITCBMM (2018c). 3.2 - the retina.

[154] Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M.-a., Morito, Y., Tanabe, H. C., Sadato, N., & Kamitani, Y. (2008). Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, 60(5), 915–929.

[155] Mozafari, M., Reddy, L., & VanRullen, R. (2020). Reconstructing natural scenes from fmri patterns using bigbigan. In *2020 International joint conference on neural networks (IJCNN)* (pp. 1–8).: IEEE.

[156] Mueller, K. N., Carter, M. C., Kansupada, J. A., & Ponce, C. R. (2023). Macaques recognize features in synthetic images derived from ventral stream neurons. *Proceedings of the National Academy of Sciences*, 120(10), e2213034120.

[157] Nanay, B. (2023). *Mental Imagery: Philosophy, Psychology, Neuroscience*. Oxford University Press.

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

[159] Nichol, A. Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., & Chen, M. (2022). Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning* (pp. 16784–16804).: PMLR.

[160] Nielsen, M. A. (2015). *Neural networks and deep learning*, volume 25. Determination press San Francisco, CA, USA.

[161] Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current biology*, 21(19), 1641–1646.

[162] Noh, H., Hong, S., & Han, B. (2015). Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision* (pp. 1520–1528).

[163] Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fmri data. *Trends in cognitive sciences*, 10(9), 424–430.

[164] Ogawa, S., Lee, T.-M., Kay, A. R., & Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *proceedings of the National Academy of Sciences*, 87(24), 9868–9872.

[165] Online, M. (2020). Lecture 19: Generative models i.

[166] OpenCourseWare, M. (2021). 7. category selectivity, controversies, and mvpa.

[167] Orban, G. A., Van Essen, D., & Vanduffel, W. (2004). Comparative mapping of higher visual areas in monkeys and humans. *Trends in cognitive sciences*, 8(7), 315–324.

[168] O'toole, A. J., Jiang, F., Abdi, H., & Haxby, J. V. (2005). Partially distributed representations of objects and faces in ventral temporal cortex. *Journal of cognitive neuroscience*, 17(4), 580–590.

[169] Owen, A. M. & Coleman, M. R. (2008). Functional neuroimaging of the vegetative state. *Nature Reviews Neuroscience*, 9(3), 235–243.

[170] Ozcelik, F., Choksi, B., Mozafari, M., Reddy, L., & VanRullen, R. (2022). Reconstruction of perceived images from fmri patterns and semantic brain exploration using instance-conditioned gans. In *2022 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8).: IEEE.

[171] Ozcelik, F. & VanRullen, R. (2023). Natural scene reconstruction from fmri signals using generative latent diffusion. *Scientific Reports*, 13(1), 15666.

[172] Pearson, J., Naselaris, T., Holmes, E. A., & Kosslyn, S. M. (2015). Mental imagery: functional mechanisms and clinical applications. *Trends in cognitive sciences*, 19(10), 590–602.

[173] Peelen, M. V. & Downing, P. E. (2017). Category selectivity in human visual cortex: Beyond visual object recognition. *Neuropsychologia*, 105, 177–183.

[174] Perrett, D., Rolls, E., & Caan, W. (1982). Visual neurones responsive to faces in the monkey temporal cortex. *Experimental brain research*, 47, 329–342.

[175] Pinti, P., Tachtsidis, I., Hamilton, A., Hirsch, J., Aichelburg, C., Gilbert, S., & Burgess, P. W. (2020). The present and future use of functional near-infrared spectroscopy (fnirs) for cognitive neuroscience. *Annals of the New York Academy of Sciences*, 1464(1), 5–29.

[176] Pividori, M., Grinblat, G. L., & Uzal, L. C. (2019). Exploiting gan internal capacity for high-quality reconstruction of natural images. *arXiv preprint arXiv:1911.05630*.

[177] Plebe, A. & de la Cruz, V. (2016). Neurosemantics.

