Hierarchical Visual Representation Shared Across Individuals

Ho, Jun Kai

Supervisor: Prof. Yukiyasu Kamitani

Department of Intelligence Science and Technology
Graduate School of Informatics
Kyoto University

This thesis is submitted for the degree of
Doctor of Philosophy

August 2023

Declaration

I, HO Jun Kai, declare that this thesis, entitled “Hierarchical visual representation shared across individuals” is original and my own work.

I confirm that:

- This work was done solely while a candidate for the research degree at the Graduate School of Informatics, Kyoto University.
- No part of this work has previously been submitted for a degree at this or any other university.
- References to the work of others have been clearly attributed. Quotations from the work of others have been clearly indicated, and attributed to them.
- In cases where others have contributed to part of this work, such contribution has been clearly acknowledged and distinguished from my own work.

Ho, Jun Kai
August 2023

Acknowledgements

At the outset, I express my heartfelt gratitude to Prof. Yukiyasu Kamitani for welcoming me into his extraordinary laboratory, and for his unwavering support throughout this enlightening journey that not only enhanced my research skills but also nurtured my personal growth. I consider it a privilege to have been a part of his research team.

I extend my sincere appreciation to Dr. Tomoyasu Horikawa and Dr. Kei Majima, whose invaluable guidance has profoundly shaped my research career.

I am grateful to all my colleagues in the lab, whose companionship made this journey enriching and enjoyable. My warm thanks to each member of the Kyoto University lab: Ken Shirakawa, Fan Cheng, Jong-Yun Park, Haibao Wang, Shangfeng Jin, Shuntaro Aoki, and Misato Tanaka, as well as those at ATR, particularly Mitsuaki Tsukamoto.

Further, I wish to acknowledge the research administrators, Yukari Kado and Yukiko Masuda, who have facilitated my journey by managing the many intricacies of research, and graciously accommodated my initial struggle with Japanese language.

My heartfelt thanks also go to the members of the Kokoro Research Center at Kyoto University, for their indispensable assistance during the experimental processes.

Lastly, I wish to convey my deepest appreciation to my family. Their unwavering belief in my abilities and consistent support played an instrumental role in my academic journey, leading me to this proud moment of graduating from Kyoto University.

Abstract

The human brain exhibits both shared anatomical characteristics and individual variations. While global structures such as the central sulcus are present in all human brains, local variations in sulci and gyri exist. Despite anatomical alignment, differences in neural representation and brain activity patterns persist due to diverse developmental environments and experiences. However, it is plausible to assume that when presented with a particular stimulus, brain activity patterns across different individuals encode relatively similar information. Therefore, in theory, the brain responses of one individual could potentially predict the responses of another individual to the same stimulus. Functional alignment, a technique that aligns brain activity patterns without considering anatomical structure, has emerged as a means to investigate the existence of neural representations shared across individuals.

The processing of visual information follows a hierarchical pathway, wherein early stages detect simple local features, while later stages encode complex global features, ultimately enabling holistic perception. However, it remains unclear whether this hierarchical and fine-grained visual representation can be effectively converted across individuals while preserving the encoded perceptual content. To address this issue, this thesis employs functional magnetic resonance imaging (fMRI), functional alignment methods, and deep neural network models. The study utilizes a functional alignment technique called the neural code converter, which predicts a target subject's brain activity pattern based on the response of a source subject to the same stimulus. The converted patterns are then analyzed through the decoding of hierarchical visual features and the reconstruction of perceived images.

Chapter 1 provides an introduction to brain differences among individuals, functional alignment, and visual features shared across individuals, along with a review of the current state of the field. Chapter 2 describes the human brain activity data and experimental design in the study. Chapter 3 demonstrates that human brain activity patterns and visual hierarchy can be converted across individuals with moderate accuracy. Chapter 4 analyzes the converted brain activity patterns through decoding of deep neural network features and visual image reconstruction. Chapter 5 explicitly compares neural code conversions trained without imposing

visual hierarchy to those that respect the visual hierarchy, highlighting the effectiveness of data-driven approaches in detecting cortical hierarchy. Chapter 6 explores the pooling of data from multiple subjects into a target subject's space, resulting in slightly improved decoders as assessed through visual image reconstruction evaluation. Finally, Chapter 7 discusses the implication of neural code converters, future directions in inter-individual visual image reconstructions and explores other visual features possibly shared across individuals.

Table of contents

List of figures xiii

1 Introduction 1

1.1 Commonalities and differences of human brains 1

1.1.1 Brain structures 2

1.1.2 Anatomical alignment 4

1.1.3 Functional topography 5

1.1.4 Functional brain activity pattern 6

1.2 Functional alignment 8

1.2.1 Hyperalignment 9

1.2.2 Neural code converter 11

1.2.3 Others 13

1.3 Previous studies on visual features shared across individuals 13

1.3.1 Image contrast 14

1.3.2 Object identity 14

1.3.3 Retinotopic organization 17

1.3.4 Semantic contents 18

1.4 Hierarchical visual processing in the ventral pathway 19

1.4.1 Contrast 21

1.4.2 Orientation 22

1.4.3 Depth 23

1.4.4 Color 24

1.4.5 Contour integration 25

1.4.6 Texture perception 26

1.4.7 Object recognition 27

1.5 Artificial deep neural network (DNN) 28

1.5.1 Basics of DNN 28

1.5.2 Brain and DNN 30
1.5.3 Why are DNNs good for modeling the visual system? 31
1.6 Thesis organization 33

2 Human brain activity data 35
2.1 Introduction 35
2.2 Basics of fMRI 36
2.3 Review of related fMRI studies on visual decoding 37
2.4 Experiments 39
2.4.1 Subjects 40
2.4.2 Visual stimuli 40
2.4.3 Experimental design 40
2.4.4 fMRI acquisition 41
2.4.5 fMRI data preprocessing 41
2.4.6 Region of interest (ROI) 42
2.4.7 Data availability 42
2.5 Data for analysis 43

3 Hierarchical neural code conversion 45
3.1 Introduction 45
3.2 Methods 47
3.2.1 Anatomical alignment 47
3.2.2 Neural code converter 48
3.2.3 Conversion accuracy 49
3.2.4 Noise ceiling estimation 50
3.2.5 Ablation analysis on neural code converters 51
3.2.6 Statistics 51
3.3 Results 52
3.3.1 Evaluation of neural code converters 52
3.3.2 Detection of the hierarchical correspondence 55
3.4 Discussion 55

4 Inter-individual DNN feature decoding and deep image reconstruction 59
4.1 Introduction 59
4.2 Methods 62
4.2.1 Deep neural networks (DNN) 62
4.2.2 DNN feature decoding analysis 63

4.2.3 Brain hierarchy (BH) score 63
4.2.4 Deep image reconstruction 64
4.2.5 Identification analysis 65
4.2.6 Other methods of functional alignment 65
4.2.7 Statistics 67
4.3 Results 68
4.3.1 DNN feature decoding 68
4.3.2 Visual image reconstruction 72
4.3.3 Robustness across functional alignment methods 76
4.4 Discussion 79

5 Effectiveness of data-driven hierarchical neural code converter - A comparison 83
5.1 Introduction 83
5.2 Methods 84
5.2.1 Visual subarea-wise conversion 84
5.2.2 Statistics 85
5.3 Results 86
5.3.1 Subarea-wise conversion accuracy 86
5.3.2 DNN feature decoding via subarea-wise conversion 86
5.3.3 Visual image reconstruction via subarea-wise conversion 86
5.3.4 Varying the number of training data 89
5.4 Discussions 96

6 Application of neural code converters to pooling data analysis 97
6.1 Introduction 97
6.2 Methods 98
6.2.1 Pooling data into a target subject 98
6.2.2 Statistics 99
6.3 Results 100
6.3.1 Pooling data into a target subject 100
6.3.2 Pooling data analysis with limited data availability 102
6.4 Discussions 105

7 General discussion 111
7.1 Summary of findings and contributions 111
7.2 Fine-scaled voxel mapping 112
7.3 The Black-Box of DNN 114

7.4 Potential of neural code converters in visual restoration 115
7.5 Saving data collection time 116
7.6 Inter-individual neural code conversion without paired stimuli 116
7.7 Other future directions 118
7.7.1 Other visual features 118
7.7.2 Decoding task optimized neural code converters 119
7.7.3 Pooling open data from other laboratories 121
7.7.4 Geometric-based alignment 121
7.7.5 Domain adaptation 122
7.8 Conclusions 123

References **125**

Appendix A Publications **137**
A.1 Manuscript 137
A.2 Presentation 137

Appendix B Code availability **139**

List of figures

1.1 Four lobes of the human cerebrum 3

1.2 T1-weighted (T1w) MRI image of the human brain 4

1.3 Illustration of retinotopic mapping using rotating wedges and expanding rings 7

1.4 Comparison of fMRI activity patterns in two subjects viewing the same image 8

1.5 Idea of functional alignment 10

1.6 Schematic of searchlight hyperalignment 11

1.7 Illustration of neural code converter 12

1.8 Visual image reconstruction by multiscale local image decoders 15

1.9 Reconstructed artificial images from the converted brain activity patterns of V1 area 16

1.10 Retinotopic mapping from the visual field to the visual cortex 17

1.11 llustration for building between-subject semantics encoding models with hyperalignment 20

1.12 Schematic of the center-surround receptive field 21

1.13 Schematic of the formation of the receptive field of a simple cell 23

1.14 Schematic of deep neural network 28

1.15 Examples of images preferred four by randomly selected units in each DNN layer 29

1.16 Deep image reconstruction 31

1.17 Illustration of preservation of the hierarchical and fine-grained visual features 32

2.1 Schematic of image presentation experiment 38

2.2 Visual images reconstructed from the fMRI brain activities 38

2.3 Reconstruction with participant's attention directed towards one of two superimposed images 39

3.1 Visual cortical hierarchy along the ventral pathway 46

3.2 Training of a neural code converter 48

3.3 Evaluations of neural code converters 50
3.4 Ablation analysis on neural code converters 51
3.5 Conversion accuracies of individual pairs 53
3.6 Conversion accuracies of 20 individual pairs 53
3.7 Single trial brain activity patterns responding to two test natural images . . 54
3.8 Cortical map of the effects of source area exclusion 56
3.9 Performance drop caused by source area exclusion for four representative individual pairs 57
3.10 Mean performance drop caused by source area exclusion across 20 individual pairs 57

4.1 Training of the DNN feature decoders 60
4.2 Inter-individual deep image reconstruction 62
4.3 Illustration of pairwise alignment and template-based pairwise alignment . 67
4.4 DNN feature decoding accuracy from the whole visual cortex (VC) for four representative individual pairs 70
4.5 Mean DNN feature decoding accuracy from the whole visual cortex (VC) across 20 individual pairs 70
4.6 Proportion of the “top visual area” across DNN units in each layer for four representative individual pairs 71
4.7 Mean proportion of the “top visual area” across DNN units in each layer over 20 individual pairs 71
4.8 Brain hierarchy (BH) score 72
4.9 Within and across-individual reconstructions from the whole visual cortex (VC) 73
4.10 Reconstructed images across individual pairs 74
4.11 Reconstructed images using varying repetitions of samples 75
4.12 Identification accuracies of representative individual pairs 77
4.13 Mean identification accuracy across all individual pairs or all subjects . . . 78
4.14 Evaluation of different methods of functional alignment on visual areas . . 78
4.15 Inter-individual DNN feature decoding accuracies via different methods of functional alignment 79
4.16 Reconstructed images obtained via different methods of functional alignment 80

5.1 Schematics of the subarea-wise conversion 85
5.2 Subarea-wise conversion accuracies for individual pairs 87

5.3 Mean and distribution of subarea-wise conversion accuracies across individual pairs 87
5.4 DNN feature decoding accuracy of natural images and artificial images for four representative individual pairs 88
5.5 Mean and distribution of subarea-wise conversion accuracies across individual pairs 89
5.6 Reconstructed natural and artificial images via whole VC conversions and subarea-wise conversions 90
5.7 Identification accuracies with the reconstructed natural and artificial images for four representative individual pairs 91
5.8 Mean identification accuracies across all individual pairs 92
5.9 Reconstructed images via neural converters trained with a varying amount of data 93
5.10 Identification accuracies of representative individual pairs with different numbers of training samples for converters 94
5.11 Mean identification accuracies across all individual pairs with varying numbers of training data for the whole VC and subarea-wise conversions 95

6.1 Basic idea of pooling data through neural code conversion 98
6.2 Illustration of the pooling procedure 99
6.3 Inter-individual DNN Feature decoding accuracies of representative individual pairs with multiple- and single-subject feature decoders 101
6.4 Mean DNN feature decoding accuracy obtained via multiple- and single-subject feature decoders across all individual pairs 102
6.5 Reconstructed natural images for the multiple- and single-subject conditions 103
6.6 Identification analyses of representative individual pairs with multiple- and single-subject feature decoders 104
6.7 Mean identification accuracies obtained via multiple- and single-subject feature decoders across all individual pairs 105
6.8 Inter-individual DNN Feature decoding accuracies of representative individual pairs with multiple- and single-subject feature decoders under limited data availability 106
6.9 Mean DNN feature decoding accuracy obtained via multiple- and single-subject feature decoders across all individual pairs under limited data availability 107
6.10 Reconstructed natural images for the multiple- and single-subject conditions under limited data availability 108

6.11 Identification analyses of representative individual pairs with multiple- and single-subject feature decoders under limited data availability 109
6.12 Mean identification accuracies obtained via multiple- and single-subject feature decoders across all individual pairs under limited data availability . 110

7.1 Schematic of the training process for the neural code converter in the DNN feature space 117
7.2 Reconstructions of illusory images from brain activity patterns 120

Chapter 1

Introduction

Human brains exhibit both individual uniqueness and shared characteristics in anatomical and functional domains. Anatomical variations often become evident in local brain structures such as gyrus and sulcus, while functional differences manifest in functional topography and distinct responses to stimuli. While some common anatomical features are readily apparent, exploring complex aspects such as structural connectivity and functional attributes necessitates the use of brain imaging techniques, with this thesis primarily focusing on functional magnetic resonance imaging (fMRI).

To uncover shared features, an alignment process is essential. Traditional fMRI studies have employed anatomical alignment as a standard preprocessing step. However, the more recent advent of functional alignment has unveiled a range of shared features across individuals. This thesis will delve into these shared features, emphasizing the role of functional alignment in their identification.

1.1 Commonalities and differences of human brains

While human brains share global structures such as hemispheres and four lobes - frontal, temporal, parietal, and occipital - they exhibit variations in size, shape, and local structures like gyri and sulci. These local differences persist even after accounting for variations in brain size and shape. To align these global and local structures, several anatomical alignment techniques have been proposed. However, these techniques have limitations as functional topographies do not always align across individuals, posing challenges for group-level studies where fMRI data are normalized to a common template.

Furthermore, brains may respond differently to the same stimulus due to varying developmental environments and experiences. However, it has been observed that brain responses synchronized across individuals when viewing a movie (Hasson et al., 2004). This suggests that the brain activity pattern in response to a stimulus in one individual could predict the brain activity pattern in another.

1.1.1 Brain structures

The human brain is an intricate network of neural connections that define who we are as individuals. Among its constituents, the cerebrum is the largest and arguably the most complex part, responsible for various sophisticated functions such as cognition, language, memory, and sensory processing. Understanding the structure of the cerebrum – including its hemispheres, lobes, gyri, and sulci – and the functionalities associated with these structures, is key to understand how the brain shapes our perceptions and behaviors.

The cerebrum is essentially split into two halves, referred to as the left and right hemispheres, joined by a robust bundle of nerve fibers called the corpus callosum. This integration allows for communication and coordination between the two hemispheres, enabling a comprehensive perception of our world by integrating information from both sides of the body. Interestingly, while the basic organization of these hemispheres is consistent across individuals, some unique functionalities are often lateralized. For instance, language processing is typically localized to the left hemisphere, while the right hemisphere often handles spatial abilities and face recognition. These lateralizations, however, are not absolute and can show variation among individuals, influenced by factors such as handedness and cultural upbringing.

Further subdividing each hemisphere, we find four primary lobes: the frontal, parietal, temporal, and occipital lobes (Figure 1.1). Each lobe hosts specific functionalities and their respective structure is generally conserved across individuals, although the size and specific organization may differ slightly. The frontal lobe, located at the front of the brain, is essential for higher cognitive functions such as decision-making, problem-solving, and planning. It also controls voluntary motor activity. Adjacent to the frontal lobe, the parietal lobe specializes in processing sensory information and maintaining spatial awareness, enabling us to navigate and interact with our surroundings effectively. The temporal lobe, situated at the sides of the brain near the ears, is primarily responsible for auditory processing and memory. Finally, the occipital lobe, located at the back of the brain, is dedicated to visual processing, translating the light that enters our eyes into images that we can comprehend.

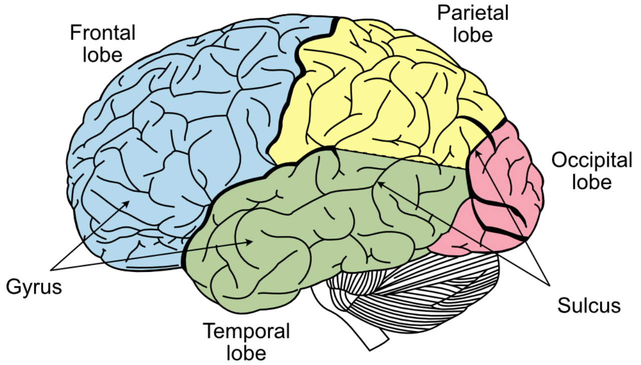

Figure 1.1: Four lobes of the human cerebrum. The frontal lobe (depicted in blue) is involved in higher cognitive functions such as decision-making, planning, and voluntary motor activity. The parietal lobe (depicted in yellow) processes sensory information and is integral to spatial awareness. The temporal lobe (depicted in green) plays a crucial role in auditory processing and memory. Lastly, the occipital lobe (depicted in red) is dedicated to visual processing, transforming the light entering our eyes into comprehensible images.

An additional defining feature of the cerebrum is its convoluted surface, marked by folds (gyri) and grooves (sulci). This folding pattern increases the surface area of the brain, facilitating a higher density of neurons and therefore increasing cognitive capacity. While the specific pattern of gyri and sulci can exhibit individual variation, the overall folding pattern, as well as major landmarks like the central sulcus separating the frontal and parietal lobes, are remarkably conserved across individuals.

Magnetic Resonance Imaging (MRI) is a non-invasive imaging technique that provides detailed, high-resolution images of the internal structures of the body, including the brain. This technology is capable of generating high-resolution images, thereby providing an intricate view of the internal structure of the brain (Figure 1.2), which is otherwise unobservable through mere external examination. This imaging technique does not merely show the gross anatomical structure, but reveals minute details such as the gray and white matter, subcortical structures, and complex neural pathways. These high-resolution images underscore the existence of individual variations, particularly in the local structures of the brain, such as differences in cortical fold size and shape, cortical thickness, or the relative dimensions of distinct brain structures.

In summary, while the cerebrum's anatomy maintains a degree of uniformity among individuals, variations in size, shape, and local structures do exist. This level of consistency allows for generalized understandings of brain functionality, correlating certain functions with specific regions within the cerebrum. Nonetheless, these individual differences present challenges when exploring brain functions beyond global structures.

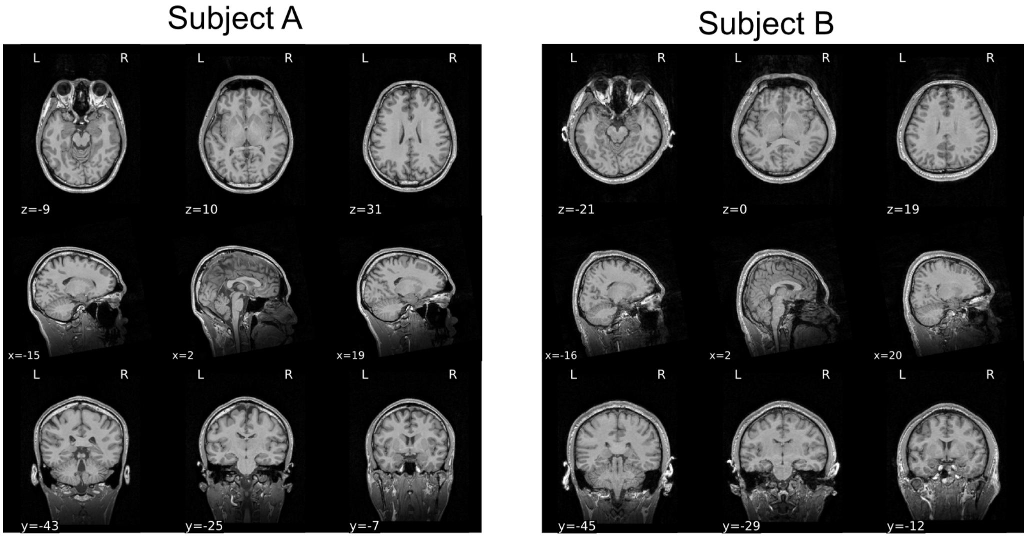

Figure 1.2: T1-weighted (T1w) MRI image of the human brain. The high contrast between gray and white matter demonstrates the intricate detail of internal brain structures, showcasing the unique capabilities of MRI technology. Variation in brain structure can be observed among the two subjects' brain. This image is provided by Kamitani laboratory, Kyoto University, with permission from the subjects.

1.1.2 Anatomical alignment

Over the years, numerous techniques, collectively termed as "anatomical alignment" or "registration", have been developed to precisely align anatomical features. These methods comprise landmark-based, volume-based, and surface-based techniques.

The landmark-based registration technique was pioneered by Tailairach as a method to standardize the alignment of brains across different individuals (Talairach & Tournoux, 1988). This method performs piecewise affine transformation to register a brain to an atlas using anatomical landmarks, including anterior and posterior commissure, midline sagittal plane, and the exterior boundary of the brain. One significant drawback of this method lies in the absence of an MRI image template for the atlas, and hence the registration is solely dependent on the discernment of anatomical landmarks. Consequently, the preference has shifted towards volume-based and surface-based registration techniques.

The volume-based registration registers 3D volumetric images to a brain template. The registration process typically commences with an initial linear and rigid registration, followed by a nonlinear registration of an image to a template. A myriad of nonlinear registration algorithms have been developed, each varying based on the deformation model, similarity metric, and regularization utilized. For instance, SyN incorporates a bi-directional diffeomorphism approach (Avants et al., 2008); SPM5 employs discrete cosine transforms; FNIRT utilizes cubic B-splines. Some of the most frequently used similarity metrics include least squares, normalized correlation, correlation ratio, mutual information, and cross-correlation between an image pair (refer to Poldrack et al., 2011 for further details).

The surface-based registration performs alignment based on surface features, such as gyri and sulci. This method mandates the reconstruction of the cerebral cortex into a cortical surface derived from the anatomical image, a process largely automated in the Freesurfer software package (Dale et al., 1999). Initially, the cortical surface is first transformed into a spherical representation, and the cortical folding pattern is represented as the average convexity on a unit sphere. The alignment is then achieved by minimizing the mean square error between the convexity of an individual and that of a template in a multi-scale manner in which a Gaussian kernel with a decreasing standard deviation is applied (Fischl et al., 1999). The *fsaverage*, a synthesis of 40 MRI scans of brains, is the most commonly employed template.

While anatomical alignment provides a valuable means of reconciling individual variances in the brain structure, this approach alone is insufficient to mitigate the disparities observed in the functional aspects of brain activity.

1.1.3 Functional topography

Functional brain topography is the specialized blueprint of our brain wherein each specific region is assigned dedicated tasks. These functions range from language comprehension, motor actions control, and visual perception, to name a few. While individual brain structures may vary significantly in size and configuration, research has consistently pointed towards an overarching pattern of functional organization that is shared across the majority of humans. This shared functional blueprint is integral in allowing neuroscientists to extrapolate insights about the neural mechanics underpinning cognition and behavior from a universal perspective, rather than limiting it to individual-specific studies.

One well-established example of shared functional topography is the organization of the motor cortex. Situated in the precentral gyrus of the frontal lobe, the primary motor cortex plays a pivotal role in coordinating our body's movements. Within this section of the brain, different body parts find specific representation, adhering to an organized "somatotopic" layout (Penfield and Boldrey, 1937; Grodd et al., 2001; Roux et al., 2020; Gordon et al., 2023). Areas controlling the face and hands are notably more extensive, a pattern observed across individuals. This pattern proves invaluable in medical and technological fields, enabling neurosurgeons to target accurately during brain surgery and aiding in the development of brain-computer interfaces for individuals with motor impairments.

In the visual cortex, orientation columns, composed of an assembly of neurons, are selectively responsive to the orientation of edges within the visual field, with different columns sensitive to different edge orientations. The pattern of these orientation columns is often referred to as a "pinwheel" arrangement, wherein neurons tuned to all possible orientations are represented within a small area of the cortex (Blasdel & Salama, 1986). This intricate design allows for a wide spectrum of orientations to be encoded, enabling the brain to process the complexity and richness of the visual environment. Each pinwheel's center corresponds to a singular point in the visual field, and moving radially outward from this center results in a systematic change in the preferred orientation, yielding a map of orientation preference across the cortical surface.

Another example of functional topography in the visual cortex is the retinotopy, which is the mapping of visual input from the retina to neurons in the visual cortex, maintaining a spatially organized representation of the visual field (Figure 1.3). This mapping begins with the photoreceptors in the retina and continues through the optic nerve to the primary visual cortex located in the occipital lobe of the brain, where it maintains a topographical representation of the visual field. However, this map is not an exact replica of the visual field; rather, there's a phenomenon called cortical magnification, where the central area of the visual field, or the fovea, is overrepresented due to its higher density of photoreceptors.

While the existence of the functional topographies common among human brains underscores similarities in cognitive and behavioral processing across individuals, it's important to acknowledge that there are still variations on an individual level. For instance, even with anatomical alignment taken into account, the positioning of the visual motion area (known as area V5 or MT) can differ as much as approximately 20 mm across different individuals (Watson et al., 1993). This variability in individual functional representations poses substantial challenges, particularly when attempting to standardize fMRI data to a common template for large-scale group studies. Consequently, while the shared functional topography of the human brain offers insights into general patterns, a thorough understanding of human cognition must also factor in individual variations.

1.1.4 Functional brain activity pattern

Functional brain activity pattern refers to the activation of regions of the brain in response to stimuli or during cognitive processes. Neuroimaging techniques such as functional magnetic resonance imaging (fMRI) and electroencephalography (EEG) are commonly used to study the functional brain activity patterns in humans. In fMRI study, a brain activity pattern refers to the pattern of voxel responses to a stimulus.

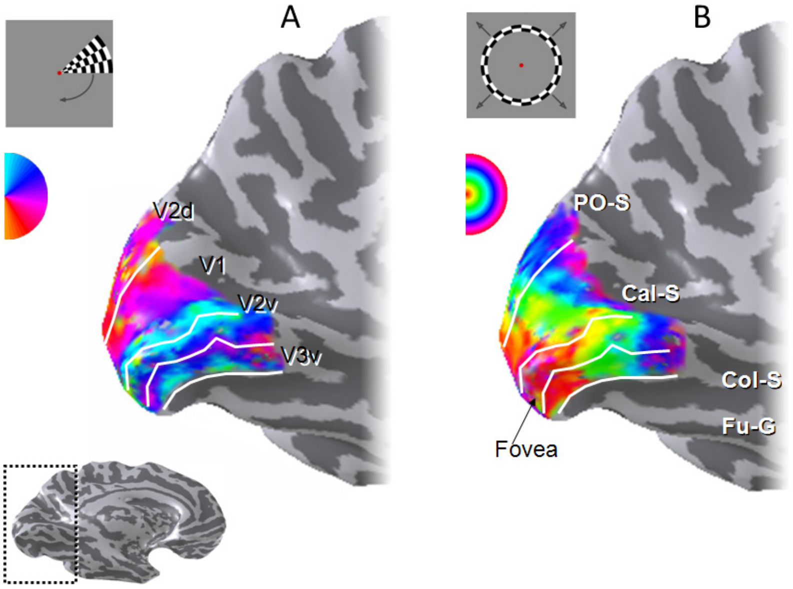

Figure 1.3: Illustration of retinotopic mapping using rotating wedges and expanding rings. This figure illustrates the principle of retinotopy. In panel A, a rotating wedge stimulus moves in a clockwise direction around a central fixation point, mapping polar angle representation in the visual cortex. In panel B, an expanding ring stimulus moves outward from the center, mapping eccentricity representation in the visual cortex. Together, these stimuli help to create a comprehensive retinotopic map of the visual field onto the brain's visual cortex. Reprinted from *Visual Cortex - Current Status and Perspectives*, 2012, Chapter 2, Brewer and Barton, Visual Field Map Organization in Human Visual Cortex, licensed under CC BY 3.0.

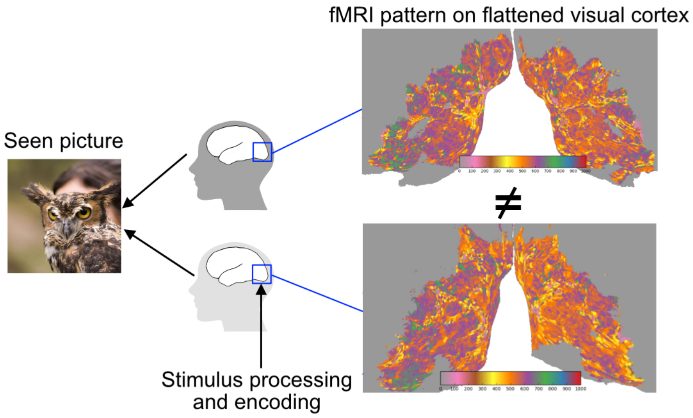
fMRI pattern on flattened visual cortex
Seen picture
≠
Stimulus processing and encoding

Figure 1.4: Comparison of fMRI activity patterns in two subjects viewing the same image. Even when two subjects view the identical picture, their brain processing of the visual stimulus results in different fMRI activity patterns, highlighting the individual variability in brain responses.

Functional brain activity patterns can vary depending on the task or state being studied. For example, when a person is engaged in a visual task, such as watching a movie or looking at a picture, there is increased neural activity in the visual cortex, which is responsible for processing visual information. Similarly, when a person is engaged in a hearing task, such as music listening, there is increased neural activity in the auditory cortex located on the superior temporal gyrus. Functional brain activity patterns encode rich information about the stimulus. With appropriate techniques, the stimulus information can be decoded and even reconstructed, providing insights into the cognitive processing mechanisms.

However, an intriguing feature of these functional brain activity patterns lies in their variability across individuals. While certain functional brain networks show shared characteristics among many individuals, there is significant variability in the brain activity patterns between different people (Figure 1.4). This variability poses a unique challenge to neuroscientists. It complicates the process of drawing generalized conclusions about brain function or developing decoding models that work reliably across a range of individuals. As such, the study of functional brain activity patterns requires careful consideration of individual differences to fully understand the intricate workings of the human brain.

1.2 Functional alignment

In fMRI studies, the brain response of an individual to a given stimulus is characterized by a distinct pattern of fMRI activity that encodes the information pertaining to the stimulus. These fMRI activity patterns for each subject can be represented as high-dimensional vectors, with each voxel serving as a dimension within the vector space. Consequently, each vector in this space represents an fMRI activity pattern that corresponds to a specific stimulus. Stimuli exhibiting common characteristics, such as being quadrupeds and possessing fur, demonstrate brain response vectors that are spatially proximate to each other within the vector space. For instance, the brain response vectors corresponding to a dog and a cat would be closer to one another compared to a response vector associated with a house. In essence, the response vectors encompassing all available stimuli collectively form a response manifold, which is embedded within the vector space. It is important to note that the response manifolds of different subjects generally do not overlap, indicating the presence of variability in responses across individuals.

Functional alignment is based on the underlying assumption that it is possible to achieve alignment of subjects' response vectors for a specific stimulus within a common model space (Figure 1.5), which is also known by various terms in the literature, including common space, common template, shared feature space, or simply shared/common space (Haxby et al., 2011; Chen et al., 2015; Bazeille et al., 2021). This alignment process entails learning the relationships among voxels across subjects using a machine learning model. A prerequisite for this approach is the availability of an fMRI training dataset containing the responses of each subject to a predetermined set of stimuli. Each subject's dataset is represented as a data matrix, where each row corresponds to the fMRI activity pattern evoked by a specific stimulus.

1.2.1 Hyperalignment

Hyperalignment is one of the earliest methods to tackle the problem of functional alignment (Haxby et al., 2011; Haxby et al., 2020 for a recent review). Hyperalignment aligns subjects' brain response vectors corresponding to a stimulus by performing a series of Procrustean transformations (Schönemann, 1966). The algorithm consists of three iterative steps:

1. In the first iteration, the hyperalignment algorithm first selects an initial target subject whose fMRI responses are used as a template, then aligns the second subject's fMRI responses to the template using Procrustes transformation. The template is then updated as the mean of the current template and the newly aligned fMRI responses. The same procedure is repeated for additional subjects.

2. In the second iteration, each subject's original response is aligned to the mean aligned responses of other subjects. The mean aligned response is recalculated and treated as a template.

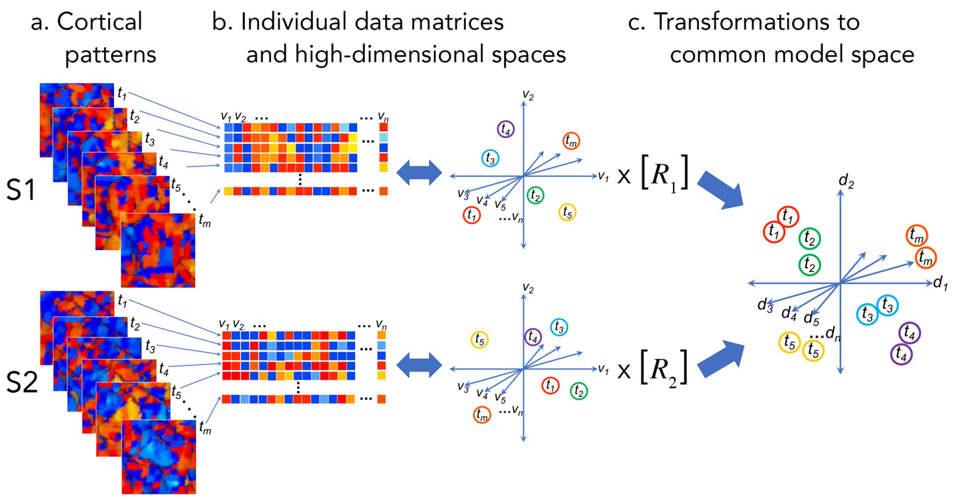
a. Cortical patterns
b. Individual data matrices and high-dimensional spaces
c. Transformations to common model space
S1
S2
t_1 t_2 t_3 t_4 t_5 t_m
v_1 v_2 ... ... v_n
v_1 v_2 v_3 v_4 v_5 ...v_n
x $[R_1]$
x $[R_2]$
d_1 d_2 d_3 d_4 d_5 ...d_n

Figure 1.5: Idea of functional alignment. The brain responses vectors to the same stimulus in the high-dimensional individual spaces can be brought aligned in the common model space. Reprinted from *Elife*, 2020;9: e56601, Haxby et al., Hyperalignment: Modeling shared information encoded in idiosyncratic cortical topographies, Page No. 3, licensed under CC BY 4.0

3. In the last step, each subject's response is aligned to the template, and an orthogonal transformation matrix is obtained for each subject.

Hyperalignment often does not work well in high-dimensional data when dealing with a large region of interest (ROI) with more voxels than training samples. To partially tackle this issue, Chen et al. (2015) proposed joint SVD-hyperalignment that first performs dimensionality reduction by SVD, followed by hyperalignment in the lower-dimensional feature space. However, they are still not optimal for the whole cortex analysis that generally involves hundreds of thousands of voxels. A variant of hyperalignment, which is called searchlight hyperalignment, tackles this issue by using disks of searchlights that cover the whole cortex (Kriegeskorte et al., 2006; Guntupalli et al., 2016). The disk could be a three-dimensional volume or a two-dimensional surface, depending on the type of fMRI data (Oosterhof et al., 2011). Hyperalignment algorithm is then performed in each searchlight disk, and a local transformation matrix is obtained (Figure 1.6). A whole-cortex transformation matrix is obtained by aggregating all local transformation matrices of each searchlight disk. Nevertheless, this method is computationally costly, and the matrix aggregation procedure destroys the local structure imposed within each searchlight disk, thus the aggregated matrix is not orthogonal in general. Despite this disadvantage, searchlight hyperalignment is often the go-to method in the whole-brain functional alignment.

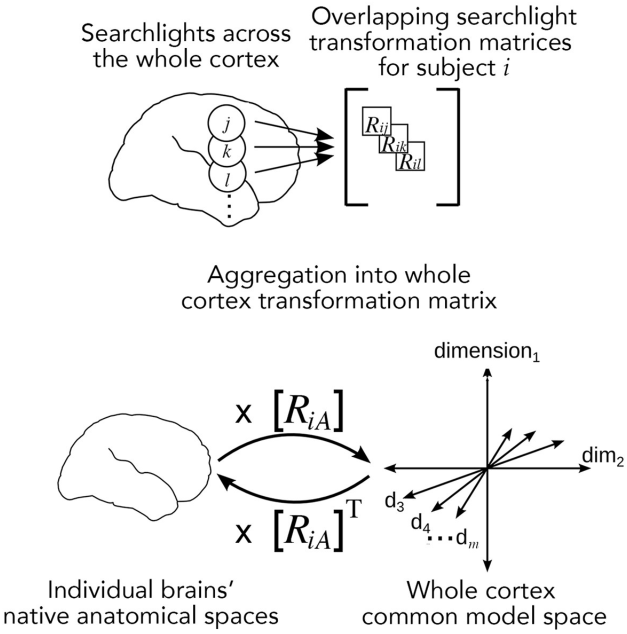
Searchlights across the whole cortex
Overlapping searchlight transformation matrices for subject i
j
k
l
R_{ij}
R_{ik}
R_{il}
Aggregation into whole cortex transformation matrix
dimension$_1$
dim$_2$
x $[R_{iA}]$
x $[R_{iA}]^T$
d$_3$
d$_4$...d$_m$
Individual brains' native anatomical spaces
Whole cortex common model space

Figure 1.6: Schematic of searchlight hyperalignment. Reprinted from *Elife*, 2020;9: e56601, Haxby et al., Hyperalignment: Modeling shared information encoded in idiosyncratic cortical topographies, Page No. 6, licensed under CC BY 4.0

1.2.2 Neural code converter

The concept of neural code converter was introduced around the same time as hyperalignment in 2011, as described by Yamada et al. (2011 & 2015). While both approaches share the principle of functional alignment, they differ in terms of underlying assumptions and model implementations. The neural code converter functions as a pairwise transformation between two subjects, whereas hyperalignment estimates a common template and supports bidirectional transitions between individual space and the common template. The neural code converter is trained to predict fMRI activity patterns of a target subject from measured fMRI activity patterns of a source subject, with both subjects presented to the same sequence of stimuli (Figure 1.7). The converter is then tested with a test stimulus, such as a geometric shape.

Furthermore, the original study primarily focused on the early visual area (V1) and assumed that the activity of a voxel should exhibit similarity to a limited number of voxels in V1 from another subject. To address this assumption, the neural code converter method employs sparse regression, utilizing automatic relevance determination (ARD) prior within Bayesian estimation of the weights. This approach offers greater flexibility compared to the Procrustean transformation employed in hyperalignment, which imposes orthogonality constraint. Despite lacking orthogonality, the neural code converter's flexibility has demonstrated enhanced

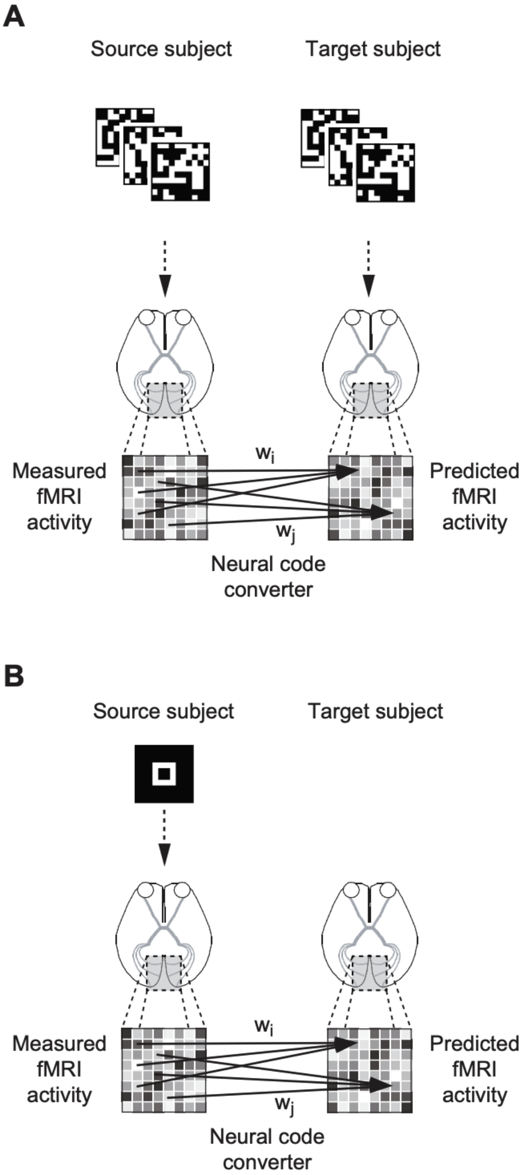

Figure 1.7: Illustration of neural code converter. Panel A indicates the training of converters with a pair of subjects presented to a sequence of stimuli. Panel B shows the conversion of a source subject's fMRI activity pattern to a target subject's brain space. Reprinted from *NeuroImage*, 2015;113, Yamada et al., Inter-subject neural code converter for visual image representation, Page No. 290, Copyright (2015), with permission from Elsevier.

prediction accuracy (Yamada et al., 2011 & 2015). Additionally, the transformation matrix in a neural code converter is generally non-invertible, in contrast to hyperalignment. Generally speaking, a neural code converter need not be a sparse regression. Any regression model that can establish a statistical relationship between subjects' voxels can be termed a neural code converter. In my study, I mainly used the Ridge-based neural code converter.

1.2.3 Others

The field of functional alignment has been developed for a decade, and various methods have been proposed in order to improve the accuracy of alignment. Some methods generalize hyperalignment algorithm, for example, deep hyperalignment (Yousefnezhad & Zhang, 2017) and hybrid hyperalignment (Busch et al., 2021). Other methods, such as optimal transport, canonical correlation analysis (Zhuang et al., 2020 for a review in neuroscience application), and shared response model (Chen et al, 2015), use different approaches to transform or align brain responses. For an evaluation of functional alignments, please refer to Yousefnezhad et al. (2021) and Bazeille et al. (2021).

1.3 Previous studies on visual features shared across individuals

The human visual system operates as a complex mechanism that translates intricate visual features into a holistic perception. These visual features span from basic elements such as edges and image contrast, to more complex attributes like semantics and the identity of objects. In addition to encoding these visual features, the human visual system also arranges them into global organizations. An important instance of such an organization is the retinotopic organization, which maintains a spatial map of visual information throughout the processing pathway in the visual cortex. These features, organized in particular patterns, are processed and interpreted to construct a detailed mental image of our surroundings.

In light of the individual differences outlined previously, the question of whether these visual features are universally shared cannot be resolved without detailed scrutiny. Functional alignment emerges as a potent tool in this regard, enabling the examination of whether a particular visual feature or neural representation is common across individuals. The fundamental concept involves training either an encoding or decoding model on a given subject, and subsequently applying this model to a different subject whose brain activity patterns have been aligned with the former. If the model successfully generalizes to the new subject, it indicates that the visual features learnt by the model are indeed shared among individuals. In the following, I will present several research studies that further elucidate this point.

1.3.1 Image contrast

The investigation of decoding visual perception in the human brain has consistently been an engaging and active area of research (Kay et al., 2008; Miyawaki et al., 2008; Güçlütürk et al., 2017; Shen et al., 2019a & 2019b; Han et al., 2019; Seeliger et al., 2018). One notable early study in this field focused on visual image reconstruction of simple contrast patterns, wherein the multi-scale contrast of an image was decoded using multi-voxel decoders (Miyawaki et al., 2008; Figure 1.8). The stimulus images were 10 × 10 checkerboard patches comprising random images, geometric shapes, and alphabet characters. The researchers employed four scales of local image bases that covered the entire image. At each position within the image, the multi-voxel decoders, trained with hundreds of random images, predicted the contrast at each scale from fMRI signals of the primary visual cortex. By combining the locally predicted contrasts, the visual image of geometric and alphabet shapes could be reconstructed. This finding demonstrated that the primary visual cortex encodes information related to local image contrast.

In further research by Yamada et al. (2011 & 2015), a neural code converter was trained using fMRI responses to random patches. The neural code converter was then employed to convert fMRI responses from a source subject space to a target subject space using structured patches such as squares, square rings, plus signs, crosses, and large square rings (refer to section 1.2.2: Neural code converter). The converted fMRI responses were subsequently reconstructed into images using Miyawaki's reconstruction algorithm. The reconstructed images generated from the predicted or converted fMRI responses of a source subject closely resembled those obtained from the measured fMRI responses of a target subject (Figure 1.9). These findings indicated that the information pertaining to image contrasts is shared across individuals.

1.3.2 Object identity

The human visual system could recognize objects in a variety of situations, in spite of the variation in the physical stimulus. Mishkin and Ungerleider (1982) proposed the concept of "what" and "where" pathways within the visual system that process different types of information. The "what" pathway, also known as the ventral visual pathway, encompasses the occipital lobes and several regions within the temporal lobe. Neurons within the posterior

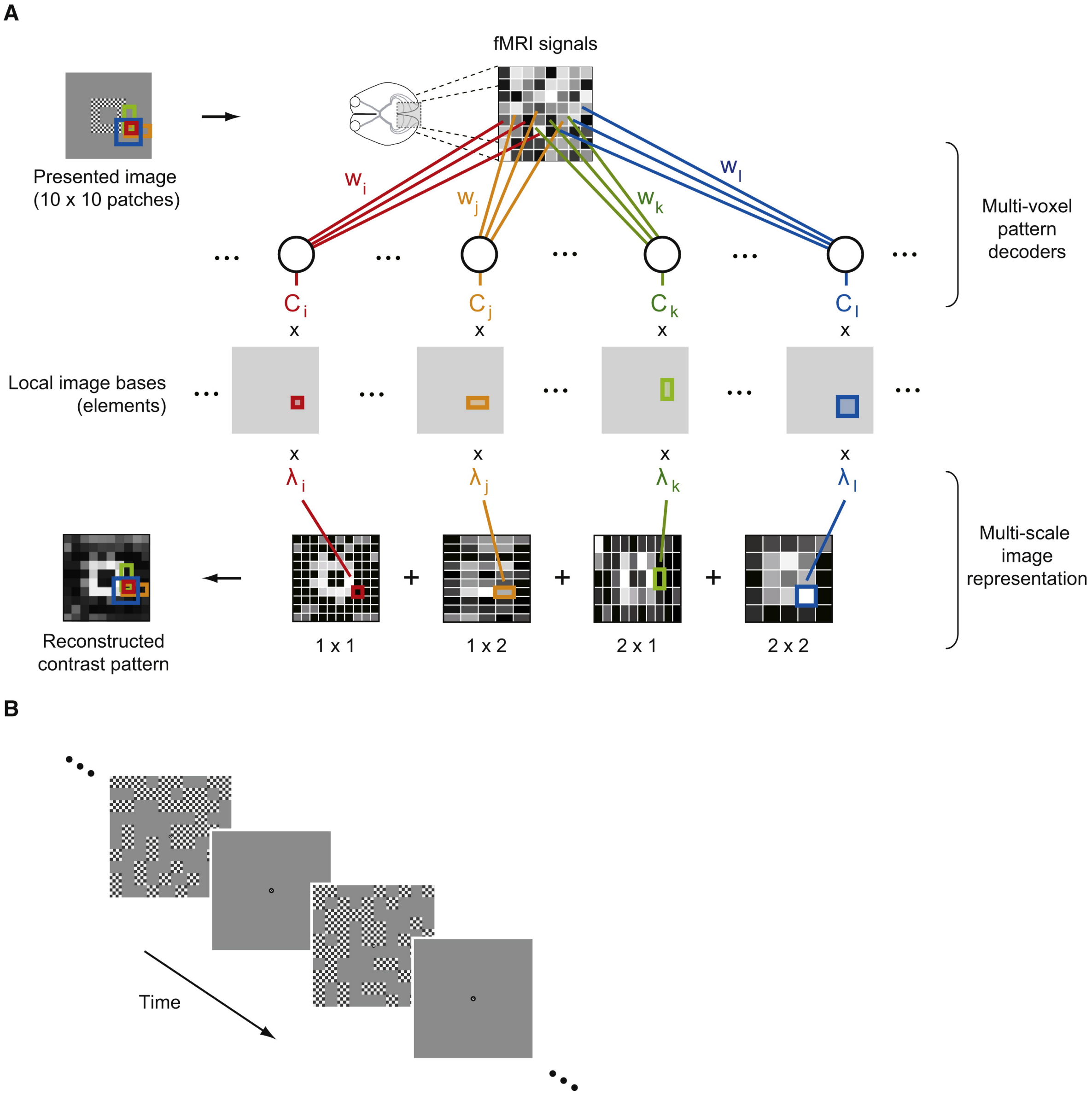

Figure 1.8: Visual image reconstruction by multiscale local image decoders. Panel A indicates the prediction and combination of multiscale local image bases, and the reconstruction of the contrast pattern. Four decoders corresponding to four local image bases (1×1, 1×2, 2×1, 2×2) are trained at each location of an image. Panel B shows the experiment design with the stimuli flashed over the time. Reprinted from *Neuron*, 2008;60, Miyawaki et al., Visual Image Reconstruction from Human Brain Activity using a Combination of Multiscale Local Image Decoders, Pages No. 917, Copyright (2008), with permission from Elsevier.

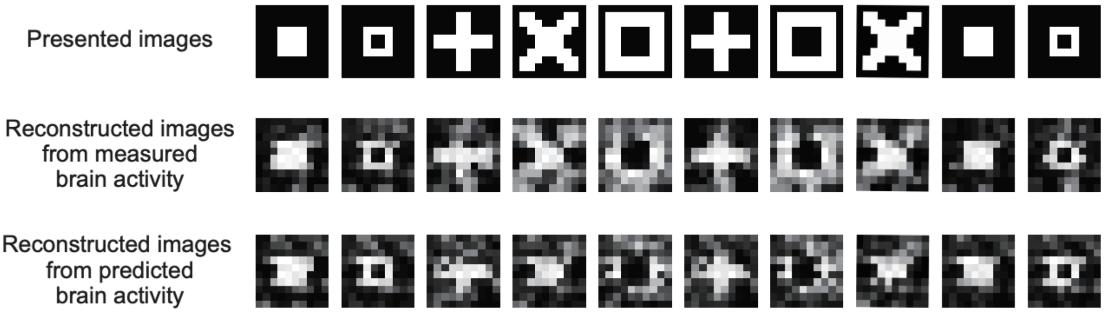
Presented images
Reconstructed images from measured brain activity
Reconstructed images from predicted brain activity

Figure 1.9: Reconstructed artificial images from the converted brain activity patterns of V1 area. Reprinted from *NeuroImage*, 2015;113, Yamada et al., Inter-subject neural code converter for visual image representation, Page No. 296, Copyright (2015), with permission from Elsevier.

region of this pathway exhibit selectivity towards basic features such as edges and contrasts. In contrast, neurons in the inferior temporal (IT) cortex demonstrate selectivity towards more complex features, such as human body parts (Desimone et al., 1984). Thus, it is widely accepted that neural representations within the IT cortex encode information pertaining to object identity.

Within the ventral stream, certain regions exhibit selectivity towards specific visual features that play a crucial role in object recognition. For instance, the lateral occipital complex (LOC) is instrumental in shape recognition (Kourtzi & Kanwisher, 2000), while the fusiform face area (FFA), located in the ventral part of the temporal lobe, is specifically involved in the recognition of faces (Kanwisher et al., 1997). Additionally, the parahippocampal place area (PPA), situated within the parahippocampal cortex, is crucial for scene perception (Epstein & Kanwisher, 1998).

In the research conducted by Haxby et al. (2011), hyperalignment was employed on fMRI response vectors derived from the visual area, which encompasses the FFA, PPA, and a region of the inferior temporal cortex (IT) (refer to section 1.2.1: Hyperalignment). The hyperalignment procedure used a training dataset consisting of fMRI responses collected while subjects viewed a full-length movie. Subsequently, the additional fMRI datasets of the subjects, acquired while they were exposed to face and simple object stimuli, were transformed into the estimated common model space obtained through hyperalignment.

To investigate the shared neural representations of object identities, a linear support vector machine (SVM) was trained using the transformed dataset of a particular subject, with the aim of predicting the stimulus category, including male faces, female faces, monkey faces, dog faces, shoes, chairs, and houses. The performance of the SVM was evaluated

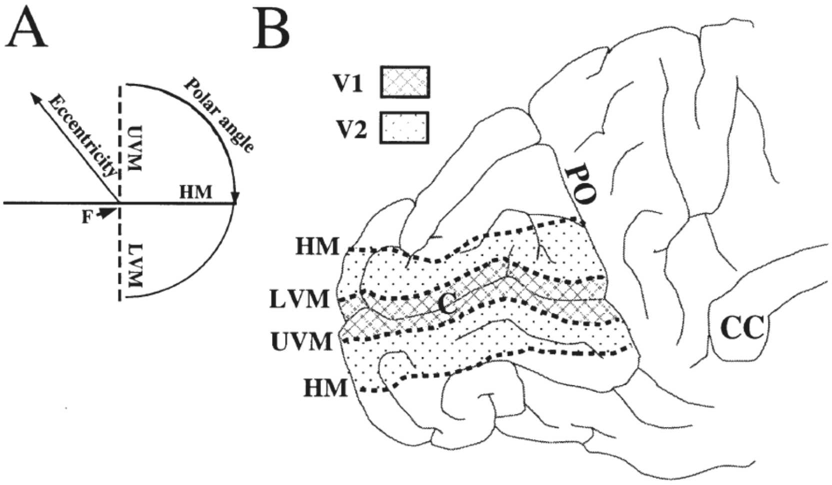

Figure 1.10: Retinotopic mapping from the visual field to the visual cortex. Panel A shows the visual field defined by polar coordinate. Panel B shows the corresponding mapping from the visual field to the visual areas V1 and V2. Reprinted from *Cerebral Cortex*, 1997;7(2), Engel et al., Retinotopic organization in human visual cortex and the spatial precision of functional MRI, Page No. 182, Copyright (1997), with permission from Oxford University Press.

using two distinct scenarios: within-subject classification, where the SVM was tested on the same subject's dataset, and between-subject classification, where the SVM was tested on a different subject's dataset. Notably, between-subject classification yielded results comparable to within-subject classification, indicating that the complex features underlying object identity are indeed shared across individuals.

1.3.3 Retinotopic organization

Retinotopy is the systematic mapping of visual stimuli from the retina to the visual cortex, as described by Engel et al. (1994 & 1997). This organization is distinctly seen in the primate primary visual cortex, known as V1. The visual field is defined utilizing a polar coordinate system, where the dimension extending from the center to the periphery is termed eccentricity, while the dimension traversing from the upper vertical meridian (UVM) through the horizontal meridian (HM) to the lower vertical meridian (LVM) is known as the polar angle (see Figure 1.10).

Within V1, retinotopic mapping manifests in the following manner: as a visual stimulus moves along the eccentricity dimension from the center, neural activity undergoes a spatial shift from the posterior to the anterior cortex. Likewise, as the stimulus traverses the polar angle dimension, progressing from UVM to LVM, neural activity within the calcarine sulcus undergoes a transition from the ventral to the dorsal part. Furthermore, retinotopic mapping exhibits a contralateral organization, whereby stimuli presented in the left or right visual field elicit corresponding neural activity in the right or left visual cortex, respectively. As a result, the primary visual cortices of the two brain hemispheres jointly cover the entire visual field.

The initial evidence for the preservation of retinotopic organization after functional alignment was provided by Yamada et al. (2011 & 2015; see section 1.2.2: Neural code converter). Specifically, they observed that the transformation matrix consistently assigned higher weights to source subject voxels exhibiting similar eccentricity and polar angle as the target voxel. This finding implies that when a stimulus traverses the visual field of the source subject, it generates a "pseudo" neural activity that replicates the retinotopic pattern within the corresponding location of the target subject's visual cortex. However, their investigation examined solely on the preservation of retinotopic organization in area V1. Subsequently, Bilenko and Gallant (2016) and Guntupalli et al. (2016) extended these findings by a different way and presented further evidence supporting the preservation of retinotopic organization following functional alignment beyond area V1.

1.3.4 Semantic contents

Natural scenes present a rich and dynamic environment filled with a diverse array of objects, actions, and interactions. The complexity of these scenes goes beyond the mere identification of individual objects and requires a deeper understanding of the semantic relationships and contextual information present. The semantic representations associated with natural scenes are not confined to specific localized regions in the brain but rather distributed across extensive neural networks (Huth et al., 2016).

The distributed nature of semantic representations in the brain can be attributed to the broad range of underlying concepts involved. For example, when perceiving a natural scene, our brain processes not only the objects present but also the actions being performed and the relationships between objects. Consider a scene depicting a group of people having a picnic in a park. Our understanding of the scene goes beyond recognizing the individuals and the picnic items; it also involves comprehending the social interaction, the leisurely atmosphere, and the spatial context of the park environment. All these interconnected semantic components contribute to the holistic interpretation of the scene.

In a study by Van Uden et al. (2018), the researchers measured fMRI brain responses from 18 subjects as they freely watched a naturalistic audiovisual movie (Figure 1.11). Each imaging volume was assigned word embeddings (word2vec) related to agents, actions, objects, and scenes to capture semantic information. The researchers performed searchlight hyperalignment using the collected fMRI data to estimate a common model space. They transformed data from 17 of 18 subjects to a left-out subject's space through the common model space, and trained a semantic encoding model to predict the left-out subject's fMRI responses. The researchers used a separate dataset not involved in the hyperalignment process to compare the prediction accuracy with the actual fMRI responses from the left-out subject, by evaluating the correlation between them. Remarkably, the between-subject model exhibited similar accuracy to that of the within-subject model, in which the encoding model was trained only with the fMRI responses of the left-out subject. This indicated that despite individual life experience variations, the semantic space organizing semantic concepts in certain patterns is universally shared.