[178] POLDRACK, R. A. (2018). *The New Mind Readers: What Neuroimaging Can and Cannot Reveal about Our Thoughts*. Princeton University Press.

[179] Polimeni, J. R., Fischl, B., Greve, D. N., & Wald, L. L. (2010). Laminar analysis of 7 t bold using an imposed spatial activation pattern in human v1. *Neuroimage*, 52(4), 1334–1346.

[180] Priest, S. (1991). *Theories of the Mind*. New York, N.Y., USA: Penguin Books.

[181] Prince, J. S., Alvarez, G. A., & Konkle, T. (2023). Lesioning category-selective units in silico yields functionally specialized deficits. *Journal of Vision*, 23(9), 5657–5657.

[182] Putnam, H. (1980). The nature of mental states. In *The Language and Thought Series* (pp. 223–231). Harvard University Press.

[183] Qiao, K., Chen, J., Wang, L., Zhang, C., Tong, L., & Yan, B. (2020). Biggan-based bayesian reconstruction of natural images from human brain activity. *Neuroscience*, 444, 92–105.

[184] Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045), 1102–1107.

[185] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).: PMLR.

[186] Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

[187] Radford, A. & Narasimhan, K. (2018). Improving language understanding by generative pre-training. In *OpenAI*.

[188] Rakhimberdina, Z., Jodelet, Q., Liu, X., & Murata, T. (2021). Natural image reconstruction from fmri using deep learning: A survey. *Frontiers in neuroscience*, 15, 795488.

[189] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

[190] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-shot text-to-image generation. In *International conference on machine learning* (pp. 8821–8831).: Pmlr.

[191] Ramsey, W. (2022). Eliminative Materialism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2022 edition.

[192] Ratan Murty, N. A., Bashivan, P., Abate, A., DiCarlo, J. J., & Kanwisher, N. (2021). Computational models of category-selective brain regions enable high-throughput tests of selectivity. *Nature communications*, 12(1), 5540.

[193] Razavi, A., Van den Oord, A., & Vinyals, O. (2019). Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32.

[194] Reddy, L. & Kanwisher, N. (2007). Category selectivity in the ventral visual pathway confers robustness to clutter and diverted attention. *Current Biology*, 17(23), 2067–2072.

[195] Reddy, L., Tsuchiya, N., & Serre, T. (2010). Reading the mind's eye: decoding category information during mental imagery. *Neuroimage*, 50(2), 818–825.

[196] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779–788).

[197] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

[198] Ren, Z., Li, J., Xue, X., Li, X., Yang, F., Jiao, Z., & Gao, X. (2021). Reconstructing seen image from brain activity by visually-guided cognitive representation and adversarial learning. *NeuroImage*, 228, 117602.

[199] Revonsuo, A. (2009). *Consciousness: The science of subjectivity*. Psychology Press.

[200] Robinson, H. (2023). Dualism. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2023 edition.

[201] Roelfsema, P. R., Denys, D., & Klink, P. C. (2018). Mind reading and writing: The future of neurotechnology. *Trends in cognitive sciences*, 22(7), 598–610.

[202] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10684–10695).

[203] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18* (pp. 234–241).: Springer.

[204] Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6), 386.

[205] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088), 533–536.

[206] Ryberg, J. (2017). Neuroscience, mind reading and mental privacy. *Res Publica*, 23(2), 197–211.

[207] Sacks, O. (2003). The mind's eye. *New Yorker*, 28, 48–59.

[208] Saha, S., Mamun, K. A., Ahmed, K., Mostafa, R., Naik, G. R., Darvishi, S., Khandoker, A. H., & Baumert, M. (2021). Progress in brain computer interface: Challenges and opportunities. *Frontiers in Systems Neuroscience*, 15, 578875.

[209] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Gontijo-Lopes, R., Ayan, B. K., Salimans, T., Ho, J., Fleet, D. J., & Norouzi, M. (2022). Photorealistic text-to-image diffusion models with deep language understanding. In A. H. Oh, A. Agarwal, D. Belgrave, & K. Cho (Eds.), *Advances in Neural Information Processing Systems*.