1.4 Hierarchical visual processing in the ventral pathway

Human vision follows a hierarchical process that seamlessly translates light waves into vivid, meaningful perception. The magic begins in the retina, where light stimulates photoreceptor cells to convert this physical energy into electrical signals. These signals embark on a complex journey through the visual system's intricate layers, offering us a glimpse into our surrounding world.

The visual system follows a hierarchical organization that progressively breaks down and processes visual information in stages. Early stages involve processing basic visual elements such as color, brightness, and edge orientation, primarily in the retina and primary visual cortex (V1). Information then advances to higher visual areas, where more complex attributes such as object recognition, motion perception, and spatial awareness are processed. This stage-wise processing allows the system to construct a comprehensive visual understanding from simple components, resembling a pyramid of cognition with intricate details at the base and broader concepts at the peak.

This layered hierarchy, where higher levels integrate and interpret information from lower ones, serves as the cornerstone of our visual perception. It allows us to navigate, interact with, and make sense of our environment. The architecture of the hierarchical visual processing system is not only a testament to the sophistication of human physiology but also provides a blueprint for developing advanced artificial vision systems. In this section, I introduced

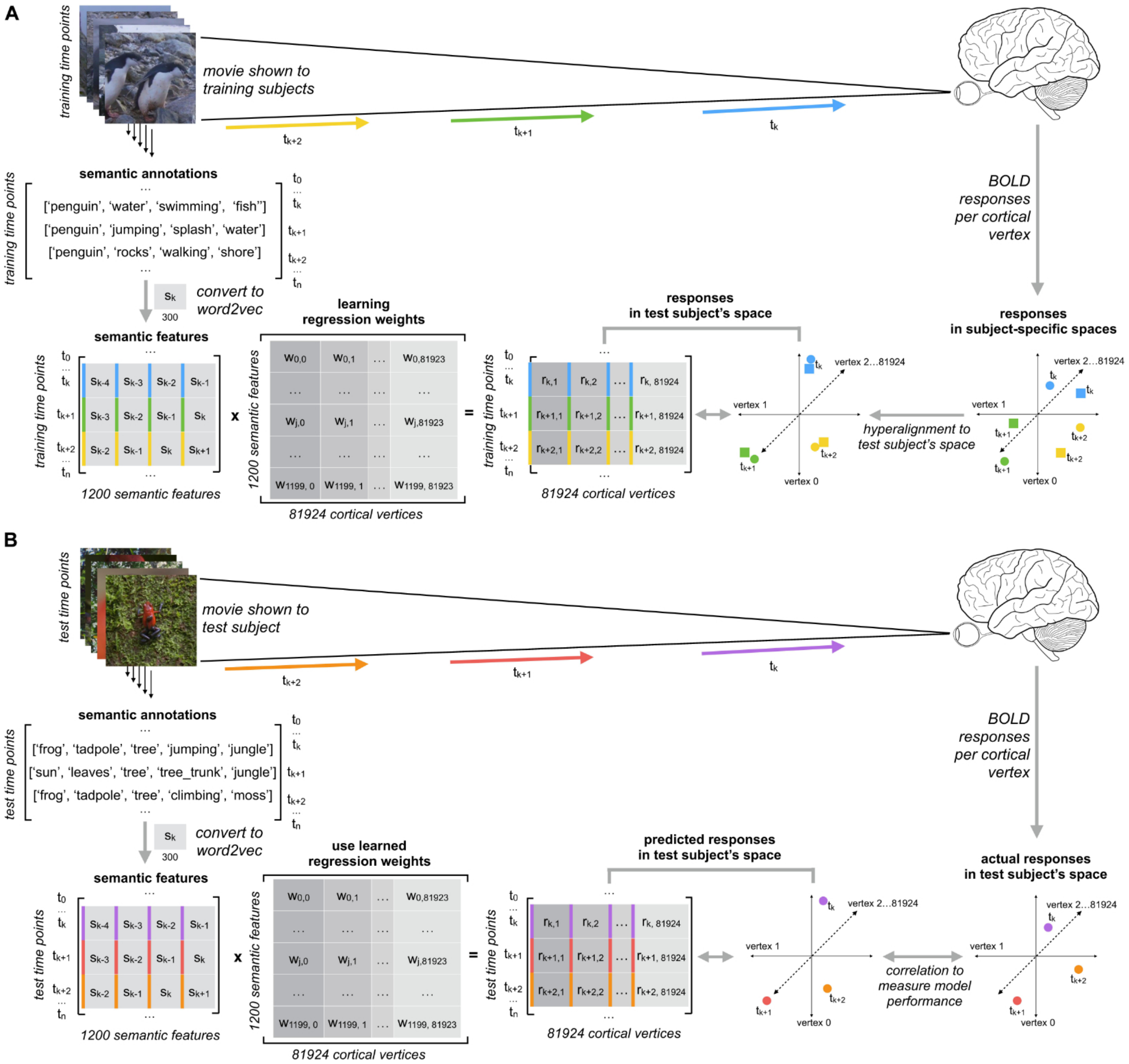
A
training time points
movie shown to training subjects
t_{k+2}
t_{k+1}
t_k
semantic annotations
training time points
['penguin', 'water', 'swimming', 'fish"]
['penguin', 'jumping', 'splash', 'water']
['penguin', 'rocks', 'walking', 'shore']
t_0 … t_k t_{k+1} t_{k+2} … t_n
BOLD responses per cortical vertex
S_k 300
convert to word2vec
semantic features
learning regression weights
responses in test subject's space
responses in subject-specific spaces
S_{k-4} S_{k-3} S_{k-2} S_{k-1}
S_{k-3} S_{k-2} S_{k-1} S_k
S_{k-2} S_{k-1} S_k S_{k+1}
1200 semantic features
x
$W_{0,0}$ $W_{0,1}$ … $W_{0,81923}$
$W_{j,0}$ $W_{j,1}$ … $W_{j,81923}$
$W_{1199,0}$ $W_{1199,1}$ … $W_{1199,81923}$
1200 semantic features
81924 cortical vertices
=
$r_{k,1}$ $r_{k,2}$ … $r_{k,81924}$
$r_{k+1,1}$ $r_{k+1,2}$ … $r_{k+1,81924}$
$r_{k+2,1}$ $r_{k+2,2}$ … $r_{k+2,81924}$
81924 cortical vertices
vertex 2…81924
vertex 1
vertex 0
hyperalignment to test subject's space
B
test time points
movie shown to test subject
t_{k+2}
t_{k+1}
t_k
semantic annotations
test time points
['frog', 'tadpole', 'tree', 'jumping', 'jungle']
['sun', 'leaves', 'tree', 'tree_trunk', 'jungle']
['frog', 'tadpole', 'tree', 'climbing', 'moss']
BOLD responses per cortical vertex
S_k 300
convert to word2vec
semantic features
use learned regression weights
predicted responses in test subject's space
actual responses in test subject's space
1200 semantic features
81924 cortical vertices
correlation to measure model performance

Figure 1.11: Illustration for building between-subject semantics encoding models with hyperalignment. Panel A illustrates the model training process: the training subjects' data are first mapped into the test subject's space, and then a linear regression model is trained to predict responses in the test subject's space. Panel B depicts the model testing process, where a left-out movie, viewed by the test subject, is used to predict fMRI responses within the test subject's space. Reproduced from *Frontiers in Neuroscience*, 2018;12: 437, Van Uden CE et al., Modeling Semantic Encoding in a Common Neural Representational Space, Page No. 5, licensed under CC BY 4.0 / Cropped from original.

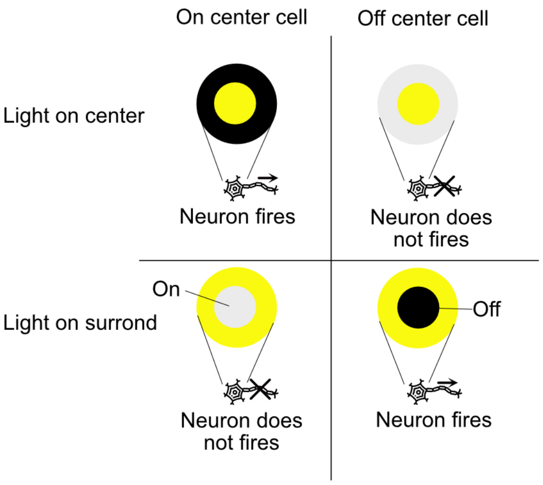

Figure 1.12: Schematic of the center-surround receptive field.

several visual features in the ventral visual stream, with a focus on the hierarchical nature of their processing.

1.4.1 Contrast

At the heart of visual interpretation lies the ability to distinguish objects from their backgrounds and recognize the boundaries between different areas in our field of vision. This fundamental ability, known as contrast perception, is crucial for detecting edges and delineating forms in our visual surroundings.

Contrast perception commences at the level of the retina, where specialized cells called photoreceptors detect the light that reaches our eyes. These cells, comprising rods and cones, convert light into neural signals, which are processed by subsequent layers of cells in the retina. The retinal ganglion cells, particularly, play a key role in contrast perception due to their unique receptive field structure.

Each ganglion cell possesses a receptive field that comprises two parts - a central region (center) and a surrounding region (surround). Depending on the cell type, light stimulation in the center could either excite or inhibit the cell's firing, while stimulation in the surround has the opposite effect (Figure 1.12). This center-surround organization enhances the contrast at edges where light intensity changes significantly, facilitating the detection of boundaries.

The neural signals, now contrast-enhanced, travel via the optic nerve to the lateral geniculate nucleus (LGN) in the thalamus, where further processing occurs. The neurons in the LGN maintain the center-surround structure of their receptive fields, further enhancing the contrast information received from the retina.

Following the LGN, this information is relayed to the primary visual cortex (V1), where neurons known as simple and complex cells take over. These cells, particularly sensitive to edges and orientation, combine the contrast information from multiple LGN inputs. Through this integration, they can detect and enhance contrast along specific orientations, a critical step for shape recognition and pattern detection.

In summary, contrast perception is an intricate and vital process, seamlessly coordinating between different stages of the visual pathway. From the initial detection of light in the retina to complex processing in the visual cortex, the collective effort results in our ability to distinguish boundaries, recognize shapes, and perceive patterns. It is a testament to the sophistication of our visual system and its ability to decode the nuanced tapestry of our visual world.

1.4.2 Orientation

The primary visual cortex (V1), also known as the striate cortex, is the first cerebral cortex region to receive visual input. Two main cell types within V1, simple and complex cells, work together to interpret and respond to orientation.

Simple cells, first discovered by David Hubel and Torsten Wiesel (Hubel and Wiesel, 1959; 1962), respond most strongly to oriented edges and bars of specific orientations within their receptive fields. Each simple cell possesses a distinct receptive field that is spatially organized into “on” and “off” regions (Figure 1.13). The receptive field of simple cells is suggested to be formed by overlapping multiple receptive fields of LGN cells. When light falls onto an “on” region, it stimulates the cell, and when it falls onto an “off” region, it inhibits the cell. This arrangement allows simple cells to respond maximally when a line or edge of a specific orientation is present in their receptive field. Thus, they serve as the building blocks for edge detection and orientation selectivity, critical aspects of visual perception.

On the other hand, complex cells, another discovery by Hubel and Wiesel, show a higher degree of abstraction in their responses. Unlike simple cells, complex cells respond to oriented edges and bars across a broad spatial area and are insensitive to the exact location of the stimulus in the receptive field. Furthermore, they often exhibit a preference for motion in a particular direction. This insensitivity to the specific position and sensitivity to the direction of motion allows complex cells to process more global aspects of the visual scene, such as object movements and broader shapes.

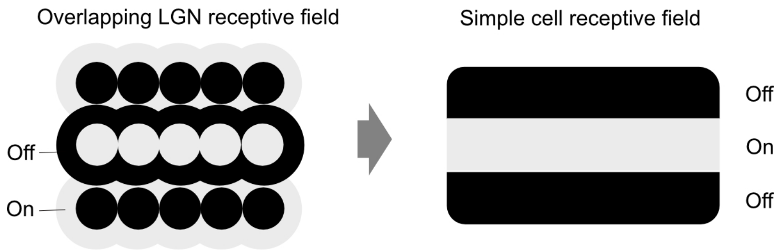

Figure 1.13: Schematic of the formation of the receptive field of a simple cell.

Together, simple and complex cells in the visual cortex play complementary roles in visual perception. Simple cells initiate the process by detecting edges and orientations, effectively delineating objects' boundaries in the visual field. Complex cells then take this processed information a step further, integrating these boundaries over larger areas and tracking their movement over time.

In conclusion, the discovery and study of simple and complex cells have significantly contributed to our understanding of the visual system. They serve as prime examples of how the brain processes information hierarchically, beginning with simple features and progressively constructing a more comprehensive representation of the world. Their precise functions in vision provide insights not only into the human visual system but also into principles of neural computation and organization that can be extended to other sensory systems and cognitive functions.

1.4.3 Depth

As inhabitants of a three-dimensional world, our ability to perceive depth is vital for interpreting the spatial relationships between objects, estimating distances, and navigating through our environment. The underlying biological mechanisms enabling depth perception are complex, engaging multiple stages of visual processing, culminating in certain specialized areas of the visual cortex.

Depth perception initiates with the reception of visual stimuli by the eyes. Due to their horizontal separation on our faces, each eye views the world from a slightly different angle, resulting in two slightly different images being projected onto each retina. This discrepancy, known as binocular disparity, provides one of the most critical cues for depth perception.

The process begins in the retina and is then transmitted via the optic nerve to the lateral geniculate nucleus, and eventually to the primary visual cortex (V1). While V1 does have neurons that respond to binocular disparity (Poggio et al., 1985), the real depth computation appears to occur beyond this initial stage of cortical processing.

This further processing of depth information takes place in specialized visual areas, including V2, V3, and also in an area known as V5/MT (Parker, 2007). These areas contain neurons that respond selectively to particular degrees of binocular disparity, effectively enabling them to compute the relative depth of objects in our visual field.

In addition to binocular disparity, other cues such as perspective, shadows, relative size, and occlusion, contribute to depth perception. While these cues are primarily monocular, meaning they can be interpreted with just one eye, they still influence the processing in these disparity-selective regions of the cortex, offering a more complete and robust perception of depth.

In conclusion, depth perception is a multifaceted process that encompasses both binocular and monocular cues and engages multiple regions of the visual cortex. The processing of these depth cues, particularly binocular disparity, allow us to navigate effectively through our three-dimensional world. This intricate interplay between the different stages of visual processing underscores the complexity of our visual system and its remarkable ability to create a rich, three-dimensional interpretation of our surroundings.

1.4.4 Color

Color perception is a fundamental component of our visual system. At the onset of color perception are the photoreceptor cells in the retina known as cones. Humans typically have three types of cone cells, each sensitive to different wavelength ranges that broadly correspond to the colors blue, green, and red. The differential stimulation of these cones by various wavelengths of light sets the stage for color perception.

Following the initial detection, the signals from the cones are processed further by the retinal ganglion cells. These cells generate responses based on the differences in signals from the various types of cones, encoding the color information into two dimensions: brightness (black-white) and color (blue-yellow and red-green). This encoded information is then sent via the optic nerve to the LGN in the thalamus for further processing.

From the LGN, the color-coded signals are relayed to the primary visual cortex (V1), where they are interpreted by specialized cells. The neurons in V1, especially the double-opponent cells, respond to color contrasts within their receptive fields, which allows them to identify color edges and small color patterns (Livingstone & Hubel, 1984). This processing of color contrast is crucial for recognizing color boundaries and enhancing color perception.

Beyond V1, higher areas of the visual cortex, such as the V4 (Zeki, 1973) and the inferior temporal cortex (Komatsu et al., 1992), also play a crucial role in color perception. Particularly, area V4 has been linked to color constancy - our ability to perceive the consistent color of an object despite changes in lighting conditions (Wild et al., 1985). These areas integrate the color information with other visual attributes, contributing to our ability to recognize objects and perceive a coherent visual scene.

In conclusion, color perception is a multifaceted process that spans several stages of the visual pathway. From the initial detection by the cones in the retina to sophisticated processing in the visual cortex, this color decoding capability endows us with a rich and nuanced understanding of our environment. It adds depth to our visual experience, allows us to discern objects, and even provides cues about the emotional state of our surroundings. This colorful journey underlines the complexity and precision of our visual system in transforming light into a visual feast of colors.

1.4.5 Contour integration

Contour integration begins in the primary visual cortex (V1) where the first stage of visual processing occurs. Here, neurons known as simple and complex cells respond to specific orientations of light and dark contrasts in their receptive fields. While these cells are adept at detecting local features such as edges and bars, they lack the capacity to perceive how these isolated elements relate to one another to form larger structures.

This is where the process of contour integration steps in. In areas beyond V1, such as the secondary visual cortex (V2) and other associated areas (Anzai et al., 2007; Hegdé and Van Essen, 2000; Ito and Komatsu, 2004), more complex cells respond to aligned edges and can start to link these edges together to form perceived contours. This process is thought to be aided by lateral connections between neurons that allow them to “communicate” and establish relationships between their receptive fields (Yen & Finkel, 1998).

These neurons exhibit a property called “end-stopping” (Hubel & Wiesel, 1965), meaning they prefer line segments of a specific length. End-stopped cells are critical for contour integration as they help determine where a line or edge ends, enabling the visual system to differentiate between separate contours in the visual field.

The integration of contours is a critical intermediate step in the hierarchy of visual processing, forming the bridge between low-level feature detection and high-level object recognition. It aids in distinguishing objects from their backgrounds and recognizing the boundaries of objects, ultimately contributing to our perception of shapes, scenes, and structures.

In conclusion, contour integration is a powerful tool that our visual system employs to make sense of the complex visual world. By stitching together individual elements into larger structures, contour integration allows us to perceive the outlines of objects and understand their form. This process not only illustrates the complexity of our visual system but also showcases its remarkable ability to transform fragments of visual information into coherent and recognizable images.

1.4.6 Texture perception

The neural mechanisms driving texture perception function in a hierarchical manner within the human visual system. The initial processing of visual information, including basic texture features, occurs in V1. Cells in this region respond to simple attributes such as edge orientation and spatial frequency, elements that are essential to texture perception. However, the understanding of texture transcends these rudimentary characteristics; our perception integrates these features into a more holistic understanding of surfaces and materials. Subsequent stages of visual processing in areas such as V2 and V4 are believed to play a crucial role in this integration (Kastner et al., 2000; Puce et al., 1996. Here, neurons respond to more complex patterns and combinations of features, essentially “composing” the textures we perceive from the simpler elements detected by V1.

Texture assists in segmenting a visual scene (Julesz, 1981), helping us differentiate between objects and their surroundings, an essential aspect of successful object recognition. Furthermore, the consistency of a particular texture across the surface of an object provides crucial information about the object’s shape and spatial orientation. It is through this intricate interplay between texture perception and object recognition that we can navigate and understand our richly textured world.

On the other hand, the role of texture is also important in object recognition (Vaina, 1987). Texture provides essential cues about an object’s identity and its spatial properties, contributing significantly to our perception of depth and three-dimensionality. By offering information about surface characteristics and material composition, texture allows us to differentiate between objects that may otherwise share similar shapes or colors. For instance, the visual system can distinguish between a marble statue and a wooden carving of identical form, primarily through differences in their textures.

Nevertheless, texture, representative of mid-level visual features, remains more enigmatic compared to low-level visual features. The primary reason for this knowledge gap is the inherent complexity associated with formulating a mathematical model for texture, an endeavor that is relatively more straightforward in the context of low-level visual features. Thus, the systematic exploration and modeling of texture, as a mid-level visual feature, presents a significant challenge in the realm of visual perception research.

1.4.7 Object recognition

Object recognition, lied on top of the ventral visual pathway, is a task that synthesizes information from multiple sources and stages to generate coherent perceptions. This section describe the specific roles of the lateral occipital complex (LOC), fusiform face area (FFA), parahippocampal place area (PPA), and inferior temporal (IT) cortex in contributing to the phenomenon of object recognition.

The LOC is a central player in the object recognition process, primarily involved in the detection and perception of shapes (Kourtzi & Kanwisher, 2000). Situated in the occipito-temporal region, this cortical area displays heightened activity in response to a variety of objects, signifying its critical role in the recognition and interpretation of an array of visual stimuli. By facilitating the discernment of distinct shapes, the LOC paves the way for invariant object recognition, allowing us to identify objects regardless of alterations in perspective or size.

In contrast, the FFA is renowned for its specialization in face perception (Kanwisher et al., 1997). Positioned within the fusiform gyrus on the ventral surface of the brain, the FFA is particularly activated during tasks involving facial recognition, demonstrating its crucial role in discerning the complex and unique features that define individual faces. This specialization underscores the evolutionary importance of face recognition in social communication and interaction.

Adjacent to the FFA lies the PPA, a region within the medial temporal lobe known for its predilection for scene recognition (Epstein & Kanwisher, 1998). The PPA exhibits robust activity in response to landscapes, cityscapes, or rooms, suggesting its role in spatial awareness and navigation. This area aids us in understanding the spatial layout or context of an environment, contributing to the greater cognitive map of our surroundings.

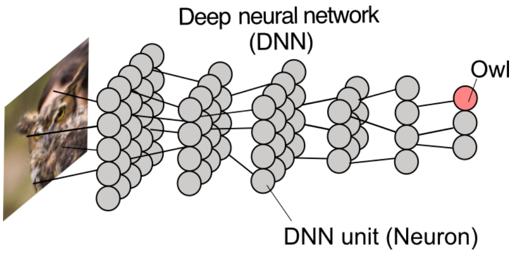

Figure 1.14: Schematic of deep neural network

Finally, the IT cortex, situated in the ventral stream of the brain, is involved in the final stages of object recognition. The neurons in this region respond to complex, high-level visual features, assembling the various elements processed in preceding regions into a cohesive and detailed object representation.

In summary, object recognition emerges from the synergistic activity of these distinct yet interconnected visual areas. The LOC, FFA, PPA, and IT cortex each perform specialized functions, processing different aspects of visual stimuli. Yet, it is the consolidation of these processed elements that culminates in the formation of an object's identity. The brain's ability to synthesize this wealth of information into a coherent perception underscores the intricate complexity of the visual processing system and its remarkable capacity for object recognition.

1.5 Artificial deep neural network (DNN)

1.5.1 Basics of DNN

Deep Neural Networks (DNNs) have dramatically transformed the landscape of artificial intelligence, rapidly becoming the cornerstone of numerous cutting-edge technological advancements. Inspired by the sophisticated circuitry of the human brain, these complex models employ layers of interconnected nodes or neurons to process, learn, and infer from a vast array of information. These layers, stacked in a hierarchical structure, constitute the "deep" in Deep Neural Networks (Figure 1.14), enabling them to model high-level abstractions in data with a remarkable degree of accuracy.

At their heart, DNNs are an exemplification of machine learning's core principle: learning from data. They excel at identifying patterns and relationships within datasets, dynamically adjusting their internal parameters based on the input they receive. This enables them to draw connections that may be too subtle, complex, or multi-dimensional for traditional algorithms to detect. Furthermore, DNNs are equipped with the powerful capability of automatic

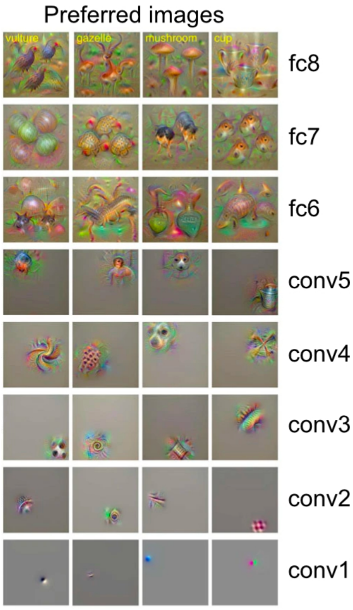

Figure 1.15: Examples of images preferred by four randomly selected units in each DNN layer. Reproduced from *Nature Communications*, 2017;8: 15037, Horikawa and Kamitani, Generic decoding of seen and imagined objects using hierarchical visual features, Page No. 5, licensed under CC BY 4.0 / Cropped from original.

feature extraction. The DNN units in the lower-layers always prefer simple features, such as orientation, exhibiting a striking similarity with the simple cells in the primary visual cortex, while the DNN units in the higher layers prefer complex attributes, such as object identity (Figure 1.15). Unlike their more conventional counterparts that rely on explicitly programmed features, DNNs can independently discover and learn the significant features needed to make accurate predictions.

In recent years, one of the most potent applications of DNNs has been in the realm of computer vision, particularly object recognition (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014). Tasked with the detection and classification of objects within digital images or videos, object recognition is a significant step towards creating machines that perceive the world as humans do. DNNs, with their ability to understand complex patterns, have significantly amplified the accuracy of object recognition. The layers of a DNN gradually distill raw input into meaningful abstract representations, enabling the system to discern objects and patterns that are often invisible to the human eye.

1.5.2 Brain and DNN

Previous sections has illuminated a number of similarities between the human visual system and Deep Neural Networks (DNNs). Notably, both of these systems interpret and process data in a hierarchical fashion, assembling features from preceding inputs. This suggests that DNNs hold considerable promise as effective models for emulating the neural representations present in the visual cortex.

In order to investigate whether the visual system operates in a manner analogous to DNNs, Yamins et al. (2014) trained a hierarchical DNN to achieve human-level performance in recognition tasks involving the classification of animals, boats, cars, and other objects. Despite not being directly constrained by neural data, the features extracted from the output layer of the DNN were found to be predictive of neural responses in the inferior temporal (IT) cortex, while features from the middle layer were predictive of neural responses in area V4. This outcome suggests that DNNs, optimized for object recognition, have the potential to serve as predictive models for neural processing.

Furthering this notion, Güçlü and van Gerven (2015) extended the analysis to the ventral visual pathway by predicting fMRI responses using activations from deep neural networks, specifically the AlexNet and VGG models. The voxel responses in lower and higher visual areas exhibited a favorable correlation with features extracted from corresponding lower and higher layers of the CNN. This investigation was further substantiated from a decoding perspective, where the DNN layer features were predicted based on the fMRI responses (Horikawa & Kamitani, 2017). This suggests that the lower and higher visual areas encode visual features that are either correlated or share similarities with the features extracted from corresponding lower and higher layers of the CNN.

Subsequent research conducted by Güçlü and van Gerven (2017) expanded upon their earlier investigations to include the dorsal visual pathway, which includes areas V1, V2, V3, V3A, V3B, and MT. They demonstrated that deep neural network (DNN) features derived from videos could effectively predict the hyperaligned responses of the dorsal visual pathway within a common model space. This finding hints at a degree of shared representation across individuals within the dorsal visual pathway. However, it is important to note that the hierarchical organization along the dorsal visual pathway was assumed in the region-of-interest (ROI) hyperalignment, thereby making it difficult to definitively claim that this “hierarchy” is universally shared. Consequently, the question of whether visual features along the visual pathway and their associated "hierarchy" are shared across individuals remains unresolved.

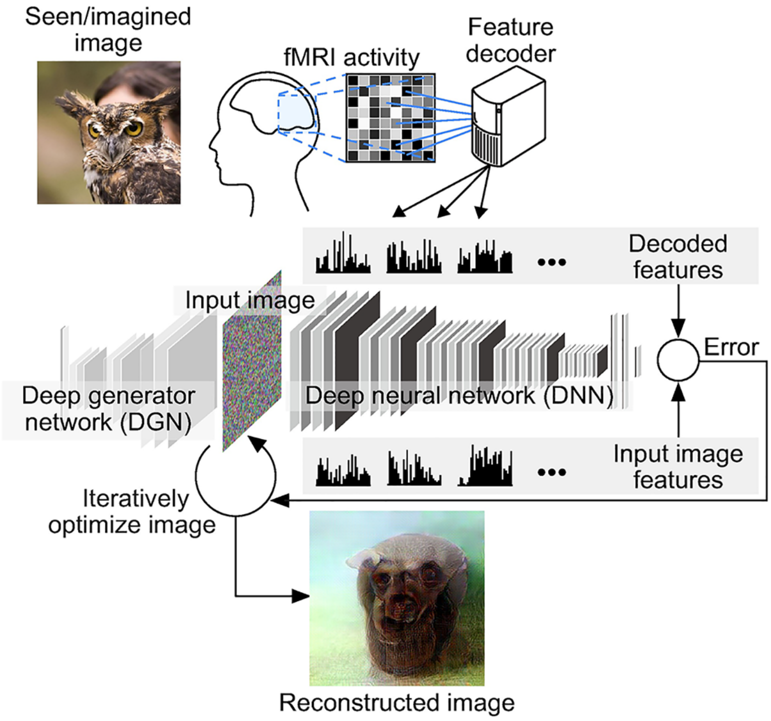

Figure 1.16: Deep image reconstruction. DNN features of a seen image is predicted from the subject's fMRI activity by feature decoders. A reconstructed image is generated by iteratively minimizing the error between the image features and the decoded features. Through this iterative process, the hierarchical visual information retained in the decoded features is incorporated into the reconstructed image. Reprinted from *PLoS Computational Biology*, 2019;15: 1006633, Shen et al., Deep image reconstruction from human brain activity, Page No. 3, licensed under CC BY 4.0.

Building upon the work of Horikawa and Kamitani (2017), the decoded hierarchical DNN features can be further reconstructed as visual images (Figure 1.16; Shen et al., 2019a). The successful reconstruction of natural images implies that the decoded DNN features encompass a rich set of intricate hierarchical visual features. Subsequently, visual image reconstruction has emerged as a prominent research area, with several alternative methods proposed (Güçlütürk et al., 2017; Shen et al., 2019b; Han et al., 2019; Seeliger et al., 2018; see Rakhimberdina et al., 2021 for a survey). However, it remains an open question as to whether the hierarchical and fine-grained visual representation enabling visual image reconstruction are universally shared across individuals (Figure 1.17).

1.5.3 Why are DNNs good for modeling the visual system?

The section 1.4 presents an exploration of visual features, from those elementary in nature (low-level) to those imbued with complexity (high-level). While this introduction does not

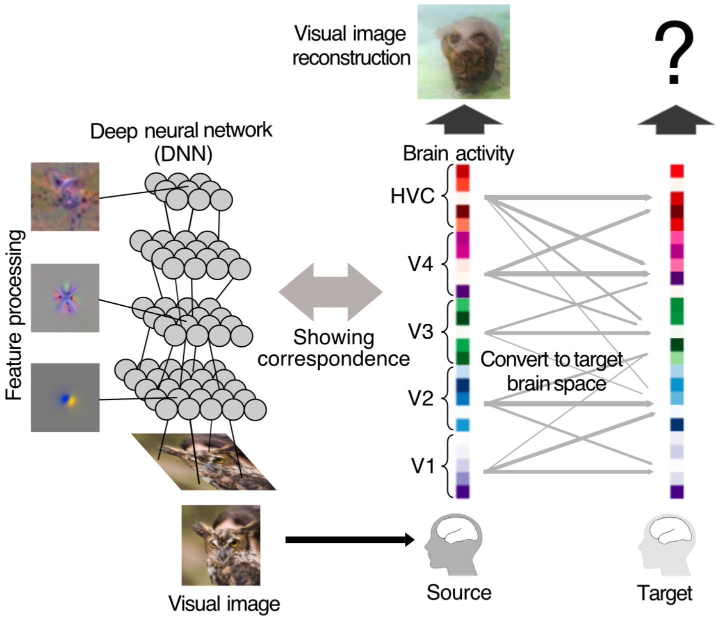

Figure 1.17: Illustration of preservation of the hierarchical and fine-grained visual features.

purport to present a comprehensive review of hierarchical visual representation, it offers an illustration of the progressive process by which our brain constructs our perception. A notable ambiguity within this spectrum of visual representations lies in the domain of middle-level visual representation, a realm that remains relatively underexplored (Peirce, 2015).

For low-level features, mathematical models have been devised, offering insights into the mechanisms through which these features derive representations from earlier inputs (Hubel and Wiesel, 1962; Adelson and Bergen, 1985). High-level visual features, conversely, lack a physiological-based model, yet humans are undeniably capable of articulating the semantic elements within a scene, such as identifying if there is tea in a cup.

There is an ongoing debate regarding the existence of mid-level visual representation. Many researchers maintain the view that these features pose significant challenges to modeling efforts or representation through conventional human intuition and language. The absence of a robust model for mid-level visual features further complicates the task of ascertaining whether a hierarchical visual representation is a universal phenomenon shared among individuals.

Deep Neural Networks (DNNs), designed with the objective of object recognition, process images in a hierarchical fashion akin to the workings of the human brain. In addition, DNNs exhibit layer-specific preferences for distinct types of images, which range from basic oriented edges to complex attributes such as faces, culminating in object identification.

The features of the lower and upper layers of a DNN echo the attributes of low-level and high-level visual features, respectively. The characteristics of a DNN's middle layers may potentially serve as viable models for approximating mid-level visual features. While it remains uncertain whether the features of DNN's middle layers are an exact match for mid-level visual features, existing studies indicate that these DNN features can predict neural activity in the mid-visual areas, such as V4 (Yamins et al., 2014; Güçlü and van Gerven, 2015; Horikawa and Kamitani, 2017). This points to the potential of middle-layer DNN features serving as a credible model.

In summary, the utilization of DNNs for modeling visual features provides a distinct advantage, particularly in the context of mid-level visual features, which are notoriously difficult to model using traditional mathematical models or semantic-based approaches. Leveraging the capabilities of DNNs could help demystify the complexities of mid-level visual features and contribute towards a more profound understanding of hierarchical visual representation.

1.6 Thesis organization

The thesis is organized as follows to investigate whether fine-grained hierarchical visual information that enables visual image reconstruction can be retained after functional alignment. Chapter 2 introduces the human brain activity data and the experiment details. In Chapter 3, neural code conversions are performed between individuals, followed by an examination of the automatically detected cortical hierarchy. In Chapter 4, the converted fMRI brain activity patterns are then translated into hierarchical DNN features, evaluating the extent of hierarchy using brain hierarchy scores. The preservation of rich visual information is demonstrated through the reconstruction of decoded DNN features into visual images. Chapter 5 explicitly compares neural code conversions trained without imposing visual hierarchy to those respecting the visual hierarchy, highlighting the ability of data-driven approaches to detect the cortical hierarchy effectively. In Chapter 6, data from multiple subjects are pooled into a target subject's space, leading to slightly improved decoders as observed through visual image reconstruction evaluation. Lastly, Chapter 7 discusses the implication of neural code converters, future directions concerning inter-individual visual image reconstructions and explores other visual features possibly shared across individuals.

Chapter 2

Human brain activity data

2.1 Introduction

The human brain activities can be measured by a diversity of neuroimaging modalities, comprising non-invasive methods such as functional magnetic resonance imaging (fMRI), electroencephalogram (EEG), magnetoencephalography, and invasive approaches, such as electrocorticography (ECoG). Each of these methodologies offers unique advantages and limitations in relation to spatial and temporal resolution. The emphasis of my research investigation is placed on the fMRI modality.

Functional magnetic resonance imaging (fMRI) is a specialized variant of MRI designed to measure brain activities. Unlike conventional MRI, which yields a static illustration of brain anatomy, fMRI generates dynamic representations, elucidating the interaction and functionality of diverse brain regions over a given time span. This capacity to measure brain activity renders fMRI a potent instrument for cognitive studies. fMRI facilitates the capture of brain activities during task performance, such as watching a movie, or during periods of rest. This feature presents an opportunity to evaluate how the brain reacts to various stimuli.

Shen et al. (2019a) conducted an experiment wherein participants were exposed to thousands of natural images while their brain activities were monitored using fMRI. This procedure enabled the researchers to collect a wide range of brain responses to a variety of natural scenes. Employing machine learning and artificial intelligence algorithms, the researchers were successful in reconstructing visual images from the brain activities in the visual cortex. This research design served as the basis for my thesis.

This chapter aims to provide a fundamental understanding of fMRI, an examination of preceding studies closely related to the experimental design in this thesis, an in-depth explanation of the experimental procedure, and the processes used in the pre-processing and preparation of the fMRI data for subsequent analyses.

2.2 Basics of fMRI

Magnetic Resonance Imaging (MRI) operates based on the principles of nuclear magnetic resonance. At its core, the technology leverages the inherent property of atomic nuclei to absorb and emit radio frequency energy when placed in an external magnetic field. In the context of an MRI scanner, the magnetic field causes the hydrogen atoms in the body's water and fat molecules to align in one direction. Upon application of a radio frequency pulse, these atoms are excited and tipped out of alignment. As they return to their equilibrium state, they emit signals that can be picked up by a receiver. The timing and intensity of these signals depend on the type of tissue and its environment, providing a nuanced map of the body's internal structures. This non-invasive, radiation-free technique thus allows for detailed visualization of the body's anatomy and tissues, making it an indispensable tool in modern diagnostics and research.

fMRI is an extension of the traditional MRI technology, providing a dynamic map of brain activity rather than a static image of anatomy. The evolution of fMRI traces its roots to groundbreaking work by pioneering scientists in the early 1990s. During the early 1990s, John Belliveau led groundbreaking research which demonstrated that MRI could be used to detect regional changes in cerebral blood flow in response to visual stimulation (Belliveau et al., 1991). Around the same time, another significant discovery was made by Seiji Ogawa. Ogawa's research revealed that oxygenated and deoxygenated blood exhibit distinct magnetic properties, a finding that laid the foundation for Blood Oxygen Level Dependent (BOLD) contrast (Ogawa et al., 1990).

BOLD contrast is a key principle underlying fMRI. This technique is predicated on the observation that oxygenated and deoxygenated hemoglobin - the molecule responsible for carrying oxygen in blood - have differing magnetic properties. When a particular region of the brain becomes active, the demand for oxygen in that area increases, leading to a rise in blood flow. As more oxygenated blood arrives, the balance between oxygenated and deoxygenated hemoglobin shifts, resulting in a change in the local magnetic field. It is this change that fMRI measures, thereby allowing it to provide real-time maps of brain activity.

The discovery of BOLD contrast has revolutionized neuroscience, making it possible to non-invasively explore the workings of the brain in unprecedented detail.

The 3T MRI scanner, widely utilized in both clinical and research settings, offers superior image quality and shorter scan times compared to its predecessors. It provides highly detailed images, which are particularly beneficial for examining minute structures of the brain. The advent of the 7T MRI scanner has elevated this capability further. Its high-field strength dramatically improves signal-to-noise ratio and image resolution, enabling detailed visualization of minute brain structures and networks, and subtle pathological changes that might be overlooked at lower field strengths. It also improves the sensitivity and specificity of functional MRI studies by enhancing the blood oxygen level-dependent (BOLD) contrast. However, the widespread use of 7T scanners is currently limited due to factors such as cost, availability, and the specific technical expertise required for operation and data interpretation.

Despite its numerous benefits, fMRI has faced criticism. Key among the concerns is the fact that fMRI measures hemodynamic responses (blood flow changes) as a proxy for neural activity, which is an indirect method that does not capture the precise cellular and molecular dynamics of neuronal communication. Furthermore, the BOLD contrast imaging has inherent limitations, including susceptibility to physiological noise and uncertainty in the exact spatial localization of the source of the signal. Issues of statistical methodology in fMRI studies have also been raised (Monti, 2011), with some studies reported to have inadequate statistical power or improper multiple comparison correction, leading to potential false positives. Lastly, there is the challenge of interpreting the meaning of the observed activations, which is often confounded by complex brain processes and networks, leading to oversimplified or overgeneralized interpretations of fMRI results.

2.3 Review of related fMRI studies on visual decoding

In the present thesis, the main concentration lies on the analysis of fMRI brain activity during participation in a visual task viewing sequences of images. The experimental design is grounded heavily on preceding fMRI research, particularly those by Shen et al. (2019a) and Horikawa and Kamitani (2022), whose datasets partially constitute the data used in this thesis.

Shen et al. (2019a) collected data from three subjects engaged in viewing sequences of natural images sourced from the ImageNet database (Deng et al., 2009). These three subjects correspond to Subjects 1-3 in the present study. The participants were exposed to three

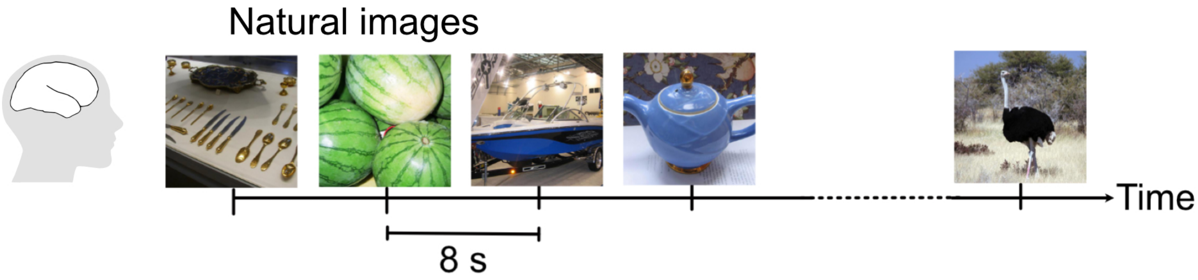

Figure 2.1: Schematic of image presentation experiment. Each image display lasted for 8 seconds in a stimulus block.

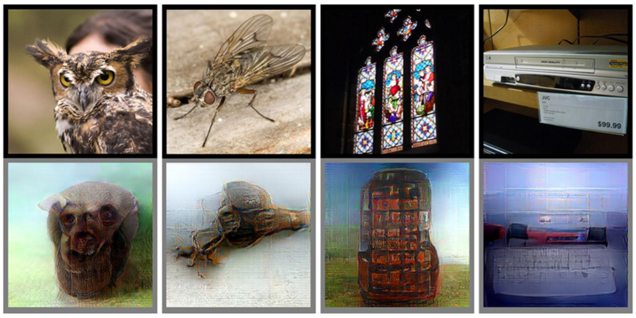

Figure 2.2: Visual images reconstructed from the fMRI brain activities. The top row is the presented images and the bottom row is the reconstructed images. Reproduced from *PLoS Computational Biology*, 2019;15: 1006633, Shen et al., Deep image reconstruction from human brain activity, Page No. 4, licensed under CC BY 4.0 / Cropped from original.

sessions of image presentation experiments encompassing the train natural-image session, the test natural-image session, and the test artificial-shape session. Each visual image featured a central fixation point and was flashed at a rate of 1 Hz. Each image display lasted for 8 seconds in a stimulus block with four volume scans (Repetition time [TR] = 2 s, Figure 2.1). Participants were directed to maintain fixation on the central point and click a button when two sequential blocks presented identical images. By employing a machine-learning model trained on the fMRI data from the train natural-image session, visual images in the test natural-image session and test artificial-shape session were reconstructed with discernible features (Figure 2.2).

Horikawa and Kamitani (2022) addressed the challenge of reconstructing visual images under the influence of attentional modulation. This investigation adopted the experimental design of Shen et al. (2019a) and collected fMRI data from seven subjects. Subjects 1-5 in their research were exactly Subjects 1-5 in the present thesis. They gathered test data from all seven subjects. Participants were asked to focus on one of the two superimposed images. The training data used for the reconstruction model mirrored the fMRI data in the

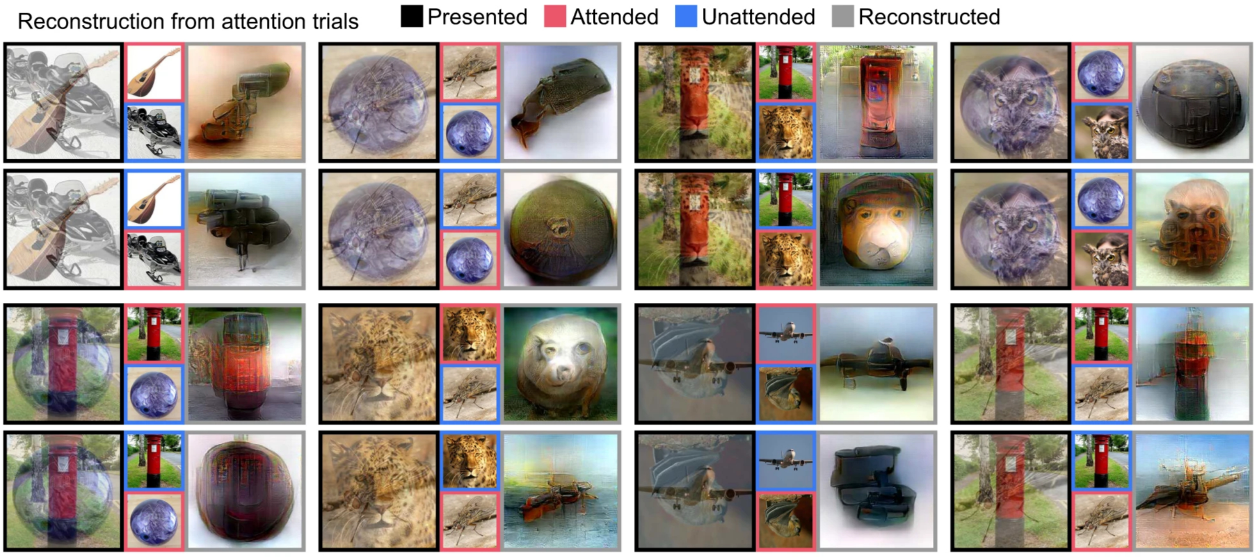

Figure 2.3: Reconstruction with participant's attention directed towards one of two superimposed images. Reproduced from *Communications Biology*, 2022;5: 34, Horikawa and Kamitani, Attention modulates neural representation to render reconstructions according to subjective appearance, Page No. 4, licensed under CC BY 4.0 / Cropped from original.

train natural-image session in Shen et al. (2019a). Notably, only the images the subjects attended to were reconstructed, a finding that highlighted the potent influence of attention in human vision (Figure 2.3).

The aforementioned studies illustrate the viability of converting the complexity of visual information encoded in fMRI brain activity into visual images. Notably, in contrast to some conventional neuroscience analytical tools, Horikawa and Kamitani (2022) employed the visual image reconstruction technique as a means to explicitly probe attentional modulation in human vision. In a similar vein, I used the fMRI data collected in the image presentation experiment and adopted the visual image reconstruction technique as a tool to assess whether fine-grained hierarchical visual information is shared across individuals.

2.4 Experiments

This section describes the details of the image presentation experiment and the fMRI data used in this thesis. The content of this section is based on the section 4.1: *fMRI datasets* and the section 4.2: *Regions of interest (ROIs)* of Ho et al. (2023).

2.4.1 Subjects

In this study, Subject 1–3 correspond to the three subjects in Shen et al. (2019a) and the dataset was reused. Subject 4 (male, age 22) and Subject 5 (male, age 27) participated in the additional experiments for the test natural-image and artificial-image sessions. The dataset of the training natural-image session of Subject 4 and 5 was reused from Horikawa and Kamitani (2022). All subjects provided written informed consent for participation in the experiments, in accordance with the Declaration of Helsinki, and the study protocol was approved by the Ethics Committee of Advanced Telecommunications Research Institute International (ATR).

2.4.2 Visual stimuli

The natural image stimuli are identical to those used in Horikawa and Kamitani (2017). The images were selected from 200 representative categories in the ImageNet dataset (2011, fall release; Deng et al., 2009). The natural training images were 1,200 images taken from 150 object categories, and the natural test images were 50 images taken from the remaining 50 object categories. The artificial image stimuli used in Shen et al. (2019a) consisted of 40 combinations of five shapes (square, small frame, large frame, plus sign, and cross sign) and eight colors (red, green, blue, cyan, magenta, yellow, white, and black).

2.4.3 Experimental design

Following Horikawa and Kamitani (2017), and Shen et al. (2019a), fMRI signals were measured while subjects viewed a sequence of visual images. The visual images had a central fixation spot and were flashed at a frequency of 1 Hz. Each presentation of an image lasted for 8 s in a stimulus block with four volume scans (Repetition time [TR] = 2 s). The subjects were instructed to maintain fixation on the central fixation spot and click a button when two sequential blocks presented the same image.

The test natural-image session and test artificial-shape session consisted of 24 and 20 runs, respectively. Each run consisted of 55 and 44 stimulus blocks comprising 50 and 40 blocks of different images, and 5 and 4 randomly interspersed repetition blocks, along with additional 32-s and 6-s rest periods at the beginning and the end. The 50 natural images and 40 artificial images were presented in random order in each run.

2.4.4 fMRI acquisition

Functional MRI data were obtained at the Kokoro Research Center of Kyoto University using a 3.0-Tesla Siemens MAGNETOM Verio scanner. An interleaved T2*-weighted gradient-echo echo-planar imaging (EPI) scan was performed to acquire functional images of the entire brain. The imaging parameters were as follows: TR=2000 ms, TE=43 ms, flip angle=80 deg, FOV=192 × 192 mm, voxel size=2 × 2 × 2 mm, slice gap=0 mm, number of slices=76, and multiband factor=4. Additionally, T1-weighted magnetization-prepared rapid acquisition gradient-echo fine-structural images of the entire head were also obtained using the following parameters: TR=2250 ms, TE=3.06 ms, TI=900 ms, flip angle=9 deg, FOV=256 × 256 mm, and voxel size=1 × 1 × 1 mm.

2.4.5 fMRI data preprocessing

The following description is provided by fMRIPrep (https://fmriprep.org/en/1.2.1/citing.html). The results included in this thesis are based on the data preprocessed using fMRIPrep version 1.2.1 (Esteban et al., 2019) and a Nipype-based tool (Gorgolewski et al., 2011 & 2017). Each T1w (T1-weighted) volume was corrected for INU (intensity non-uniformity) using N4BiasFieldCorrection v2.2.0 (Tustison et al., 2010) and skull-stripped using antsBrainExtraction.sh v2.2.0 (using the OASIS template). Brain surfaces were reconstructed using recon-all from FreeSurfer v6.0.1 (Dale et al., 1999), and the brain mask estimated previously was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical gray-matter of Mindboggle (Klein et al., 2017). Spatial normalization to the ICBM 152 Nonlinear Asymmetrical template version 2009c (Fonov et al., 2009) was performed through nonlinear registration with the antsRegistration tool of ANTs v2.2.0 (Avants et al., 2008), using brain-extracted versions of both T1w volume and template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using fast (Zhang et al., 2001; FSL v5.0.9).

Functional data were slice time corrected using 3dTshift from AFNI v16.2.07 (Cox, 1996) and motion corrected using mcflirt (FSL v5.0.9; Jenkinson et al., 2002). This was followed by co-registration to the corresponding T1w using boundary-based registration (Greve & Fischl, 2009) with 9 degrees of freedom, using bbregister (FreeSurfer v6.0.1). Motion correcting transformations, BOLD-to-T1w transformation, and T1w-to-template (MNI) warp were concatenated and applied in a single step using antsApplyTransforms (ANTs v2.2.0) using Lanczos interpolation.

Physiological noise regressors were extracted by applying CompCor (Behzadi et al., 2007). Principal components were estimated for the two CompCor variants: temporal (tCompCor) and anatomical (aCompCor). A mask to exclude signals with cortical origin was obtained by eroding the brain mask, ensuring that it only contained subcortical structures. Six tCompCor components were then calculated including only the top 5% variable voxels within that subcortical mask. For aCompCor, six components were calculated within the intersection of the subcortical mask and the union of CSF and WM masks calculated in T1w space, after their projection to the native space of each functional run. Frame-wise displacement (Power et al., 2014) was calculated for each functional run using the implementation of Nipype.

Many internal operations of fMRIPrep use Nilearn (Abraham et al., 2014), principally within the BOLD-processing workflow. For more details of the pipeline see http://fmriprep.readthedocs.io/en/latest/workflows.html.

The coregistered data to the T1w space were then re-interpolated to 2 × 2 × 2 mm voxels. The data samples were first shifted by 4-s (two volumes) to compensate for the hemodynamic delay, followed by regression to remove nuisance parameters such as a constant baseline, linear trend, and six head motion parameters from each voxel amplitude for each run. The data samples were then despiked to reduce extreme values (beyond ±3 standard deviations for each run) and were averaged within each 8-s trial (four volumes).

2.4.6 Region of interest (ROI)

The primary ROIs in this investigation include V1, V2, V3, V4 and higher visual cortex (HVC). Regions V1, V2, V3, and V4 were delineated using the standard retinotopy experiment (Engel et al., 1994; Sereno et al., 1995) in each subject's naive brain space. The HVC was defined as a contiguous region covering the lateral occipital complex (LOC), fusiform face area (FFA), and parahippocampal place area (PPA), which were identified using conventional functional localizers (Kourtzi and Kanwisher, 2000; Kanwisher et al., 1997; Epstein and Kanwisher, 1998). The whole visual cortex (VC) was defined as the combined regions of V1, V2, V3, V4, and HVC.

2.4.7 Data availability

The data that support the findings in this thesis is available from OpenNeuro repository:

- https://openneuro.org/datasets/ds001506/versions/1.3.1 for the dataset of the training natural-image session, the test natural-image session, and the test artificia-image session for Subject 1–3
- https://openneuro.org/datasets/ds003430/versions/1.2.0 for the dataset of the training natural-image session for Subject 4 and 5
- https://openneuro.org/datasets/ds003993/versions/1.0.0 for the dataset of the test natural-image and the artificial-image sessions for Subject 4 and 5

2.5 Data for analysis

The fMRI data from five subjects as reported in previously published studies (Shen et al., 2019a; Horikawa and Kamitani, 2022) were subjected to analysis in the subsequent chapters. Additional data was gathered for two out of the five subjects, the details of which are outlined in section 2.4: Experiments. The content of this section is based on the section 2.1: *fMRI data* of Ho et al. (2023).

To recapitulate, the train natural image session encompassed the repeated presentation of 1,200 natural images, each occurring five times. As for the test data, the test natural image session involved the repeated presentation of 50 natural images, each repeated 24 times. Furthermore, in the test artificial image session, a collection of 40 artificial images (simple geometric shapes) was presented, with each image displayed 20 times. Artificial images were introduced to evaluate the extent to which models trained on natural images exhibit generalization capability towards dissimilar image types. The fMRI data were averaged within each 8-s stimulus block (four fMRI volumes shifted by 4 s to account for hemodynamic delays). Consequently, after preprocessing the data (see section 2.4.5: fMRI data preprocessing), the dataset comprised of 6,000 (5 × 1,200) training samples, 1,200 (24 × 50) test samples consisting of natural images, and an additional 800 (20 × 40) test samples comprising artificial images.

In the decoding and reconstruction analyses, test samples were further averaged across repetitions (blocks) for each image, unless stated otherwise. Although some of the training data and the test data were collected at different times, separated by more than several months or even a year, the trained model generalized well across the datasets, as demonstrated in Shen et al. (2019a).

Chapter 3

Hierarchical neural code conversion

3.1 Introduction

The brain activity patterns collected during the viewing of naturalistic visual stimuli encode rich information about the presented stimuli, as described in section 1.1.4. These brain activity patterns manifest additional systematic organization, the significance of which is critical for proficient cognitive functions. The fMRI data introduced in Chapter 2 thus affords an avenue for the comprehensive investigation of both the encoded informational content and the underlying organizations. The content of this chapter is based on the section 1: *Introduction* and the section 2.2: *Neural code conversion* of Ho et al. (2023).

Sensory information is generally thought to be processed through a hierarchical pathway that detects topographically organized simple local features in the early stages and then progressively complex global features in the later stages, leading to holistic perception. In the ventral visual pathway, a stimulus is initially processed in the striate cortex (V1) to extract simple features, such as edges (Hubel and Wiesel, 1962), and is then further processed in the extrastriate cortices (V2–V4) and higher visual cortex (HVC) to detect more complex visual features (Figure 3.1), such as shape and face attributes, eventually identifying objects and scenes (Mishkin & Ungerleider, 1982). Whereas general principles such as topography and hierarchy appear to govern the organization of the visual cortex (VC), individual brains differ in both macroscopic anatomy and the fine-grained organization of feature representations. These individual differences make it challenging to relate visual cortical activity and perceptual content by simple mapping rules common across individuals.

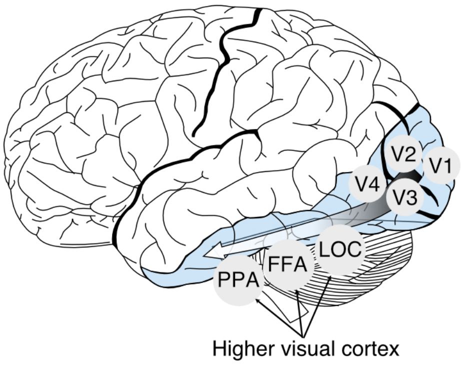

Figure 3.1: Visual cortical hierarchy along the ventral pathway. This diagram illustrates the simplified hierarchical sequence of visual processing in this study, beginning with the primary visual cortex (V1), continuing through the V2, V3, and V4 visual areas, and ending at the higher visual cortex (HVC) encompassing lateral occipital complex (LOC), fusiform face area (FFA), and parahippocampal place area (PPA).

Methods for the anatomical and functional alignment of different individuals' brains have been developed in decades of functional magnetic resonance imaging (fMRI) studies to account for individual differences. Human brain anatomy differs across individuals in terms of shape, size, and local anatomical landmarks. Functional brain area parcellation that clusters voxels/vertices with similar properties produces similar brain areas on the individual level but still exhibit distinct topological features (Blumensath et al., 2013; Laumann et al., 2015). The visual areas delineated by the retinotopy principle (Engel et al., 1994; Sereno et al., 1995) are often similar but not the same across individuals. Anatomical alignment can mitigate anatomical differences by matching anatomical features between brains (Fischl et al., 2008; Van Essen, 2004 & 2005), but it still cannot perfectly align the functional topography across individuals (Watson et al., 1993).

Functional alignment adopts an anatomy-free approach by learning statistical relationships between subjects' brain activity patterns (Haxby et al., 2011; Yamada et al., 2011 & 2015; Chen et al., 2015; Bilenko and Gallant, 2016; Guntupalli et al., 2016). The basic idea of functional alignment is that the subjects' brain activity patterns for a specific stimulus can be brought aligned such that the individual differences can be factored out (see section 1.2: Functional alignment). Methodologies of functional alignment include pairwise alignments between two subjects, such as a neural code converter (Yamada et al., 2011 & 2015), and template-based alignments, in which a shared template among subjects is constructed, such as hyperalignment (Haxby et al., 2011). Both the pairwise and template-based alignments necessitate a dataset of brain data, wherein subjects view either a sequence of preset natural images or a natural movie in order to capture a variety of brain responses to a wide range of natural scenes.