[210] Sarch, G. H., Tarr, M. J., Fragkiadaki, K., & Wehbe, L. (2023). Brain dissection: fmri-trained networks reveal spatial selectivity in the processing of natural images. *bioRxiv*, (pp. 2023–05).

[211] Schoenmakers, S., Barth, M., Heskes, T., & Van Gerven, M. (2013). Linear reconstruction of perceived images from human brain activity. *NeuroImage*, 83, 951–961.

[212] Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., & Komatsuzaki, A. (2021). Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

[213] Scotti, P., Banerjee, A., Goode, J., Shabalin, S., Nguyen, A., Dempster, A., Verlinde, N., Yundler, E., Weisberg, D., Norman, K., et al. (2024). Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. *Advances in Neural Information Processing Systems*, 36.

[214] Seeliger, K., Güçlü, U., Ambrogioni, L., Güçlütürk, Y., & van Gerven, M. (2018). Generative adversarial networks for reconstructing natural images from brain activity. *NeuroImage*, 181, 775–785.

[215] Seth, A. (2021). *Being you: A new science of consciousness*. Penguin.

[216] Shen, G., Dwivedi, K., Majima, K., Horikawa, T., & Kamitani, Y. (2019a). End-to-end deep image reconstruction from human brain activity. *Frontiers in computational neuroscience*, 13, 21.

[217] Shen, G., Horikawa, T., Majima, K., & Kamitani, Y. (2019b). Deep image reconstruction from human brain activity. *PLoS computational biology*, 15(1), e1006633.

[218] Shirakawa, K. (2023). Critical assessment of generative ai methods and natural image datasets for visual image reconstruction from brain activity.

[219] Silverberg, A. (1992). Putnam on functionalism. *Philosophical Studies*, 67, 111–131.

[220] Simonyan, K. & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[221] Skinner, B. F. (1965). *Science and human behavior*. Number 92904. Simon and Schuster.

[222] Smart, J. J. C. (2024). Materialism | definition, theories, history,& facts.

[223] Smith, A. M. (2019). *From sight to light: The passage from ancient to modern optics*. University of Chicago Press.

[224] Smith, K. (2013). Reading minds. *Nature*, 502(7472), 428.

[225] Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2020). Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

[226] Spiridon, M. & Kanwisher, N. (2002). How distributed is visual category information in human occipito-temporal cortex? an fmri study. *Neuron*, 35(6), 1157–1165.

[227] Srinivasan, R. & Nunez, P. (2012). Electroencephalography. In V. Ramachandran (Ed.), *Encyclopedia of Human Behavior (Second Edition)* (pp. 15–23). San Diego: Academic Press, second edition edition.

[228] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1), 1929–1958.

[229] St-Yves, G. & Naselaris, T. (2018). Generative adversarial networks conditioned on brain activity reconstruct seen images. In *2018 IEEE international conference on systems, man, and cybernetics (SMC)* (pp. 1054–1061).: IEEE.

[230] Stevens, W. D., Kravitz, D. J., Peng, C. S., Tessler, M. H., & Martin, A. (2017). Privileged functional connectivity between the visual word form area and the language system. *Journal of Neuroscience*, 37(21), 5288–5297.

[231] Sun, J., Li, M., Chen, Z., Zhang, Y., Wang, S., & Moens, M.-F. (2024). Contrast, attend and diffuse to decode high-resolution images from brain activities. *Advances in Neural Information Processing Systems*, 36.

[232] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–9).

[233] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818–2826).

[234] Takagi, Y. & Nishimoto, S. (2023). High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14453–14463).

[235] Tan, M. & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105–6114).: PMLR.

[236] Thakor, N. (2012). Building brain machine interfaces—neuroprosthetic control with electrocorticographic signals. *Newsletter. IEEE Life Sciences*.