Functional alignment methods have revealed common neural representations across individuals that are concealed under substantial individual variations in brain responses (see section: 1.3: Visual features). However, previous investigations have often focused on a few specific features, such as object categories, image contrast, retinotopy, and semantics (Haxby et al., 2011; Yamada et al., 2011 & 2015; Bilenko and Gallant, 2016; Van Uden et al., 2018), leaving it unclear whether distinct levels of hierarchical fine-grained neural representations can be converted across individuals while preserving the encoded perceptual contents. Furthermore, previous studies have separately performed alignments on different brain areas using rough anatomical correspondences across individuals (Güçlü & van Gerven, 2017). It remains unknown whether data-driven methods trained on fMRI data can automatically detect hierarchical representations of distinct levels of visual features common across individuals.

In this chapter, the primary objective is to investigate the conversion of the brain activity patterns across individuals and the feasibility of detecting the hierarchical correspondence of distinct levels of visual features between individuals. For this purpose, I employed a functional alignment method known as neural code converter (Yamada et al., 2011 & 2015; see also section 1.2.2: Neural code converter) to convert brain activities. A neural code converter is trained with brain activity patterns in the whole visual cortex of a pair of source and target subjects viewing a sequence of images (Figure 3.2). In particular, no explicit information about visual hierarchy is imposed at the training stage. Careful evaluations using pattern and profile correlations were undertaken to examine the performance of the neural code conversions. Subsequently, an ablation study was conducted to investigate if the neural code converters successfully learned the hierarchical correspondence of visual subareas between individuals.

3.2 Methods

3.2.1 Anatomical alignment

Anatomical alignment was used as a benchmark for methods of functional alignment. The content of this section describes the alignment details and is based on the section 4.3: *Anatomical alignment* of Ho et al. (2023).

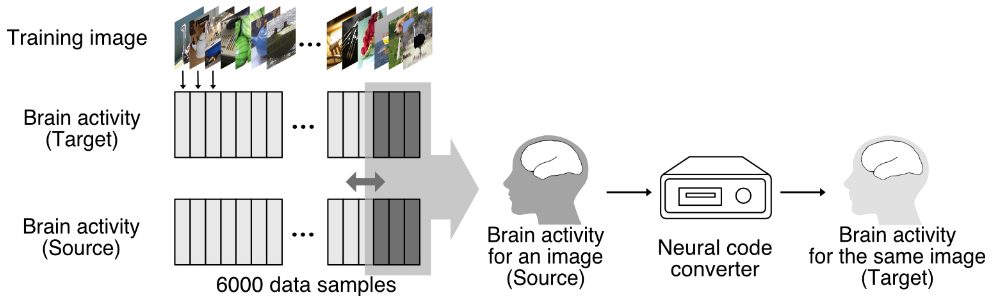

Figure 3.2: Training of a neural code converter. A converter model was trained on a subset of 6,000 samples of fMRI data responses to an identical stimulus sequence from both the source and target subject. Brain activity patterns in the whole visual cortex was used as input and no explicit information about cortical hierarchy is provided at the training stage.

The subjects' structural and functional images were nonlinearly normalized to a standard space: the ICBM 152 Nonlinear Asymmetrical template version 2009c (MNI152NLin2009cAsym [MNI space]; see section 2.4.5: fMRI data preprocessing). The T1w reference image was spatially normalized to MNI space by the ANTs (Avants et al., 2008) and the functional data were coregistered to this normalized T1w reference image. The coregistered data were then re-interpolated to 2 × 2 × 2 mm voxels. Furthermore, ANTs were used to normalize the ROI masks of V1, V2, V3, V4, and HVC in their native space to the brain in MNI space. In the inter-individual analysis, if a voxel of a source subject and a voxel of a target subject shared the same coordinates, the fMRI activity of the source voxel was considered to be that of the corresponding target voxel. Thus, the voxels of a source subject covered by a ROI mask of a target subject were selected as the input to the model.

3.2.2 Neural code converter

The neural code converter (Yamada et al., 2011 & 2015) described in the section 1.2.2 was adopted as the main method of functional alignment in this thesis. This section describes the detailed algorithm and is based on the section 4.4.1: *Neural code converter* of Ho et al. (2023).

The neural code converter model for each pair of subjects comprised a set of regularized linear regression models (ridge regression), each trained to predict the activities of an individual voxel of one subject (target) from the brain activity patterns of another subject (source) given the same stimuli. A converter takes a source subject's brain activity pattern $\mathbf{x}_i \in \mathbb{R}^m$ consisting of m voxels' values, and predicts the target brain activity pattern $\mathbf{y}_i = \mathbf{W}\mathbf{x}_i + \mathbf{b}$, where $\mathbf{y}_i \in \mathbb{R}^n$ is the converted brain activity pattern consisting of n voxels' values; $\mathbf{W} \in \mathbb{R}^{n\times m}$ is the conversion matrix and $\mathbf{b} \in \mathbb{R}^n$ is the bias vector. The converter is trained to minimize the objective function

$$\sum_{i}^{N} ||\mathbf{y}_i - (\mathbf{W}\mathbf{x}_i + \mathbf{b})||^2 + \lambda ||\mathbf{W}||^2, \tag{3.1}$$

where $\mathbf{y}_i$ is the measured target subject's brain activity pattern for the i-th sample, N is the number of training samples, λ is the regularization parameter, and $|| \cdot ||$ represents the Frobenius norm.

To optimize the performance of visual image reconstruction, I conducted fine-tuning of the regularization parameter via a 5-fold cross-validation approach on the training dataset. During each fold, the brain activity within the validation set underwent conversion to the target individual's brain space, after which it was subsequently decoded into DNN features. The decoded DNN features were used to calculate an identification accuracy that measured how well a decoded DNN feature pattern can identify the true stimulus between two alternatives (see section 4.2.5: Identification analysis). The regularization parameter was optimized in a grid-search manner to maximize the identification accuracy, which is linked to the performance of visual image reconstruction. For computational efficiency, a subset of 500 units was selected from each layer of the VGG19 model instead of using all units. The 500 units were randomly chosen due to the absence of prior knowledge regarding which specific DNN units would yield superior visual image reconstruction outcomes.

3.2.3 Conversion accuracy

The trained neural code converters were evaluated using two evaluation methods as described below. The content of this section is based on the section 2.2: *Neural code conversion* of Ho et al. (2023).

Given a trained neural code converter, two evaluation methods were performed: (a) pattern correlation, which calculates the spatial Pearson correlation coefficient between the converted and measured voxel patterns for a test image, and (b) profile correlation, which is the Pearson correlation coefficient between the sequences of converted and measured individual voxel responses to the 50 natural test images (Figure 3.3). The pattern correlation for an image was defined as the mean of 24 samples (converted) × 24 samples (measured) = 576 correlation coefficients. The profile correlation for each voxel was defined as the mean of 24 repetitions (converted) × 24 repetitions (measured) = 576 correlation coefficients. The obtained correlation coefficients were normalized by their noise ceilings to account for the noise in fMRI brain responses over repeated measurements with the same stimulus (Hsu et al., 2004; Lescroart and Gallant, 2019 ; see section 3.2.4: Noise ceiling estimation). To summarize the results, I further averaged the correlation coefficients across images and

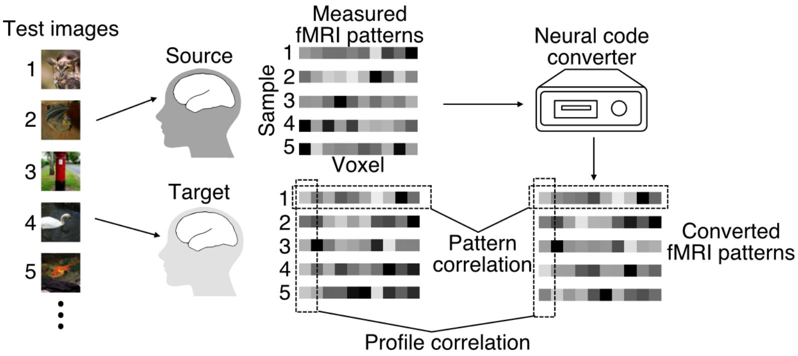

Figure 3.3: Evaluations of neural code converters. Two evaluations were performed by computing the Pearson correlation coefficients: pattern and profile correlations.

voxels for the pattern and the profile correlations, respectively, in each individual pair and each ROI.

3.2.4 Noise ceiling estimation

Repeated measures of the brain responses to an identical stimulus are subject to the measurement noise in fMRI data, inevitably degrading the prediction accuracy. To account for the noise, noise ceilings were estimated. This section describes the estimation procedure and is based on the section 4.5: *Noise ceiling estimation* of Ho et al. (2023).

I adopted the noise ceiling estimation used by Lescroart and Gallant (2019; see also Hsu et al., 2004). The noise ceiling was obtained by averaging the profile or pattern correlation coefficients between repetitions of the same stimuli within a subject. This noise ceiling estimation is based on the rationale that no model can predict better than the subject's own responses. Thus, the noise ceilings reflect the maximum performances of the converter models and were used to normalize the raw prediction accuracies of the converter models by dividing the raw accuracies by the noise ceilings.

Samples or voxels exhibiting noise ceilings falling below a threshold (defined as the 99th percentile point within the distribution derived from random pairs) were omitted from the assessment of conversion performance, as their measurement reliability was compromised. However, all voxels were encompassed within the subsequent DNN feature decoding analysis to ensure the prevention of any potential information leakage.

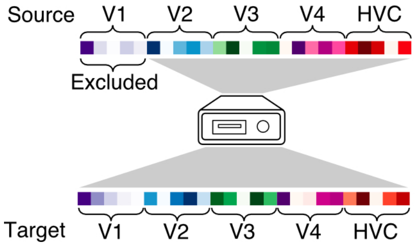

Figure 3.4: Ablation analysis on neural code converters. The analysis was performed by excluding one source visual area from the prediction of target voxel activities.

3.2.5 Ablation analysis on neural code converters

Ablation analysis on neural code converters were used to evaluate the degree of the influence of an ROI to a given target voxel. The content of this section is based on the section 2.2: *Neural code conversion* of Ho et al. (2023).

Given a trained neural code converter, excluding one of the source visual subareas (V1, V2, V3, V4, or HVC) from the input to the trained converter model can examine how the source visual areas influenced the conversion accuracy for each voxel in each target visual area (Figure 3.4). I evaluated the drop in performance (normalized profile correlation difference) relative to the performance when all source visual subareas were included (i.e., the whole VC).

3.2.6 Statistics

Statistical analyses were primarily conducted on data samples from each pair of subjects to assess the effect of individual conversions and their prevalence across pairs (Smith and Little, 2018; Ince et al., 2022). Furthermore, for summary purposes or instances where within-pair analysis was unfeasible, I performed group-level analyses using the mean values derived from 20 distinct pairs.

During the assessment of conversion within each subject pair, the mean conversion accuracy (pattern) and its corresponding 95% confidence interval were determined using pattern correlation coefficients for 50 visual stimuli. Similarly, the mean conversion accuracy (profile) and its 95% confidence interval were computed using profile correlation coefficients for all voxels. At the group level, the mean conversion accuracies (pattern/profile) from the dataset of 20 individual pairs were employed to calculate the group mean and its associated 95% confidence interval.

3.3 Results

This section presented the results of neural code conversion and the analysis of the hierarchical correspondence between individuals. The content of this section is based on the section 2.2: *Neural code conversion* of Ho et al. (2023).

3.3.1 Evaluation of neural code converters

A neural code converter model was established for each subject pair, with one subject as the target and the other as the source. This process yielded a total of 20 distinct individual conversions between five subjects. A converter model comprises a set of regularized linear regression models (ridge regression), each trained to predict the activity of each voxel of the target subject's brain from the source subject's brain activity pattern in a broad region of interest (ROI) that covered the lower to higher visual cortex termed VC (see section 1.2.2: Neural code converter). In this chapter, neural code converter models were trained using 2,400 training samples (two repetitions of 1,200 images) as a representative case.

VC consists of V1–V4 and ventral object-responsive areas (see section 2.4.6: Regions of interest). The higher visual cortex (HVC) is defined the continuous region covering the lateral occipital complex (LOC), fusiform face area (FFA), and parahippocampal place area (PPA). In the analyses of this section, all VC voxels were used as inputs to the converter without additional voxel selection. Conversion results were evaluated within individual ROIs (subareas) in the target subject's brain space.

Although the primary analyses focused on the samples within each conversion pair (Smith and Little, 2018; Ince et al., 2022), group results, where each data point represents an individual pair, are shown together for illustrative summary purposes. The normalized pattern correlation coefficients in individual pairs are shown for different ROIs of the target subject in Figure 3.5 (left), and their distributions across all conversion pairs are shown in Figure 3.6 (left). The mean normalized pattern correlation for the whole VC was 0.56 ± 0.06 (mean with 95% confidence interval [C.I.]) over 20 individual pairs, with the visual subareas showing comparable distributions. Examples of converted brain activity patterns are shown together with the targets' brain activity patterns in Figure 3.7. The mean normalized profile correlation for VC was 0.53 ± 0.05 over 20 individual pairs (Figure 3.5 right for individual pairs; Figure 3.6 right for group results). The subareas also yielded distributions similar to those of the VC. The conversion accuracy was modest in both pattern and profile correlations across all visual subareas but comparable to the findings in the previous study (Yamada et al., 2011 & 2015).

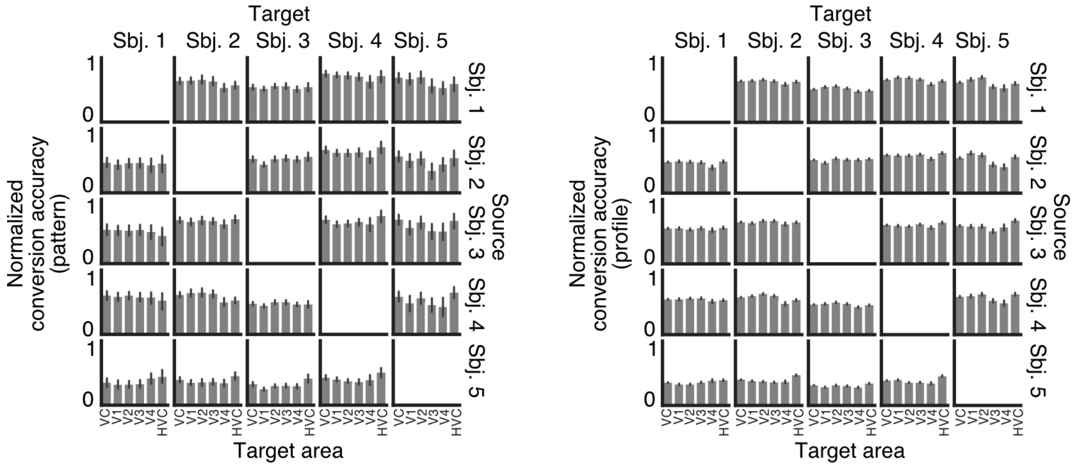

Figure 3.5: Conversion accuracies of individual pairs. The pattern correlation coefficients for 50 visual stimuli were used to calculate a mean conversion accuracy (pattern) and its 95% confidence interval. The profile correlation coefficients for voxels were used to calculate a mean conversion accuracy (profile) and its 95% confidence interval (error bars on left panel, 95% C.I. across visual images; error bars on the right panel, 95% C.I. across voxels).

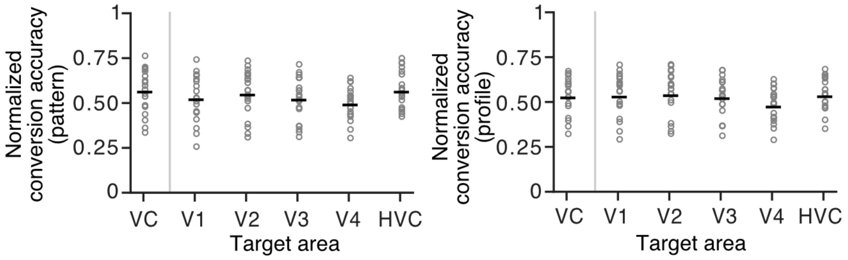

Figure 3.6: Conversion accuracies of 20 individual pairs. Distributions of the normalized pattern or profile correlation coefficients of 20 individual pairs are shown for the VC and visual subareas. Each horizontal black dash indicates the mean value; each circle represents the correlation coefficients of an individual pair.

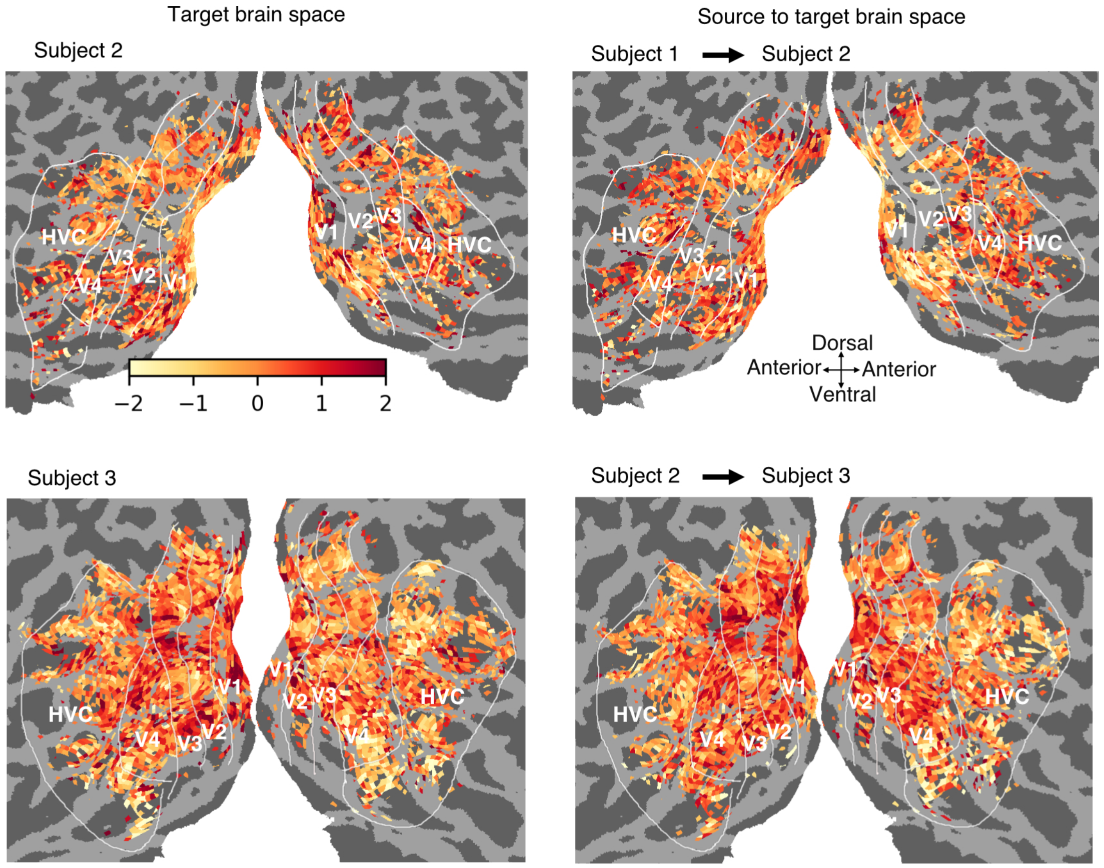

Figure 3.7: Single trial brain activity patterns responding to two test natural images. The left panels show the brain activity pattern for the golden fish in Subject 2 and the brain activity pattern for the butterfly in Subject 3 in their native brain space. The right panels show the brain activity patterns converted from the source subjects (Subject 1 and 2) to the target subjects. The activation values were normalized for ease of visualization.

3.3.2 Detection of the hierarchical correspondence

In order to assess the impact of source visual areas on the conversion accuracy for individual voxels within each respective target visual area, an ablation study was undertaken as described in section 3.2.5. The largest drop in performance of a target voxel was often caused by the exclusion of the corresponding source area (Figure 3.8). On average, the peak of performance drop shifted progressively from lower to higher excluded source areas along the hierarchy of the target areas (Figure 3.9 for results of some individual pairs; Figure 3.10 for group results). The results indicate that the machine learning-based neural code converter models automatically detect a "low-to-high" hierarchical correspondence between source and target visual areas even without explicit anatomical information.

3.4 Discussion

This chapter aimed to investigate the conversion of the brain activity patterns across individuals and the feasibility of detecting the hierarchical correspondence of distinct levels of visual features between individuals. The study started by showing that methods of pairwise functional alignment can accurately convert a source subject's brain activity into a target subject's brain space by evaluation using the pattern and profile correlations. The ablation analysis on the converters with the exclusion of voxels from various source visual subareas further showed that the converters automatically detected the hierarchical correspondences of visual subareas between individuals. The content of this section is based on the section 3: *Discussion* of Ho et al. (2023).

I have shown that the neural code converters automatically detected the hierarchical correspondence of visual subareas between two individuals without explicitly labeling the visual areas (Figures 3.8, 3.9, and 3.10). Previous studies of functional alignment have typically focused on a specific brain area, such as V1 or the inferior temporal cortex (Yamada et al., 2011 & 2015; Haxby et al., 2011). Other studies functionally aligned a large region of the cortex (Bilenko and Gallant, 2016; Van Uden et al., 2018), but their subsequent analyses addressed other research questions such as the retinotopic organization and the semantic information, leaving the hierarchical correspondences of visual subareas remained undiscussed. The results explicitly demonstrate that machine learning-based neural code converters can learn the hierarchical correspondence of visual subareas between two individuals. Furthermore, the observation that predictions can be made between different regions (for instance, source V1 predicting voxel values in target V2, as shown in Figures 3.9, and 3.10), suggests some level of shared information between neighboring areas. Nevertheless, it is noteworthy that

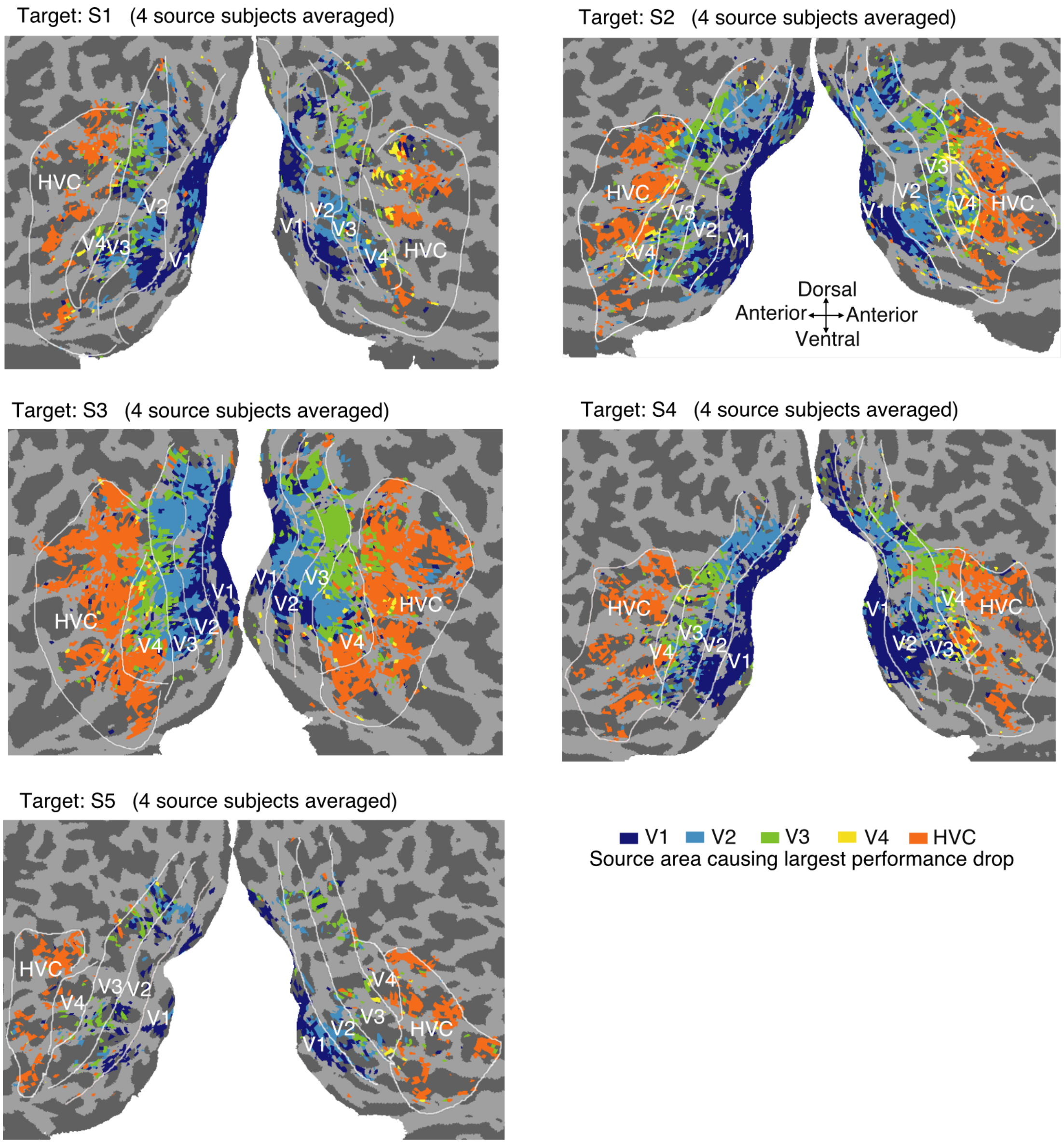
Target: S1 (4 source subjects averaged)
HVC
V2
V4
V3
V1
V2
V1
V3
V4
HVC
Target: S2 (4 source subjects averaged)
HVC
V3
V2
V1
V4
V3
V2
V1
V4
HVC
Dorsal
Anterior
Anterior
Ventral
Target: S3 (4 source subjects averaged)
HVC
V1
V2
V4
V3
V1
V3
V2
HVC
V4
Target: S4 (4 source subjects averaged)
HVC
V3
V2
V4
V1
V1
V4
HVC
V2
V3
Target: S5 (4 source subjects averaged)
HVC
V3
V2
V4
V1
V4
V2
V3
HVC
V1
V1 V2 V3 V4 HVC
Source area causing largest performance drop

Figure 3.8: Cortical map of the effects of source area exclusion. The cortical map is shown for five target subjects. Each voxel on the target brain is colored by the index of the excluded visual area that caused the largest performance drop when testing with the natural image test dataset (performance drops were averaged across four source subjects for a single target subject). Only voxels that generate reliable responses with noise ceilings above a threshold are shown (see section 3.2.4: Noise ceiling estimation).

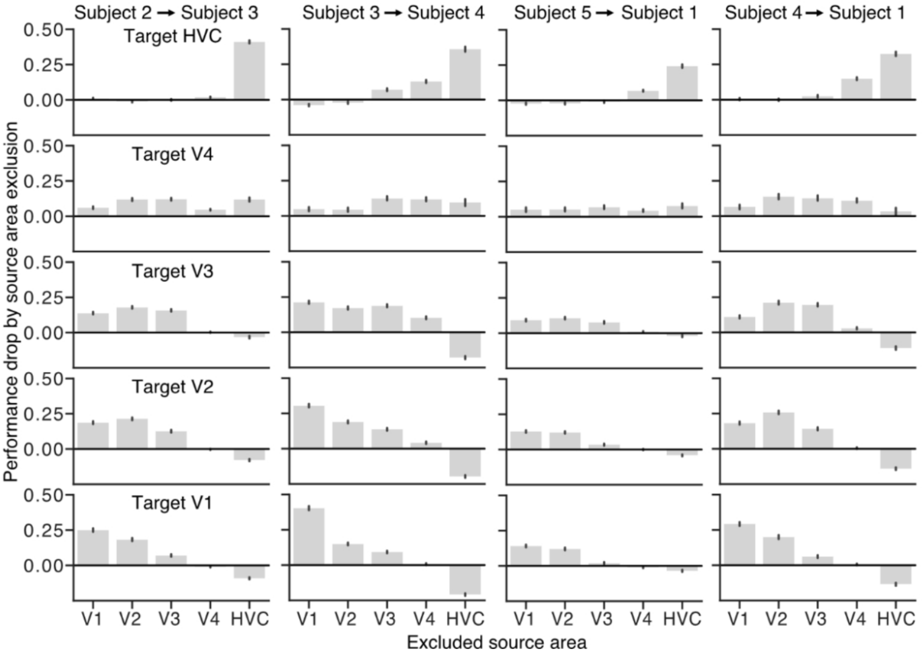

Figure 3.9: Performance drop caused by source area exclusion for four representative individual pairs. Each bar represents the mean performance drop averaged across voxels in a target area when a source area was excluded during prediction (error bar, 95% C.I. of performance drops across voxels).

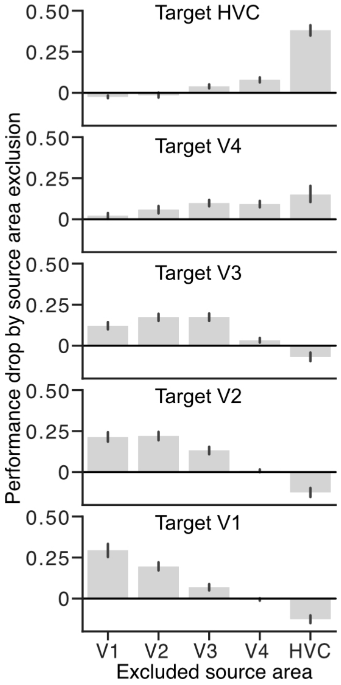

Figure 3.10: Mean performance drop caused by source area exclusion across 20 individual pairs. Each bar represents the mean performance drop averaged over 20 individual pairs (error bars, 95% C.I. from 20 individual pairs).

an increase in cortical distance between two areas correlates with a reduction in prediction accuracy, indicating that shared information resides within spatially proximate areas.

Chapter 4

Inter-individual DNN feature decoding and deep image reconstruction

4.1 Introduction

Chapter 3 demonstrated that the neural code converters can automatically detect the hierarchical correspondence of visual subareas between two individuals via purely data-driven approach. However, without a detailed analysis on the converted brain activity patterns, it remains unanswered whether the hierarchical and fine-grained visual features along the ventral pathway can be converted across individuals while preserving the encoded perceptual contents. The content of this chapter is based on the section 1: *Introduction*, the section 2.3: *DNN feature decoding*, and the section 2.4: *Visual image reconstruction* of Ho et al. (2023).

Recent progress in deep neural networks (DNNs) has facilitated comprehensive investigations into hierarchical feature representations spanning different visual cortical regions (Yamins et al., 2014; Güçlü and van Gerven, 2015 & 2017; Horikawa and Kamitani, 2017). Previous encoding and decoding studies have shown that DNNs pre-trained on natural images exhibit a correspondence between visual areas and DNN layers. These observations suggest a parallel progression, where the visual cortex, akin to DNNs, processes progressively intricate visual attributes along the ventral neural pathway. Additionally, the use of DNN-based reconstruction algorithms has led to successful reconstruction of perceptual content encoded in brain responses as images (Shen et al., 2019a & 2019b). The deep image reconstruction (Shen et al., 2019a) commences by predicting the DNN features corresponding to an image, leveraging the brain activity evoked by the image stimulus. Subsequently, an iterative

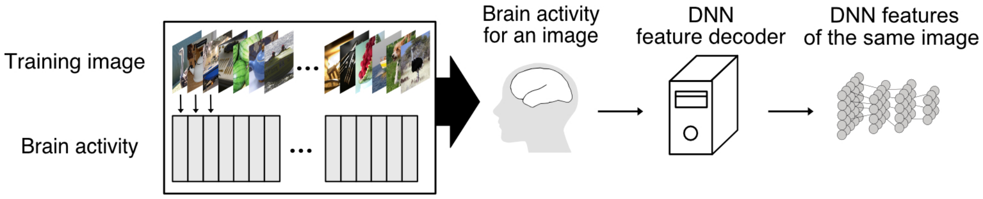

Figure 4.1: Training of the DNN feature decoders. The DNN feature decoding models were trained on the 6,000 samples of measured fMRI activities and the corresponding DNN features.

optimization procedure is employed, wherein an initial image is refined to align its DNN features with the predicted DNN features (Figure 1.16). The use of DNN feature decoding enables comprehensive evaluations of hierarchical visual representations, and visual image reconstruction affords a holistic evaluation of the precise encoding of perceptual content within brain activity patterns.

DNN feature decoding translates the brain activity patterns into the hierarchical DNN representation. This necessitates the training of DNN feature decoders, employing a dataset comprising brain data and DNN features. In this thesis, I made use of the fMRI data, collected during the image presentation experiment detailed in Chapter 2. This data was employed to train the DNN feature decoders with the objective of predicting the images' DNN features, extracted from the VGG19 DNN model (Simonyan and Zisserman, 2014; Figure 4.1). The VGG19 DNN model's 19 layers offer a holistic representation, enabling us to model the neural representation within the brain. Specifically, the intermediary layers provide a mechanism to represent the neural activity bridging the higher and lower visual areas, a task notoriously challenging to accomplish via mathematical models or modeling by semantics.

Brain activity encompass a wealth of information about the stimuli, which enables the reconstruction of the visual image through the application of reconstruction algorithms. Contemporary reconstruction algorithms frequently employ techniques derived from the realm of deep learning. Initial attempts at visual image reconstruction from brain activity patterns have been made through deep image reconstruction (Shen et al., 2019a; Figure 1.16), an optimization-based approach. Without the use of DNN feature decoding, works by Shen et al. (2019b) and Seeliger et al. (2018) employed generative adversarial networks (GANs; Goodfellow et al., 2014) to generate visual images from brain activity. Meanwhile, Han et al. (2019) leveraged a variational autoencoder for the purpose of reconstructing visual images. For a more detailed review, refer to Rakhimberdina et al. (2021). In the context of this thesis, I adopted the deep image reconstruction approach as it explicitly reconstructs images through multiple layers of hierarchical DNN features, a natural extension from the DNN feature decoding. The reconstructed images enable us to further analyze the extent of encoded stimulus information within brain activity patterns.

A model trained on one subject does not generalize to other subjects in general because of individual differences in macroscopic brain structure and fine-grained neural representations. Nevertheless, as Chapter 3 demonstrated, the neural code converter can convert brain activity patterns across individuals with moderate conversion accuracy. This provides an opportunity to analyze the converted brain activity patterns through the decoding of hierarchical DNN features and the visual image reconstruction.

In this chapter, the primary objective is to investigate the feasibility of converting the hierarchical and fine-grained visual features between individuals while preserving the encoded perceptual content. To achieve this, I utilized neural code converters to convert brain activity, and then used the decoding of hierarchical DNN features (Horikawa & Kamitani, 2017) and reconstruction of perceived images (deep image reconstruction; Shen et al., 2019a) to analyze the converted brain activity. I also trained DNN feature decoders with measured fMRI responses of the target subject as shown in Figure 4.1. Then, given the source subject's brain responses to novel stimuli, the converter transforms the brain activity into the target brain space (Figure 4.2). The DNN feature decoders, which have been pre-trained on the target subject, are employed to decode the converted brain activities. Subsequently, the decoded features are harnessed within a reconstruction algorithm to generate images. These reconstructed images are subsequently subjected to evaluation through identification analysis..

At the end of this chapter, other methods of pairwise alignment were adopted, including Procrustes transformation (Schönemann, 1966), optimal transport (Bazeille et al., 2019), and a template-based pairwise alignment via hyperalignment (Haxby et al., 2011), to replace the neural code converter in the inter-individual visual image reconstruction. The scope of the study was specifically restricted to pairwise alignment methods. Further discussions on shared templates were excluded due to the challenges associated with interpreting the correspondences of visual subareas between subjects within a shared template. In addition, the issue of optimal estimation of a template is distinct from the alignment methodologies (Bazeille et al., 2021). Instead, I exclusively employed template-based alignment for the creation of a pairwise transformation facilitated through the template, a methodology which I refer to as template-based pairwise alignment. The aim is not to exhaustively evaluate all available methods of pairwise alignment, but to show the robustness of the findings across several methods.

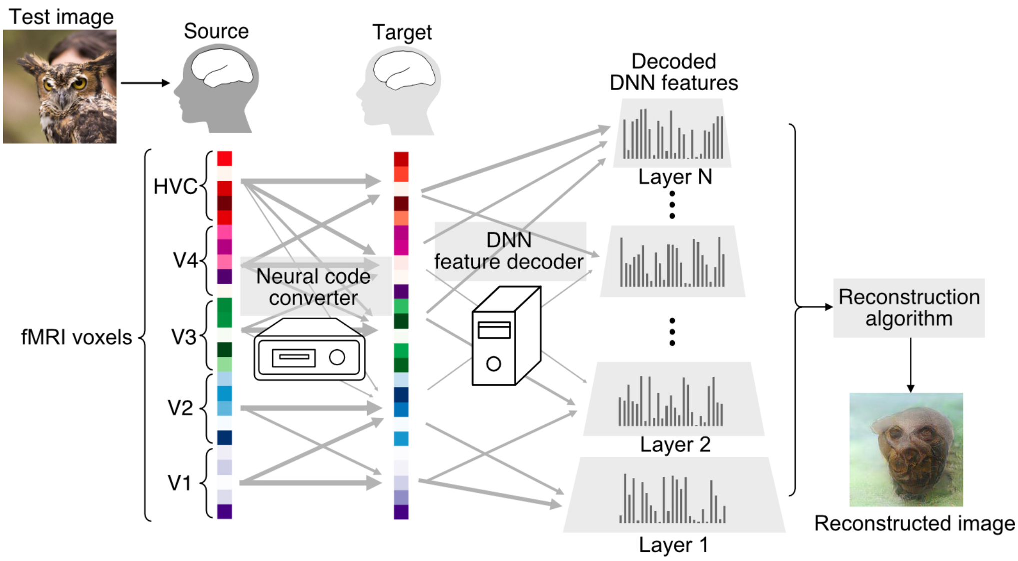

Figure 4.2: Inter-individual deep image reconstruction. The converter model converts the source subject's stimulus-induced fMRI pattern into the target subject's brain space. The converted fMRI pattern is then decoded (or translated) into a DNN feature pattern using the feature decoders. Finally, the decoded features are fed into the reconstruction algorithm to reconstruct the stimulus image perceived by the source subject.

4.2 Methods

4.2.1 Deep neural networks (DNN)

Two DNN models were used in this thesis for different purpose. The VGG19 DNN model was used for DNN feature decoding whereas AlexNet DNN model was used to extract DNN features for evaluation. This section describes the details of these two DNN models and is based on the section 4.6: *DNN model* of Ho et al. (2023).

The VGG19 DNN model (Simonyan & Zisserman, 2014) implemented using the Caffe library (Jia et al., 2014) was used for DNN feature decoding. This model is pre-trained for the 1,000-class object recognition task using the images from ImageNet (Deng et al., 2009; the pre-trained model is available from https://github.com/BVLC/caffe/wiki/Model-Zoo). The model consists of 16 convolutional layers and three fully connected layers. All the input images to the model were rescaled to 224 × 224 pixels. Following Shen et al. (2019a), outputs from individual units before rectification were used as target variables in the DNN feature decoding analysis. The number of units in each layer is as follows: conv1_1 and conv1_2, 3,211,264; conv2_1 and conv2_2, 1,605,632; conv3_1, conv3_2, conv3_3, and conv3_4, 802,816; conv4_1, conv4_2, conv4_3, and conv4_4, 401,408; conv5_1, conv5_2, conv5_3, and conv5_4, 100,352; fc6 and fc7, 4,096; and fc8, 1,000.

The AlexNet DNN model (Krizhevsky et al., 2012) implemented using the Caffe library was used to extract DNN features from the reconstructed images and the presented image. This model is also pre-trained similarly (available from https://github.com/BVLC/caffe/tree/master/models/bvlc_alexnet). The model consists of five convolutional layers and three fully connected layers. The number of units in each layer is as follows: conv1, 290,400; conv2, 186,624; conv3 and conv4, 64,896; conv5, 43,264; fc6 and fc7, 4,096; and fc8, 1,000.

4.2.2 DNN feature decoding analysis

For each individual DNN unit, I conducted training of a ridge linear regression model, herein referred to as the DNN feature decoder. This model was designed to ingest an fMRI activity pattern induced by a given stimulus as its input, subsequently generating a predictive feature value associated with the stimulus. The ridge regularization parameter was deliberately established at a value of 100, while both the feature values and voxel values underwent normalization before being employed in the training process. To enhance the training procedure, a voxel selection procedure was conducted, resulting in the identification of the uppermost 500 voxels characterized by the Pearson correlation coefficients between the sequences of feature values and voxel responses.

The performance of the trained decoders was evaluated through their application to the averaged fMRI pattern derived from repeated observations, a strategy that serves to amplify the signal-to-noise ratio of the fMRI signal. For details of the feature decoding, please refer to the works of Horikawa and Kamitani (2017 & 2022), Shen et al. (2019a), and Ho et al. (2023).

4.2.3 Brain hierarchy (BH) score

The original intent of the BH score was to quantify the extent of hierarchical resemblance between an artificial neural network and the human brain (Nonaka et al., 2021). This section briefly describes the calculation procedure and is based on the section 4.8: *Brain hierarchy (BH) score* of Ho et al. (2023).

The decoding-based BH score was used to investigate whether the hierarchical similarity is preserved after neural code conversion. The DNN features of randomly selected 1,000 units of each layer are decoded from the fMRI pattern of one of the five visual areas: V1–V4 and the HVC. For each unit, the visual area showing the best decoding accuracy was identified and was called the “top visual area.” The first layer, the last layer, and three randomly sampled intermediate layers were used to calculate a Spearman rank correlation coefficient between the hierarchical levels of the five DNN layers (coded as 0 through 4) and the top visual area (coded as V1: 0, V2: 1, V3: 2, V4: 3, and HVC: 4) across DNN units. This sampling procedure was repeated 10,000 times, and the mean Spearman rank correlation coefficient was taken as the BH score. See Nonaka et al. (2021) for more details.

4.2.4 Deep image reconstruction

An image reconstruction method (deep image reconstruction) proposed by Shen et al. (2019a) was adopted in this study. This section describes the detailed algorithms and is based on the section 4.9: *Visual image reconstruction* of Ho et al. (2023).

Deep image reconstruction optimizes pixel values of an input image based on a set of DNN features given as a target. Given the decoded DNN features from multiple layers, an image was reconstructed by solving the following optimization problem (Mahendran & Vedaldi, 2015):

$$\mathbf{v}^* = \underset{\mathbf{v}}{\operatorname{argmin}} \left(\frac{1}{2} \sum_{l}^{L} \beta_l || \phi_l(\mathbf{v}) - \mathbf{u}_{il} ||^2 \right), \tag{4.1}$$

where $\mathbf{v} \in \mathbb{R}^{224 \times 224 \times 3}$ is a vector whose elements are the pixel values of an image (width × height × RGB channels); L is the total number of layers; ϕ_l is the function that maps the image to the DNN feature vector of the l-th layer; u_{il} is the decoded DNN feature vector of the l-th layer for the i-th sample; and β_l is the parameter that weights the contribution of the l-th layer, which was set to be $1/||\mathbf{u}_{il}||^2$.

A natural image prior is applied by introducing a generative adversarial network called the deep generator network (DGN) to enhance the naturalness of the image (Nguyen et al., 2016). The optimization problem becomes

$$\mathbf{z}^* = \underset{\mathbf{z}}{\operatorname{argmin}} \left(\frac{1}{2} \sum_{l}^{L} \beta_l || \phi_l\big(\mathbf{G}(\mathbf{z})\big) - \mathbf{u}_{il} ||^2 \right), \tag{4.2}$$

where $\mathbf{G}$ is the DGN and $\mathbf{z}$ is a latent vector. The reconstructed image is obtained by $\mathbf{v}^* = \mathbf{G}(\mathbf{z}^*)$. The DGN is a pre-trained generator provided by Dosovitskiy and Brox (2016; available from https://github.com/dosovits/caffe-fr-chairs).

The solution to the above optimization problem is considered to be the reconstructed image from the brain activity pattern. Following Shen et al. (2019a), the reconstruction of natural images was executed utilizing the DGN framework. The optimization of the objective function was accomplished through the utilization of a stochastic gradient descent with momentum technique spanning 200 iterations. In contrast, for the reconstruction of artificial images, the DGN was not employed, and the optimization of the objective function was achieved via a limited-memory BFGS algorithm spanning 200 iterations (Le et al., 2011; Liu and Nocedal, 1989; Gatys et al., 2016).

4.2.5 Identification analysis

The process of identification analysis was employed as a methodology to assess and gauge the quality of image reconstruction. This section describes the analysis procedure and is based on the section 4.10: *Identification analysis* of Ho et al. (2023).

Presented images were identified using the similarity in either image pixels or DNN features, which were reshaped into an one-dimensional feature vector. The feature vector of a reconstructed image was used to compare the true feature vector of the presented image with the false alternative of another image. The comparison was counted as correctly identified if the feature vector of the reconstruction has a higher Pearson correlation coefficient with the true feature vector than with the false alternative. The identification was repeated for multiple false alternatives for each reconstruction. For the natural images, the identification was repeated with 49 alternatives for each reconstruction, resulting in 50 images × 49 comparisons = 2450 comparisons in total. The identification accuracy for a reconstructed image was defined as the proportion of correct identification.

During cross-validation to optimize the regularization parameters for the neural code converters, I used a set of decoded DNN features concatenated from multiple layers to calculate the identification accuracies and evaluate the performance. The candidate images for comparisons were a subset of the 1200 images presented in the training image session.

4.2.6 Other methods of functional alignment

This section describes other methods of function alignment, which are employed to serve as a robustness test for the finding in the chapter, and is based on the section 4.4: *Methods of function alignment* of Ho et al. (2023).

Procrustes transformation

Procrustes transformation is a transformation that includes rotation and preserves the shape of a geometric object (Schönemann, 1966). It was first applied to the functional alignment by Haxby et al. (2011). Considering the source and target subjects' brain activity patterns

$\mathbf{x}_i \in \mathbb{R}^m$ and $\mathbf{y}_i \in \mathbb{R}^n$, Procrustes transformation estimates an orthogonal transformation matrix $\mathbf{W} \in \mathbb{R}^{n \times m}$ to minimize

$$\sum_{i}^{N} ||\mathbf{y}_i - (\mathbf{W}\mathbf{x}_i)||^2, \tag{4.3}$$

with the constraint $\mathbf{W}^T\mathbf{W} = \mathbf{I}$, where N is the number of training samples. Please refer to Bazeille et al. (2021) for more details.

Optimal transport

Optimal transport is pertinent to the inquiry of efficiently transitioning a probability distribution into another probability distribution, while minimizing associated costs. It was first applied to the functional alignment in Bazeille et al. (2019). Defining $\mathbf{X} = (\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_m)$ and $\mathbf{Y} = (\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_n)$ with $\mathbf{a}_i, \mathbf{b}_j \in \mathbb{R}^N$ representing a sequence of a voxel response to N stimuli, optimal transport tries to find a transformation matrix $\mathbf{W}^*$ such that

$$\mathbf{W}^* = \underset{\mathbf{W}}{\operatorname{argmin}} \Big(\sum_{ij} \mathbf{W}_{ij} ||\mathbf{b}_i - \mathbf{a}_j||^2 - \varepsilon h(\mathbf{W}) \Big) \tag{4.4}$$

with the constraints $\sum_j \mathbf{W}_{ij} = 1/m$ and $\sum_i \mathbf{W}_{ij} = 1/n$.

The entropic term

$$h(\mathbf{W}) = -\sum_{ij} \mathbf{W}_{ij} \big(\log(\mathbf{W}_{ij}) - 1 \big) \tag{4.5}$$

regularizes the optimal transport problem and ε controls the strength of regularization. The regularization parameter was optimized as in the neural code converter. I used *fmrialign* (https://parietal-inria.github.io/fmralign-docs/index.html) package for the analysis. Please refer to Bazeille et al. (2019) for more mathematical details.

Template-based pairwise alignment via hyperalignment

The process of template-based pairwise alignment through hyperalignment initially involves the estimation of a shared template across subjects using hyperalignment (Haxby et al., 2011). Subsequently, a pairwise transformation is constructed by first mapping the brain activity of a source subject onto the established template, and then proceeding with an inverse mapping from the template to the brain space of the target subject, see Figure 4.3 for the difference between pairwise alignment and template-based pairwise alignment. In the first iteration, the hyperalignment algorithm first selects an initial target subject whose fMRI responses are used as a template, then aligns the second subject's fMRI responses to the template using Procrustes transformation. The template is then updated as the mean of the

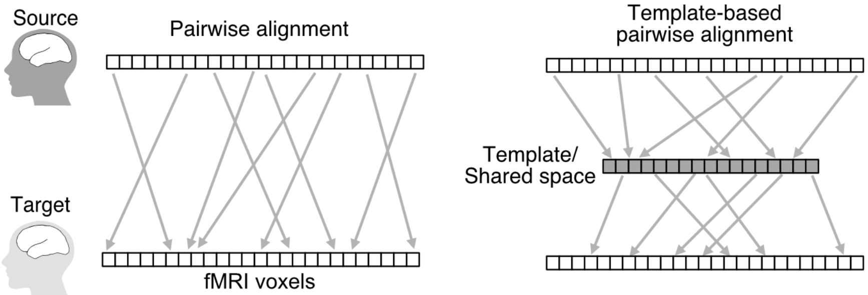

Figure 4.3: Illustration of pairwise alignment and template-based pairwise alignment. Ridge-based neural code converter, Procrustes transformation and optimal transport are pairwise alignment, and hyperalignment was used to construct a template to conduct template-based pairwise alignment. Analyses for all methods were performed with 2,400 training samples from a pair of source and target subjects. Pairwise alignment directly aligned the source subject's responses to the target subject brain space (left). Template-based pairwise alignment first mapped a source subject's responses into a template, followed by an inverse mapping into the target subject's brain space (right).

current template and the newly aligned fMRI responses. The same procedure is repeated for additional subjects. In the second iteration, each subject's original response is aligned to the mean aligned responses of other subjects. The mean aligned response is recalculated and treated as a template. In the last step, each subject's response is aligned to the template, and an orthogonal transformation matrix is obtained for each subject.

While the hyperalignment algorithm possesses the capability to estimate the shared space encompassing more than two subjects, I exclusively applied it between pairs of subjects, akin to the approach employed in the neural code converter analysis.

4.2.7 Statistics

Statistical analyses were primarily conducted on data samples from each pair of subjects to assess the effect of individual conversions and their prevalence across pairs (Smith and Little, 2018; Ince et al., 2022). Furthermore, for summary purposes or instances where within-pair analysis was unfeasible, I performed group-level analyses using the mean values derived from 20 distinct pairs. Within certain analytical contexts, the results with converted brain activity were compared with those without conversion (within-individual), in which the data from five subjects were processed in a similar way.

In the DNN feature decoding analysis, performed for each individual conversion or within each subject, the decoding accuracies (as quantified by profile correlations) of all units were employed to compute both the mean decoding accuracy and its accompanying 95%

confidence interval. At the group level, the mean decoding accuracies, emanating from either 20 individual pairs or five subjects, were harnessed to derive the group mean, accompanied by its corresponding 95% confidence interval.

When assessing the hierarchical structure, the computation of a Brain Hierarchy (BH) score was undertaken for each individual conversion or each subject (within-individual). The BH scores derived from 20 individual pairs or five subjects (within-individual) were further aggregated to obtain the group mean BH score.

During the identification analysis of reconstructed images for each specific conversion or individual subject (within), the identification accuracies associated with individual reconstructed images were harnessed to compute both the average identification accuracy and its corresponding 95% confidence interval. At the group level, the averaged identification accuracies obtained from 20 individual pairs or five subjects (within-individual) were employed to ascertain the group mean along with its accompanying 95% confidence interval.

4.3 Results

This section presented the results of DNN feature decoding and visual image reconstruction. The content of this section is based on the section 2.3: *DNN feature decoding* and the section 2.4: *Visual image reconstruction* of Ho et al. (2023).

4.3.1 DNN feature decoding

Given the trained neural code converters as described in Chapter 3, I used DNN feature decoding analysis to examine whether fine-grained representations of visual features were preserved in the converted fMRI activity patterns. The feature decoders had been trained to predict the DNN feature values of the stimuli using 6,000 training samples of a target subject's fMRI activity patterns in both the whole VC and individual visual subareas. The feature decoders were applied to the converted brain activities to predict the DNN features of the test images ("Across-functional" condition; see section 4.2.2: DNN feature decoding analysis). Following the original paper (Shen et al., 2019a), the average fMRI data over the repetitions for each test image was used as the input to feature decoders, unless stated otherwise. The decoding accuracy of each DNN unit was calculated as the Pearson correlation coefficient between the sequences of the decoded and true feature values for the test images. Additionally, I computed the mean decoding accuracy over all DNN units in each layer.

To provide a comparison, I performed the same analysis with anatomically aligned brain activity. The source subject's fMRI images were aligned to the target's anatomical template and then used for DNN feature decoding ("Across-anatomical"; see section 3.2.1: Anatomical alignment). The results were compared to those obtained from the standard within-individual decoding, where DNN features were predicted using the decoders trained on the same subject's data ("Within").

I initiated the evaluation by assessing the feature decoding performance derived from the converted fMRI activity encompassing the whole VC in the target space.The results of the neural code converter (Across-functional) showed lower but comparable performance with the within-individual results, with similar trends across layers for both in individual pairs and at the group level (Figure 4.4 for results of individual pairs; Figure 4.5 for group results). Among the three conditions, anatomical alignment (Across-anatomical) exhibited the least favorable performance, characterized by accuracies mostly falling below 0.1 across layers, both within individual pairs and at the group level. These outcomes underscore the advantage held by neural code converters over anatomical alignment in the context of DNN feature decoding.

Subsequently, I conducted decoding analyses on each DNN unit using voxels from individual visual areas (V1–V4 and HVC in the target space) and identified the visual area that yielded the highest decoding accuracy for each unit ("top visual area"), following Nonaka et al. (2021). The distribution of the top visual area across DNN units in a given layer was then computed. I observed a shift of the peak area, from lower to higher areas, along the DNN hierarchy in all conditions (Figure 4.6 for results of individual pairs; Figure 4.7 for group results). To quantify the degree of hierarchical correspondence between brain areas and DNN layers, I adopted the decoding-based brain hierarchy (BH) score, which is based on the rank correlation between the hierarchical levels of the DNN layer and the top brain area across DNN units (Figure 4.8; see section 4.2.3: Brain hierarchy (BH) score). The results of the within-individual condition replicated the previous findings with a BH score of around 0.5 (Horikawa and Kamitani, 2017; Nonaka et al., 2021). Despite the low accuracies in feature decoding with anatomical alignment (Across-anatomical; Figure 4.5), the hierarchical correspondence was largely preserved when quantified by the BH score (Figures 4.6 and 4.8). This is presumably because anatomical alignment maps a macroscopic organization of hierarchical visual areas between subjects, and the relative amount of information about hierarchy is preserved. The inter-individual conversion (Across-functional) showed a lower but substantial degree of hierarchical correspondence even though the converter was blind to cortical hierarchy information during training.

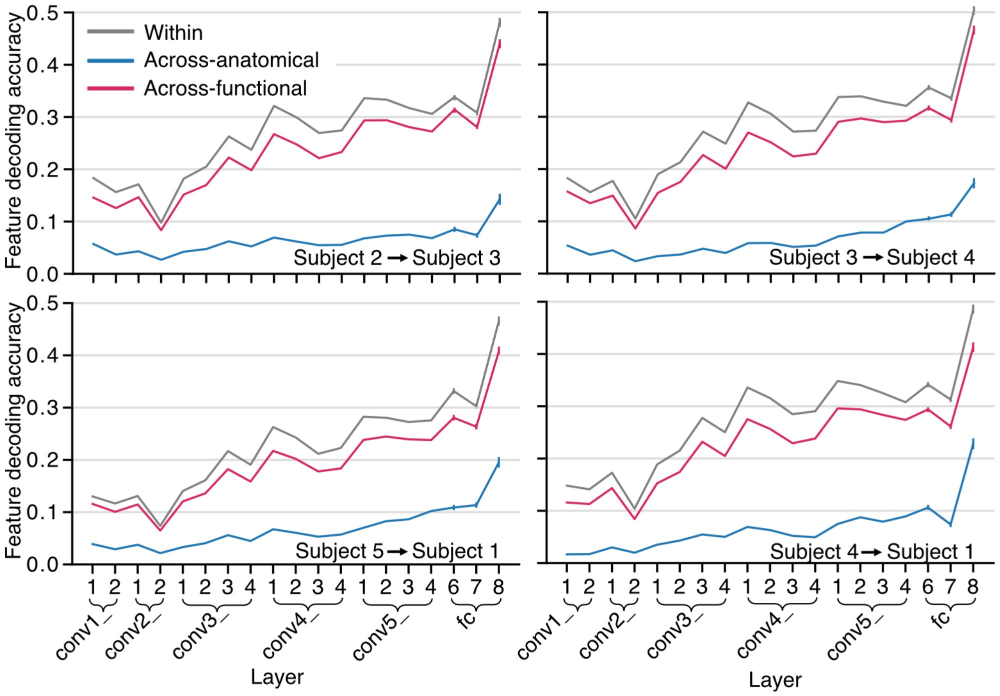

Figure 4.4: DNN feature decoding accuracy from the whole visual cortex (VC) for four representative individual pairs. Decoding accuracies for each layer of the VGG19 model are shown for the Within, Across-anatomical, and Across-functional conditions. The decoding accuracies for all DNN units in each layer were used to calculate a mean decoding accuracy and its 95% confidence interval (error bar, 95% C.I. across voxels).

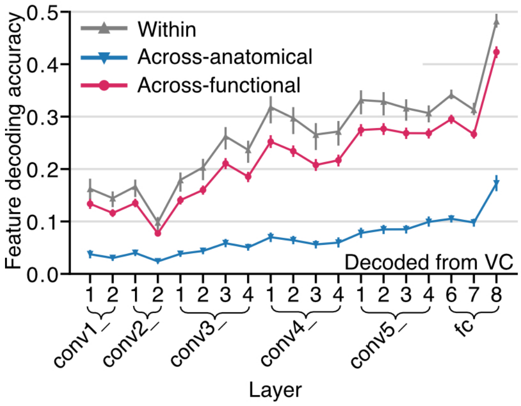

Figure 4.5: Mean DNN feature decoding accuracy from the whole visual cortex (VC) across 20 individual pairs. Decoding accuracies for each layer of the VGG19 model are shown for the Within, Across-anatomical, and Across-functional conditions (error bars, 95% C.I. from five subjects for the Within condition, and from 20 individual pairs for the Across-anatomical and Across-functional conditions).