[237] Thirion, B., Duchesnay, E., Hubbard, E., Dubois, J., Poline, J.-B., Lebihan, D., & Dehaene, S. (2006). Inverse retinotopy: inferring the visual content of images from brain activation patterns. *Neuroimage*, 33(4), 1104–1116.

[238] Tononi, G. (2004). An information integration theory of consciousness. *BMC neuroscience*, 5, 1–22.

[239] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In *International conference on machine learning* (pp. 10347–10357).: PMLR.

[240] University, S. (2017). Neural networks part 1: Setting up the architecture.

[241] Van den Oord, A., Kalchbrenner, N., Espeholt, L., Vinyals, O., Graves, A., et al. (2016). Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29.

[242] Van Den Oord, A., Kalchbrenner, N., & Kavukcuoglu, K. (2016). Pixel recurrent neural networks. In *International conference on machine learning* (pp. 1747–1756).: PMLR.

[243] Van Den Oord, A., Vinyals, O., et al. (2017). Neural discrete representation learning. *Advances in neural information processing systems*, 30.

[244] Van Essen, D. C. & Gallant, J. L. (1994). Neural mechanisms of form and motion processing in the primate visual system. *Neuron*, 13(1), 1–10.

[245] VanRullen, R. & Reddy, L. (2019). Reconstructing faces from fmri patterns using deep generative neural networks. *Communications biology*, 2(1), 1–10.

[246] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

[247] Voynov, A. & Babenko, A. (2020). Unsupervised discovery of interpretable directions in the gan latent space. In *International conference on machine learning* (pp. 9786–9796).: PMLR.

[248] Wang, A. Y., Kay, K., Naselaris, T., Tarr, M. J., & Wehbe, L. (2023). Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nature Machine Intelligence*, 5(12), 1415–1426.

[249] Wang, C., Yan, H., Huang, W., Li, J., Wang, Y., Fan, Y.-S., Sheng, W., Liu, T., Li, R., & Chen, H. (2022). Reconstructing rapid natural vision with fmri-conditional video generative adversarial network. *Cerebral Cortex*, 32(20), 4502–4511.

[250] Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4), 600–612.

[251] Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological review*, 20(2), 158.

[252] Wen, H., Shi, J., Zhang, Y., Lu, K.-H., Cao, J., & Liu, Z. (2018). Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral cortex*, 28(12), 4136–4160.

[253] Werbos, P. (1974). Beyond regression: New tools for prediction and analysis in the behavioral sciences. *PhD thesis, Committee on Applied Mathematics, Harvard University, Cambridge, MA*.

[254] Xia, W., de Charette, R., Oztireli, C., & Xue, J.-H. (2024). Dream: Visual decoding from reversing human visual system. In *Proceedings of the*

IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 8226–8235).

[255] Xu, X., Wang, Z., Zhang, E., Wang, K., & Shi, H. (2022). Versatile diffusion: Text, images and variations all in one diffusion model. *arXiv preprint arXiv:2211.08332*.

[256] Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23), 8619–8624.

[257] Zbontar, J., Jing, L., Misra, I., LeCun, Y., & Deny, S. (2021). Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning* (pp. 12310–12320).: PMLR.

[258] Zeki, S. (1983). Colour coding in the cerebral cortex: the reaction of cells in monkey visual cortex to wavelengths and colours. *Neuroscience*, 9(4), 741–765.

[259] Zeki, S. M. (1974). Functional organization of a visual area in the posterior bank of the superior temporal sulcus of the rhesus monkey. *The Journal of physiology*, 236(3), 549–573.

[260] Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2023). *Dive into deep learning*. Cambridge University Press.

[261] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 586–595).

[262] Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2223–2232).

[263] Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. L. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3).

[264] Zimmermann, M., Mars, R. B., De Lange, F. P., Toni, I., & Verhagen, L. (2018). Is the extrastriate body area part of the dorsal visuomotor stream? *Brain Structure and Function*, 223(1), 31–46.