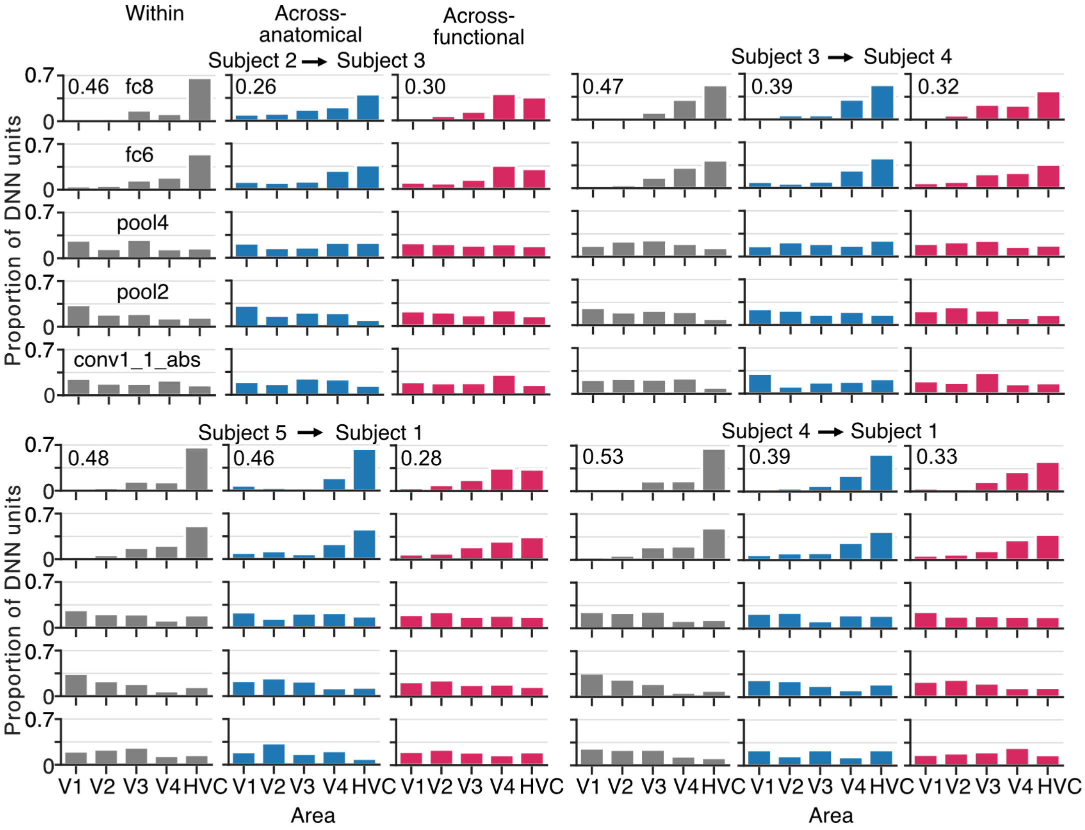

Figure 4.6: Proportion of the "top visual area" (best decodable area for each DNN unit) across DNN units in each layer for four representative individual pairs. Only five representative layers are shown. Each bar indicates the proportion of DNN units. The numbers on the top left indicate the BH scores.

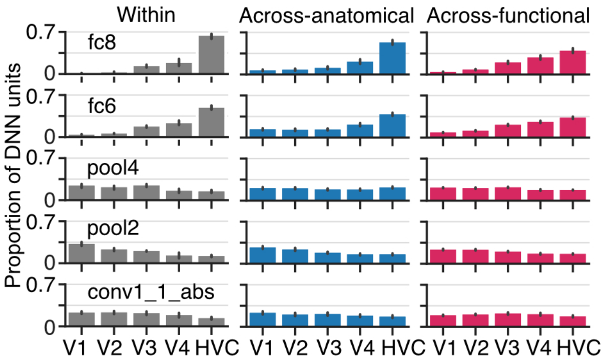

Figure 4.7: Mean proportion of the "top visual area" (best decodable area for each DNN unit) across DNN units in each layer over 20 individual pairs. Only five representative layers are shown. Each bar indicates the mean proportion of DNN units over five subjects for the Within condition or over 20 individual pairs for the Across-anatomical and Across-functional conditions (error bars, 95% C.I. from five subjects or 20 pairs.).

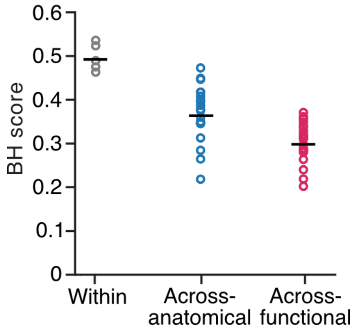

Figure 4.8: Brain hierarchy (BH) score. The horizontal black dashes indicate the mean BH score over subjects or pairs; each circle represents the BH score for a subject or a pair.

4.3.2 Visual image reconstruction

Having established the successful decoding of multiple level of DNN feature representations from converted brain activities, I next sought to investigate the feasibility of reconstructing visual images through DNN features decoded from converted brain activities using the technique of deep image reconstruction (Shen et al., 2019a; see section 4.2.4: Deep image reconstruction). This reconstruction analysis was performed not only on natural images but also extended to artificial images characterized by simple geometric shapes (see section 2.4.2: Visual stimuli).

Examples of reconstructions from the visual cortex (VC) for the Within, Across-anatomical, and Across-functional conditions are first demonstrated (Figure 4.9). The reconstructed images derived from the Within and Across-functional conditions captured the main characteristics of the presented images, including the shapes and colors of the objects, while reconstructions with anatomical alignment (Across-anatomical) showed neither a recognizable shape nor color of the objects in the presented images (see Figures 4.10 for several examples of natural images and artificial images for all individual pairs). The presented reconstructions thus far were derived from the average fMRI data over all the repetitions (24 and 20 repetitions for natural and artificial images, respectively). Despite not being the main focus of analysis, the results using the average of varying repetitions are available in Figure 4.11. Notably, it is worth highlighting that even a single repetition of fMRI sample was capable of yielding distinguishable reconstructions, wherein the visual quality exhibited enhancement with an increase in the number of repetitions.

For a quantitative assessment of the reconstruction outcomes, I performed a pairwise identification analysis in which the pixel or DNN feature pattern of a reconstruction was used to identify the true stimulus between two alternatives by choosing the one with a more correlated pattern (see section 4.2.5: Identification analysis). DNN feature patterns were

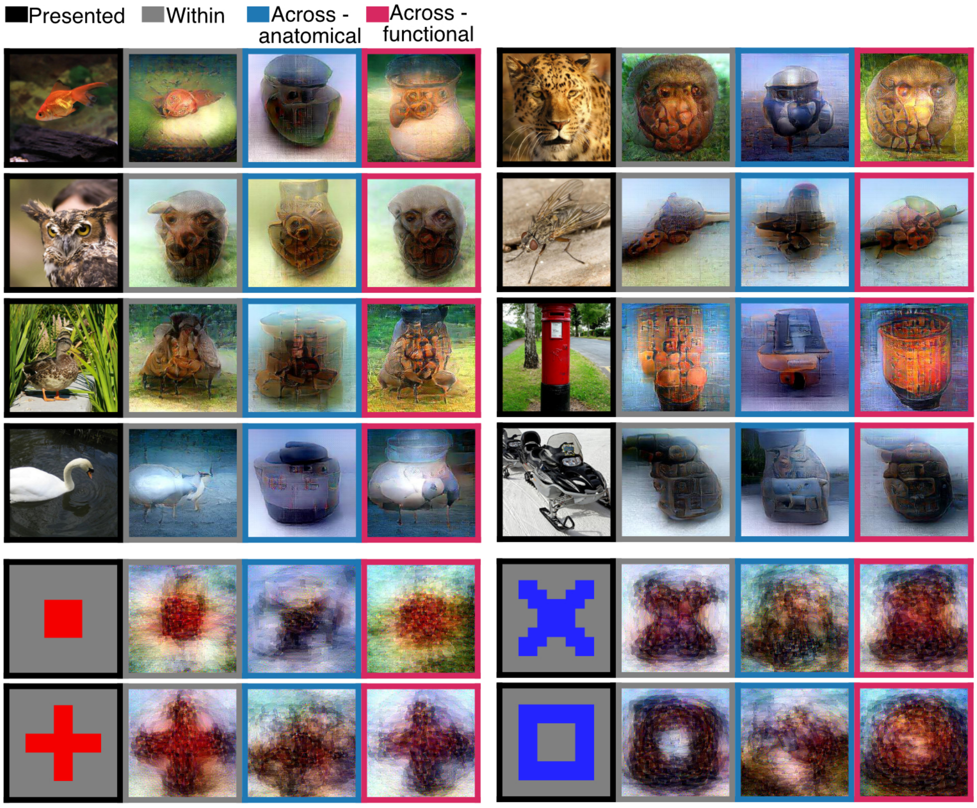

Figure 4.9: Within and across-individual reconstructions from the whole visual cortex (VC). The reconstructions shown under the three analytical conditions for each stimulus image were all from the same source subject. The results for different stimulus images are from different source subjects.

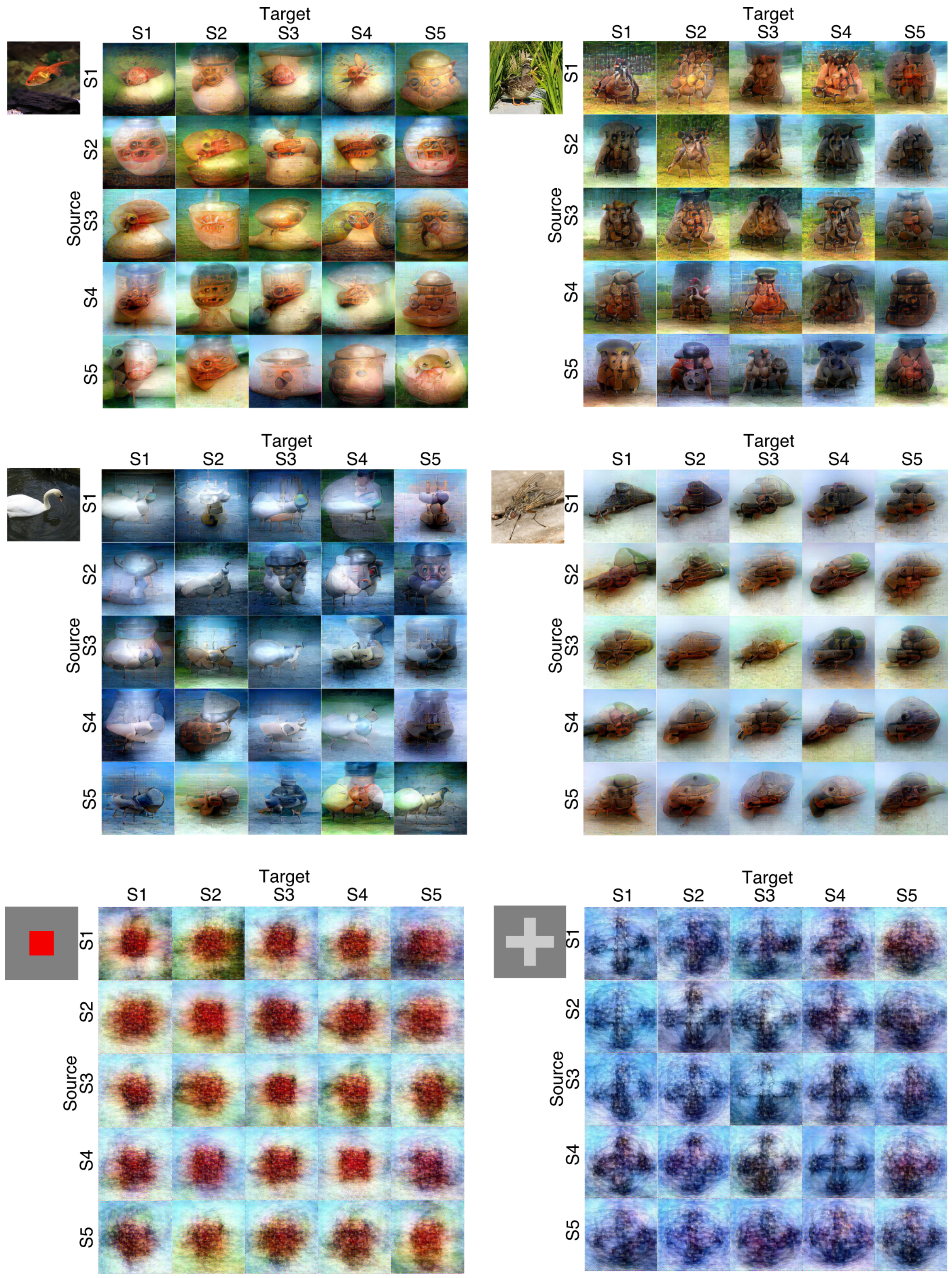

Figure 4.10: Reconstructed images across individual pairs. For each image, the diagonal images in each block are the reconstructed images in the Within condition; the off-diagonal images are the reconstructed images in the Across-functional condition with the converters trained on 2,400 training samples. All images were reconstructed from the whole visual cortex (VC).

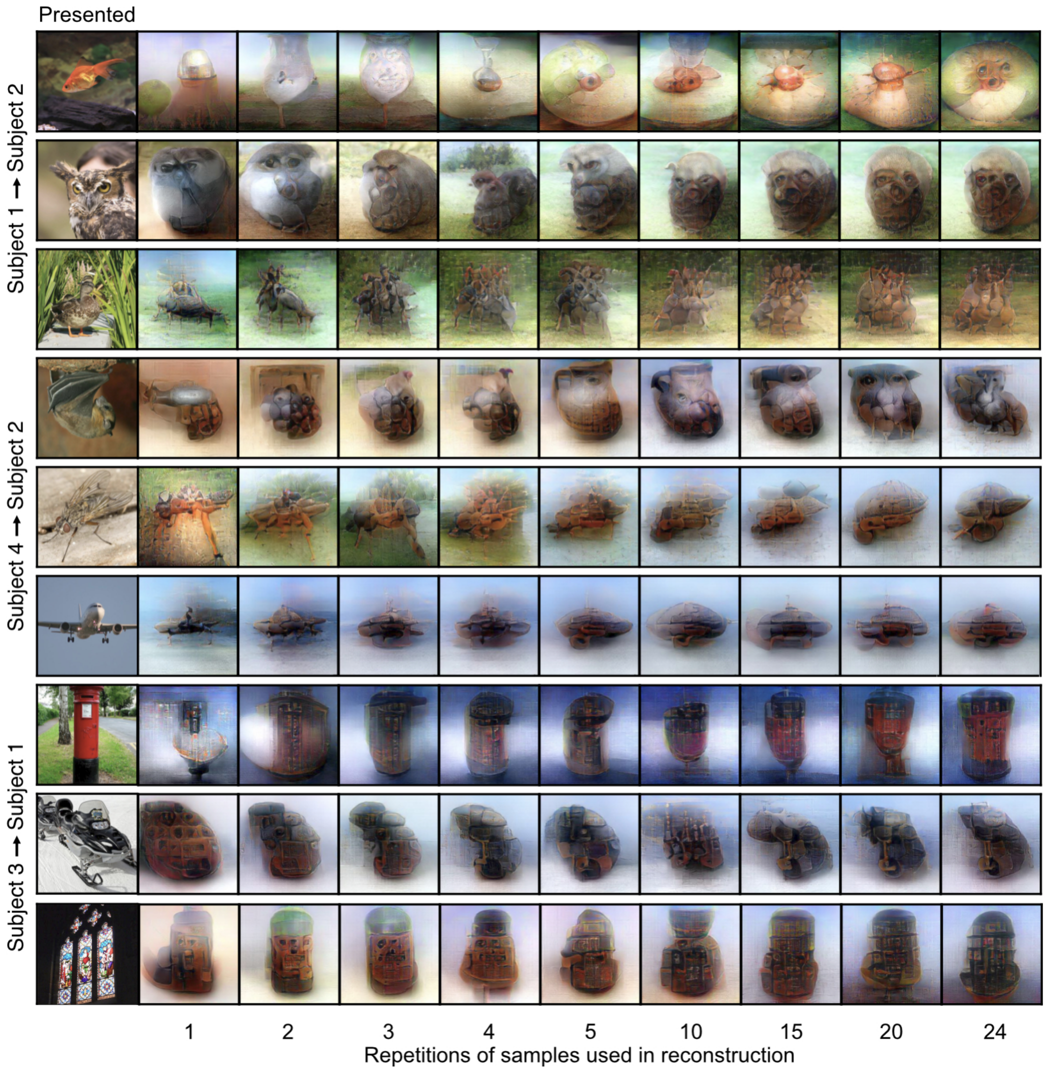

Figure 4.11: Reconstructed images using varying repetitions of samples. The converted fMRI samples corresponding to a visual image were averaged over repetitions and were reconstructed into an image. The reconstructed images were shown for three representative individual pairs. All images were reconstructed from the whole visual cortex (VC).

extracted using the AlexNet model (Krizhevsky et al., 2012), which is different from the DNN used in the reconstruction method (VGG19 model). The identification was repeated for multiple false alternatives to obtain the accuracy for each reconstruction. For group analysis, the mean identification accuracy was calculated over all reconstructions in each pair. While the within-individual condition (Within) showed overall superior performance both for natural and artificial images, neural code conversion (Across-functional) greatly outperformed anatomical alignment (Across-anatomical) both in individual pairs and at the group level (Figure 4.12 for individual pairs; Figure 4.13 for group results).

4.3.3 Robustness across functional alignment methods

Sections 3.3.2, 4.3.1, and 4.3.2 confirmed the preservation of hierarchical visual information during neural code conversion. To examine the robustness of this conclusion across different functional alignment methods, similar pairwise alignment analyses were conducted using Procrustes transformation, optimal transport, and template-based hyperalignment with 2,400 training samples. Subsequently, DNN feature decoding and visual image reconstruction were performed. The content of this section is based on the section 2.2: *Neural code conversion*, the section 2.3: *DNN feature decoding*, and the section 2.4: *Visual image reconstruction* of Ho et al. (2023).

These functional alignment techniques displayed comparable conversion accuracies. The mean normalized pattern correlations for the whole VC for Procrustes transformation, optimal transport, and template-based pairwise alignment were 0.55 ± 0.09, 0.74 ± 0.11, and 0.55 ± 0.09 (mean with 95% C.I.) across 20 individual pairs respectively. Similarly, the mean normalized profile correlations were 0.52 ± 0.07, 0.58 ± 0.08, and 0.52 ± 0.07 across 20 individual pairs respectively (Figure 4.14). Among them, optimal transport demonstrated superior conversion accuracy.

Other functional alignment methods demonstrated comparable DNN decoding accuracies. Among these, ridge-based neural code converters achieved the highest DNN decoding accuracies across the majority of layers, while optimal transport showed slightly diminished accuracy (Figure 4.15). Despite these differences in decoding, all methods led to similar visual reconstructions, albeit with optimal transport falling a bit short in comparison (Figure 4.16). This outcome substantiates the robustness of the finding, as all functional alignment methods retained the hierarchical visual information crucial for visual image reconstruction.

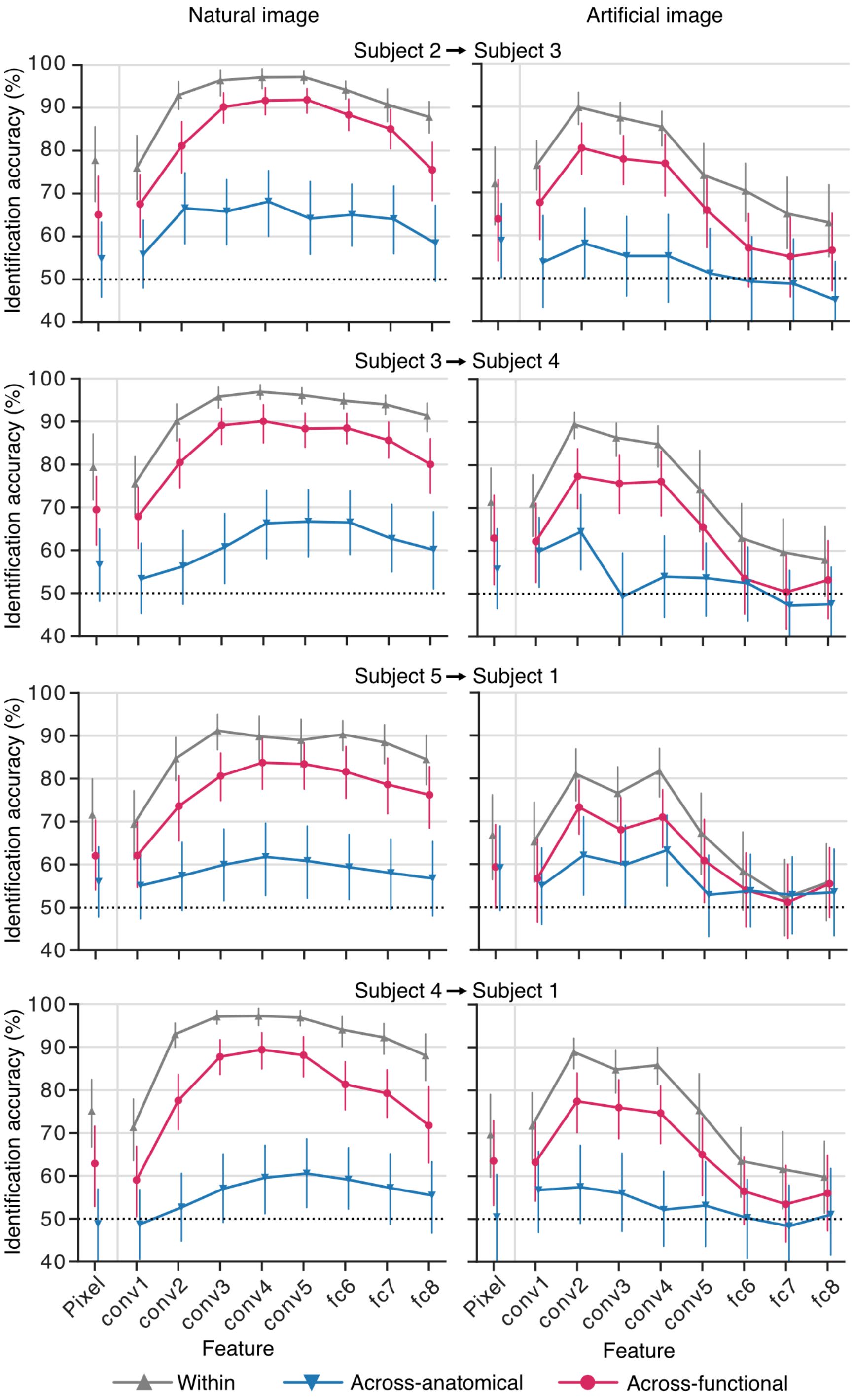

Figure 4.12: Identification accuracies of representative individual pairs. The identification accuracies for reconstructed images were used to calculate a mean identification accuracy and its 95% confidence interval for the Within, Across-anatomical, and Across-functional conditions (left, natural images; right, artificial images; error bar, 95% C.I. of identification accuracies across reconstructed images; dotted lines, chance level = 50%).

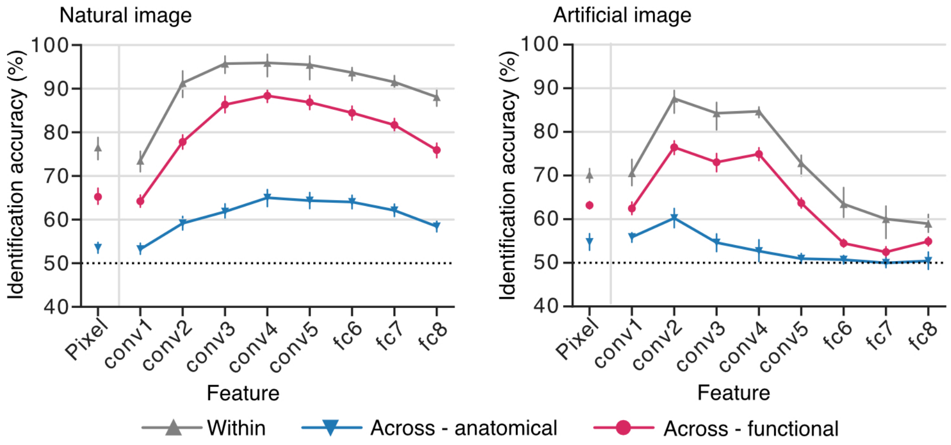

Figure 4.13: Mean identification accuracy across all individual pairs or all subjects. A mean identification accuracy was calculated over all reconstructed images for each subject or individual pair. DNN features of images were extracted from the eight layers of the AlexNet model (left, natural images; right, artificial images; error bars, 95% C.I. from five subjects or 20 pairs; dotted lines, chance level = 50%).

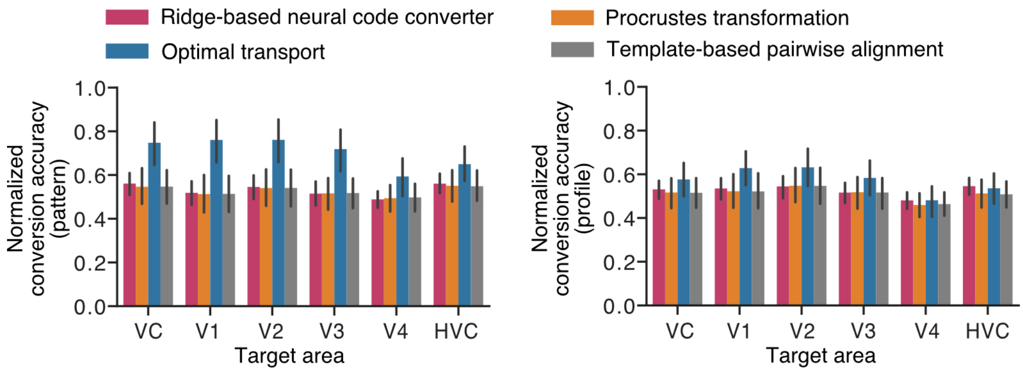

Figure 4.14: Evaluation of different methods of functional alignment on visual areas. The conversion accuracy was averaged across 20 individual pairs for Ridge-based neural code converter, Procrustes transformation, optimal transport, and template-based pairwise alignment via hyperalignment (error bars, 95% C.I. from 20 individual pairs).

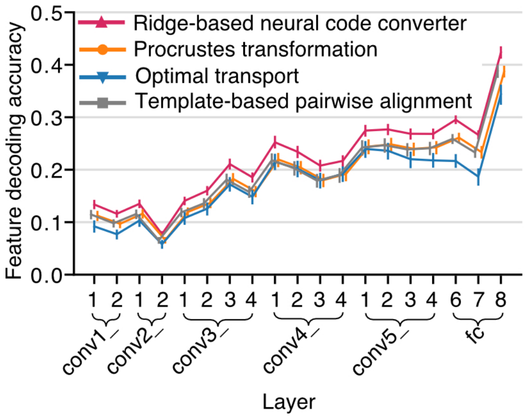

Figure 4.15: Inter-individual DNN feature decoding accuracies via different methods of functional alignment. The decoding accuracies were averaged across 20 individual pairs for Ridge-based neural code converter, Procrustes transformation, optimal transport, and template-based pairwise alignment via hyperalignment (error bars, 95% C.I. from 20 individual pairs).

4.4 Discussion

This chapter aimed to investigate whether and how hierarchical and fine-grained visual information could be converted while preserving perceptual content across individuals using methods of pairwise functional alignment. Decoding the converted brain activity into DNN features unveiled a clear correspondence between distinct visual subareas and layers within the DNN. The transformation of converted brain activity into visual images yielded reconstructions characterized by discernible shapes and colors of the objects depicted within the presented images. The analyses demonstrate that hierarchically organized fine-grained visual features that enable visual image reconstruction are preserved in the converted brain activity, allowing efficient reconstruction of visual images without training subject-specific models. The content of this section is based on the section 3: *Discussion* of Ho et al. (2023).

By decoding the converted fMRI activity patterns into DNN features and reconstructing them as visual images via the decoded DNN features (Figures 4.4, 4.5, and 4.9), I showed that hierarchically organized fine-grained visual features that enable visual image reconstruction are preserved in the neural code conversion. Previous studies have mainly focused on some specific features, such as object categories, image contrast, retinotopy, and semantics (Haxby et al., 2011; Yamada et al., 2011 & 2015; Bilenko and Gallant, 2016; Van Uden et al., 2018), but whether a set of hierarchical fine-grained features is preserved after functional alignment remained unknown. The results of DNN feature decoding on multiple levels of DNN layers

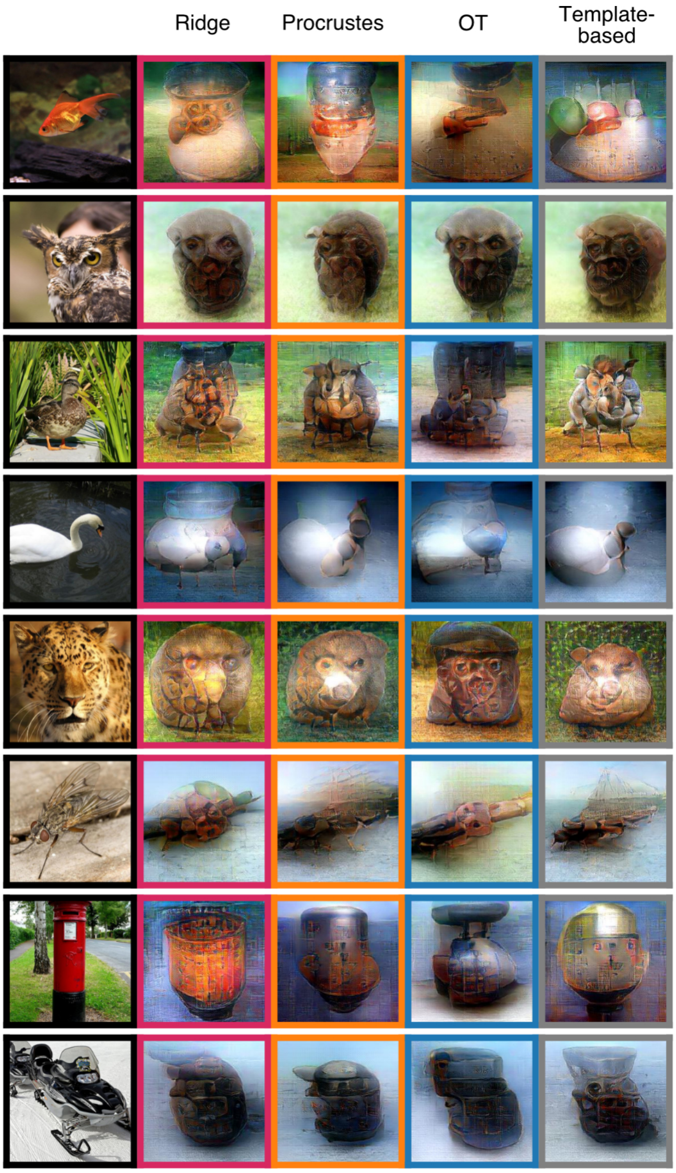

Figure 4.16: Reconstructed images obtained via different methods of functional alignment. The images were reconstructed with Ridge-based neural code converter, Procrustes transformation, optimal transport (OT), and template-based pairwise alignment via hyperalignment. All images were reconstructed from the whole visual cortex (VC). The reconstructions shown under the four analytical conditions for each stimulus image were all from the same source subject. The results for different stimulus images are from different source subjects.

showed that the converted fMRI activity patterns held multiple levels of fine-grained visual features (Figure 4.4 and 4.5). Moreover, successful visual image reconstruction further confirmed that the converted fMRI activity patterns preserved sufficient perceptual content for reconstructing visual images (Figures 4.9 and 4.10).

Furthermore, this finding is further confirmed with other methods of functional alignments (Figure 4.14, 4.15, and 4.16), showing that the finding is not specific to neural code converters but rather a general neuroscience finding. The differences in the conversion accuracies, decoding accuracies, and the reconstruction qualities across methods of functional alignments probably arises from the constraints imposed in the transformation matrices. Optimal transport imposes stronger constraints than other methods, particularly, it maps all source voxels to all target voxels exhaustively, with every voxel having an equal weight. Although it is slightly relaxed by the regularization, this constraint is unnatural for visual image reconstruction because not all voxels are equally important in the reconstruction. i.e. only a subset of voxels is critical for the reconstruction. Therefore, optimal transport using whole VC cannot guarantee those voxels are optimally converted across individuals for the reconstruction (Ho et al., 2023). Nevertheless, the differences did not prevent the conclusion that methods of functional alignment works well on preserving hierarchical visual information, providing a robustness test for my findings.

Chapter 5

Effectiveness of data-driven hierarchical neural code converter - A comparison

5.1 Introduction

In Chapter 3, neural code converters were trained using 2,400 samples and were purely data-driven without knowing the visual cortical hierarchy. Remarkably, the results of Chapter 3 demonstrate that the converters are capable of automatically detecting the visual cortical hierarchy. Furthermore, Chapter 4 demonstrate the neural code converters can preserve the hierarchical and fine-grained visual features, enabling the visual image reconstruction. This showcased the capability of data-driven approach to discover new findings in the neuroscience research.

Data-driven approach undoubtedly is a powerful tool, but it requires sufficient data to be useful. In recent years, there has been a growing trend to increase the volume of data collected for neuroscience research. In functional magnetic resonance imaging (fMRI) studies, this has been achieved either through an increase in the number of subjects (potentially up to a few ten thousands) or an increase in the number of samples (up to hundreds of thousands/a few ten hours of fMRI scan) collected per subject (Naselaris et al., 2021). Examples of large scale projects include the Human Brain Project and the Human Connectome Project. This surge in data availability has created new opportunities for data-driven studies.

In neuroscience, the data-driven approach offers a unique way to extract meaningful insights from complex and large-scale datasets. Unlike the traditional hypothesis-driven framework, which begins with a predefined theory or assumption, the data-driven approach allows researchers to discover novel patterns and relationships within the data without any prior assumptions. This is typically achieved using advanced machine learning and artificial intelligence algorithms capable of handling the immense complexity and high-dimensional nature of neuroscientific data. For example, Craddock et al. (2012) employed a data-driven clustering algorithm for brain parcellation to generate an fMRI atlas. This approach was undertaken without reliance on predetermined brain partitions and labels, offering a new perspective on brain organization.

A fundamental aspect of this thesis centers on the utilization of data-driven methodologies for understanding the hierarchical visual system. It invites an interesting comparison with a neural code conversion technique that explicitly incorporates established knowledge about the visual cortical hierarchy. Could such an approach outperform the data-driven hierarchical neural code converter in preserving perceptual contents? Given that visual hierarchy information augments the neural code converter, the hypothesis is that the performance of a neural code converter that acknowledges visual hierarchy should be superior, particularly when the data-driven approach shows limited effectiveness due to scarce data availability. The 6,000 training samples available in this study provides a good opportunity to scrutinize the effectiveness of data driven hierarchical neural code converters under different training sample conditions. How many training examples are necessary for the data-driven hierarchical neural code converter to match the efficacy of a neural code converter that respects visual hierarchy? The primary focus of this chapter, thus, is to ascertain which approach demonstrates better performance in DNN feature decoding and visual image reconstruction, and to scrutinize how the amount of training samples influences their performance.

5.2 Methods

5.2.1 Visual subarea-wise conversion

Chapter 3 introduced a method of neural code conversion that takes a whole Visual Cortex (VC) brain activity pattern from the source subject as input and predicts the activity values of voxels in the target subject. This method is referred to as "whole VC conversion." This chapter presents a different form of neural code conversion, respecting the cortical hierarchy, where the predicted activity value of a voxel in the target area is derived solely from the corresponding source area of the source subject (Figure 5.1). For instance, a voxel activity in the target's V1 is only predicted from the source subject's V1 subarea. This approach is hence termed "subarea-wise conversion." The predicted voxel activities across the five

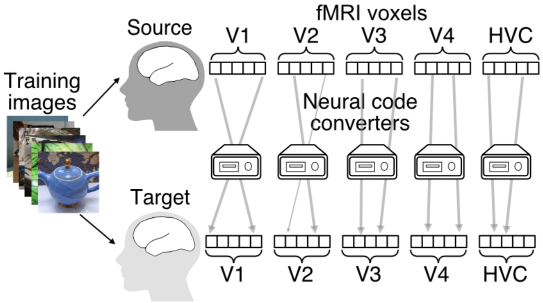

Figure 5.1: Schematics of the subarea-wise conversion. A converter model was trained on a set of fMRI responses to an identical stimulus sequence. The predicted activity values for a voxel in a target area were predicted exclusively from the source subject's brain activity patterns within the corresponding source area.

subareas (V1-V4, HVC) are integrated into a whole VC brain activity pattern, which is further decoded into DNN features and reconstructed into visual images.

5.2.2 Statistics

In comparing the subarea-wise conversion to the whole VC conversion, a comprehensive analysis was conducted through ANOVA, wherein DNN feature decoding accuracies and identification accuracies were subjected to scrutiny. In this analysis, the conversion type assumed the role of a repeated measure factor, while the DNN layer functioned as a between-subject factor. Because encompassing millions of DNN units and their corresponding decoding accuracies (profile correlations) always lead to statistical significance, the focus of group-level ANOVA was directed solely towards the DNN feature decoding accuracies, culminating in the computation of F scores, p values, and effect sizes η^2. This computation was predicated on the mean DNN feature decoding accuracies sourced from 20 individual pairs.

Regarding the identification accuracies, ANOVA analysis was conducted using accuracies of individual reconstructed image as data points. This facilitated the computation of F scores, p values, and effect sizes η^2 within each individual pair. At the group level, mean identification accuracies derived from the aggregation of images across 20 individual pairs were harnessed as data points, further contributing to the computation of F scores, p values, and the effect size η^2.

5.3 Results

This section presented the results of subarea-wise conversion and the analysis of changing training data amount. The content of this section is based on the section 2.5: *Visual subarea-wise conversion* and the section 2.6: *Varying the number of training data* of Ho et al. (2023).

5.3.1 Subarea-wise conversion accuracy

In order to investigate the influence of neural code converters that adhere to the visual hierarchy on the performance of reconstruction, I performed subarea-wise conversion that predicted the activity values of a voxel in a target area only from the source subject's corresponding source area (see section 5.2.1: Visual subarea-wise conversion). As in Chapter 3, the conversions were evaluated based on pattern and profile correlations (see section 3.2.3: Conversion accuracy). All individual pairs showed comparable conversion accuracies to the whole VC conversion, with the mean pattern correlation being 0.58 ± 0.07 and the mean profile correlation being 0.55 ± 0.06 for VC (Figure 5.2 for individual pairs; Figure 5.3 for group results). Following this, I executed DNN feature decoding and visual image reconstruction using the whole VC, and compared the results with the whole VC conversions.

5.3.2 DNN feature decoding via subarea-wise conversion

In the DNN feature decoding of the natural images, the subarea-wise conversion showed similar but slightly lower decoding accuracy than the whole VC conversion across layers in all individual pairs (Figure 5.4 left) and at the group level (Figure 5.5 left; ANOVA on the means of individual pairs, effect of conversion type with the DNN layer as a between-subject factor, $F(1, 361) = 1959$, $p < .001$, $\eta^2 = 0.84$; see section 5.2.2: Statistics). Similar results were obtained for the artificial images in some individual pairs and at the group level (Figure 5.4 right for individual pairs; Figure 5.5 right for group results; ANOVA on the means of individual pairs, $F(1, 361) = 260.6$, $p < .001$, $\eta^2 = 0.42$).

5.3.3 Visual image reconstruction via subarea-wise conversion

Reconstructed images resulting from the subarea-wise conversions demonstrated a similar visual quality to those derived from the whole VC conversions. (Figure 5.6). In the identification analysis of the natural images, only 2/20 pairs showed significantly higher accuracies for the subarea-wise conversion; 6/20 pairs showed higher significant accuracies for the

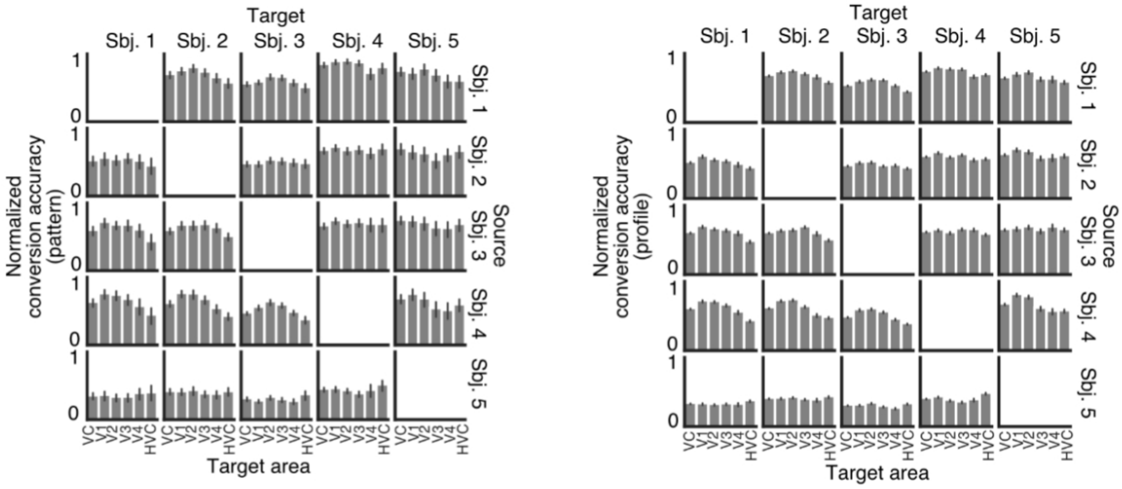

Figure 5.2: Subarea-wise conversion accuracies for individual pairs. The pattern correlation coefficients for 50 visual stimuli were used to calculate a mean conversion accuracy (pattern) and its 95% confidence interval. The profile correlation coefficients for voxels were used to calculate a mean conversion accuracy (profile) and its 95% confidence interval (right; error bars on left panel, 95% C.I. across visual images; error bars on the right panel, 95% C.I. across voxels).

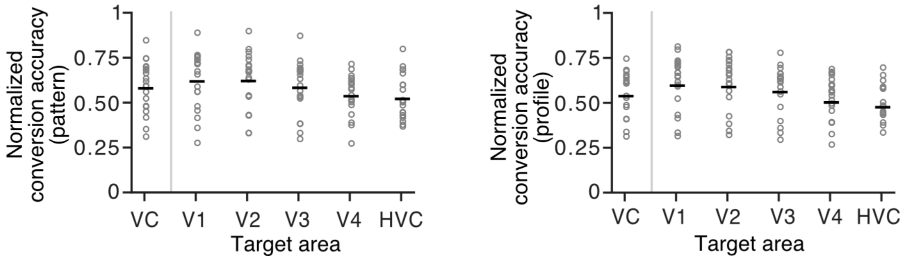

Figure 5.3: Mean and distribution of subarea-wise conversion accuracies across individual pairs. Distributions of normalized pattern or profile correlation coefficients across 20 individual pairs are shown for VC and visual subareas. Each horizontal black dash indicates the mean value over 20 individual pairs; each circle represents the correlation coefficients of an individual pair.

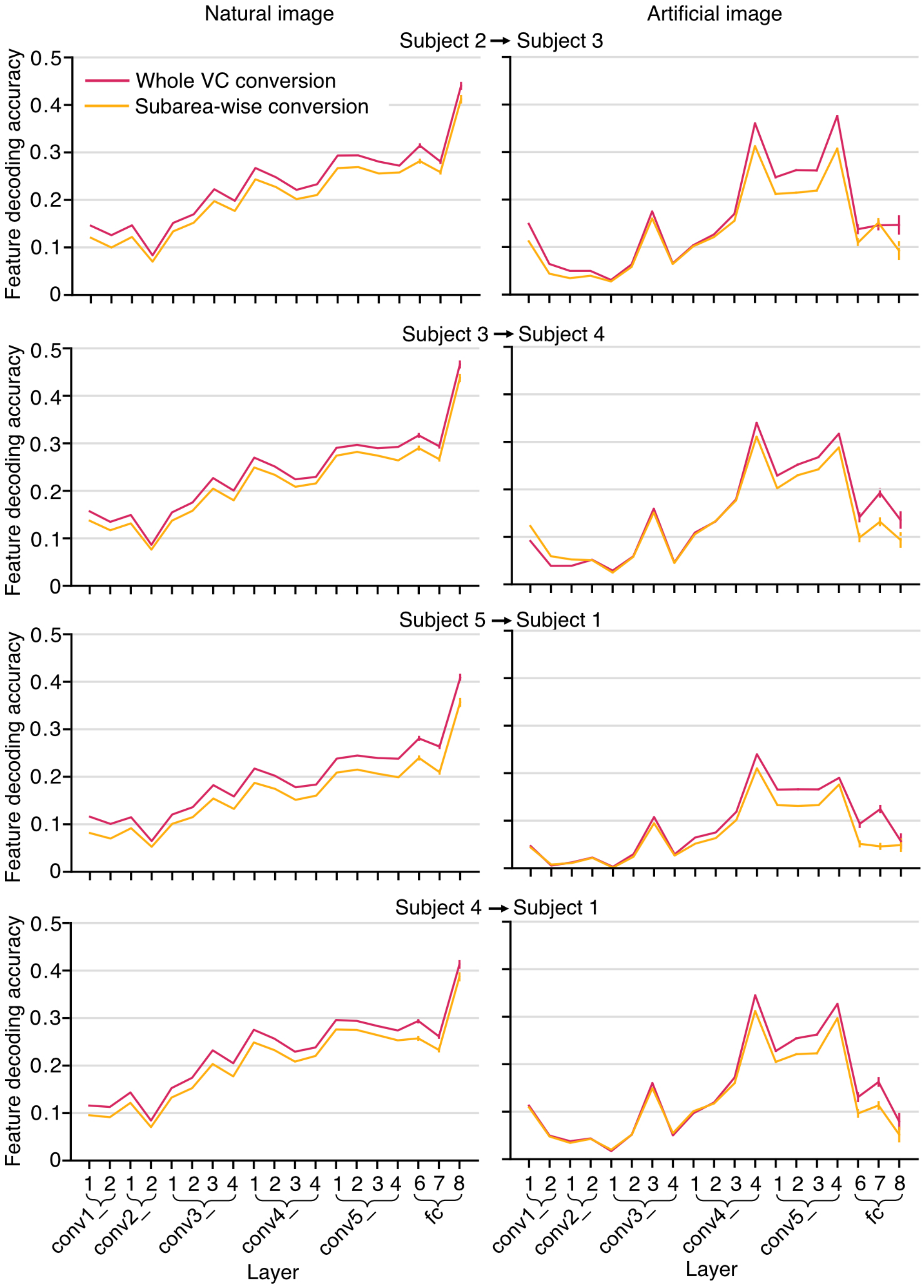

Figure 5.4: DNN feature decoding accuracy of natural images and artificial images for four representative individual pairs. The decoding accuracies for all DNN units in each layer were used to calculate a mean decoding accuracy and its 95% confidence interval (error bars, 95% C.I. across DNN units).

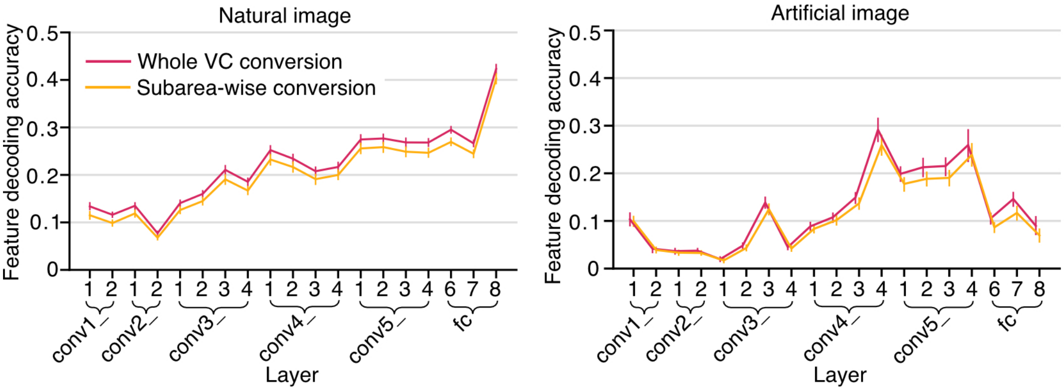

Figure 5.5: Mean DNN feature decoding accuracy across all individual pairs (error bars, 95% C.I. from 20 individual pairs).

whole conversion (Figure 5.7 left; ANOVA in individual pairs; effect of conversion type with the DNN layer feature as a between-subject factor). At the group level, the subarea-wise conversion showed lower accuracies (Figure 5.8 left; ANOVA on the means of individual pairs, $F(1, 171) = 11.2$, $p < .001$, $\eta^2 = 0.062$). In the identification analysis of the artificial images, 3/20 pairs showed higher significant accuracies for the subarea-wise conversion; 2/20 pairs showed higher significant accuracies for the whole VC conversion (Figure 5.7 right; ANOVA in individual pairs), while no statistical difference was found at the group level (Figure 5.8 right; ANOVA on the means of individual pairs, $F(1, 171) = 3.87$, p =0.051, η^2 = 0.022). These results indicate that constraining neural code conversion to respect cortical hierarchy does not seem to contribute to the improvement of visual image reconstruction. Rather, the flexibility of the mapping with the whole VC conversion could be beneficial as indicated by the slightly superior performance with the natural images.

5.3.4 Varying the number of training data

The current inter-individual analysis outcomes have been derived from the utilization of 2,400 samples for converter training. In this section, my current investigation delved into an exploration of the influence of the training sample quantity on image reconstruction quality. This exploration entailed a systematic variation of the data employed for converter training, spanning different sample sizes (300, 600, 900, 1,200, 2,400, 3,600, 4,800, and 6,000 training samples), while concurrently utilizing the complete dataset of the target subject for decoder training (6,000 samples). Particularly, a comparative analysis was executed between the whole VC and the subarea-wise conversions. The content of this section is based on the section 2.6: *Varying the number of training data* of Ho et al. (2023).

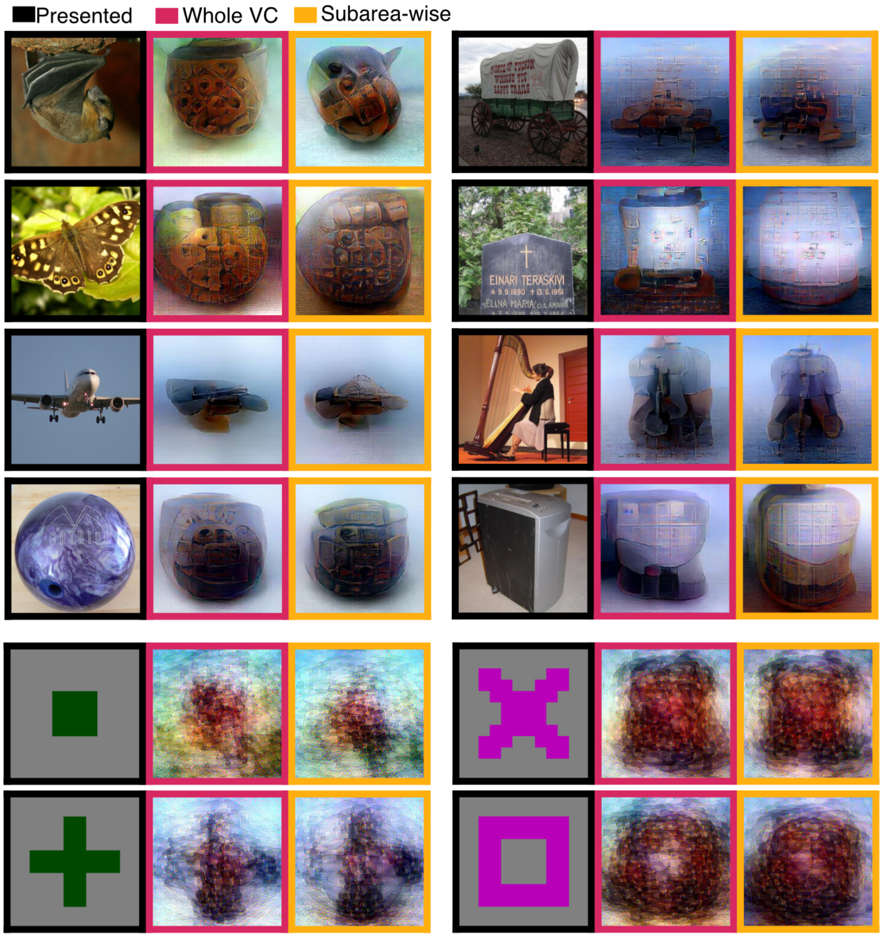

Figure 5.6: Reconstructed natural and artificial images via whole VC conversions and subarea-wise conversions. The reconstructions shown under the two analytical conditions for each stimulus image were all from the same source subject. The results for different stimulus images are from different source subjects.

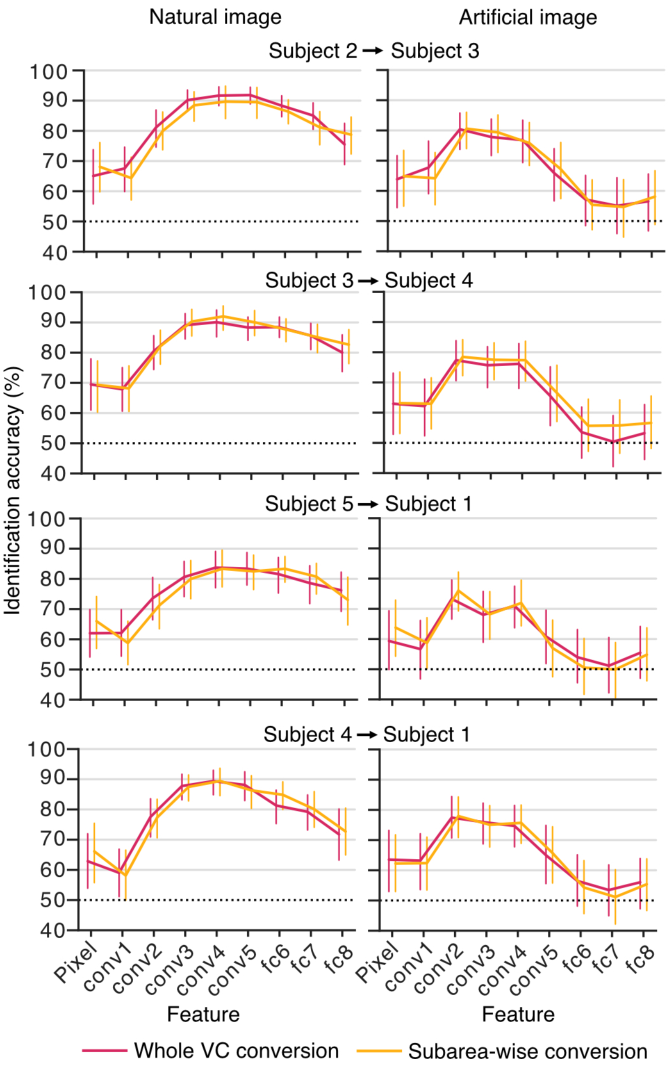

Figure 5.7: Identification accuracies with the reconstructed natural and artificial images for four representative individual pairs. DNN features of images were extracted from the eight layers of the AlexNet model. The identification accuracies for individual reconstructed images were used to calculate a mean identification accuracy and its 95% confidence interval (chance level = 50%).

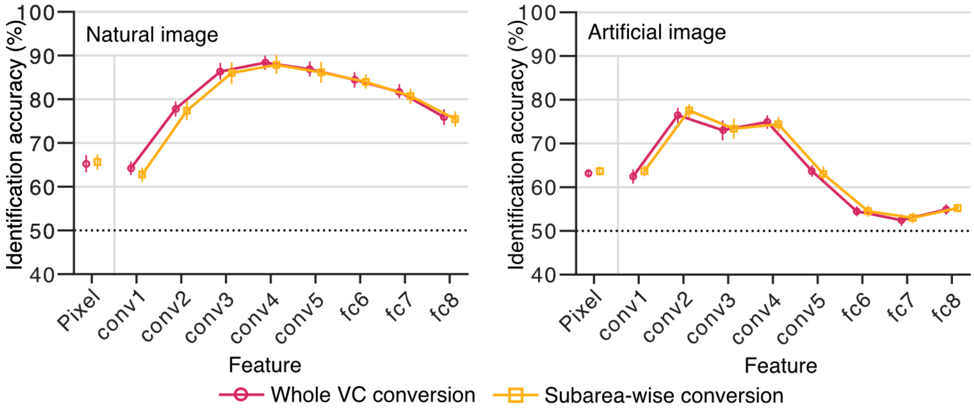

Figure 5.8: Mean identification accuracies across all individual pairs. DNN features of images were extracted from the eight layers of the AlexNet model (left, natural images; right, artificial images; error bars, 95% C.I. from 20 individual pairs; dotted lines, chance level = 50%).

The reconstructed images retained a discernible quality even with a reduction in the number of training samples. Specifically, using converters trained on 300 samples still produced recognizable images in both the whole VC and subarea-wise conversions (Figure 5.9). This result indicates that image reconstruction using converters with a small number of training data is feasible, without the need to collect a full set of fMRI data for each subject.

The reconstructed images were further evaluated using identification analysis (see section 4.2.5: Identification analysis). The identification accuracies increased with the number of training samples, approaching the accuracy of the within-individual (Within) condition (see Figure 5.10 left for individual pairs and Figure 5.11 top for group results). The subarea-wise and whole VC conversions showed similar accuracies with more than 1,200 training samples, but the subarea-wise conversion outperformed the whole VC conversion with 1,200 or fewer training samples (ANOVA within individual pairs at each training sample number, effect of conversion type, $p < .05$ in 18, 10, 4, 3, 2, 1, 4, and 1 out of 20 pairs for the eight training sample numbers, respectively; group analysis on the mean accuracies of individual pairs, $p < .05$ at 300, 600, and 900 samples; Bonferroni-corrected by eight). Similar results were obtained for the artificial images (Figure 5.10 right for individual pairs; ANOVA within individual pairs, effect of conversion type, $p < .05$ in 10, 8, 5, 2, 2, 1, 1, and 2 out of 20 pairs for the eight training sample numbers, respectively; Figure 5.11 bottom for group results; group analysis on the mean accuracies of individual pairs, $p < .05$ at 300, 600, and 900 samples; Bonferroni-corrected by eight). Overall, incorporating the cortical hierarchy constraint into neural code conversion does not yield enhanced reconstruction; however, it proves advantageous in scenarios with limited training data.

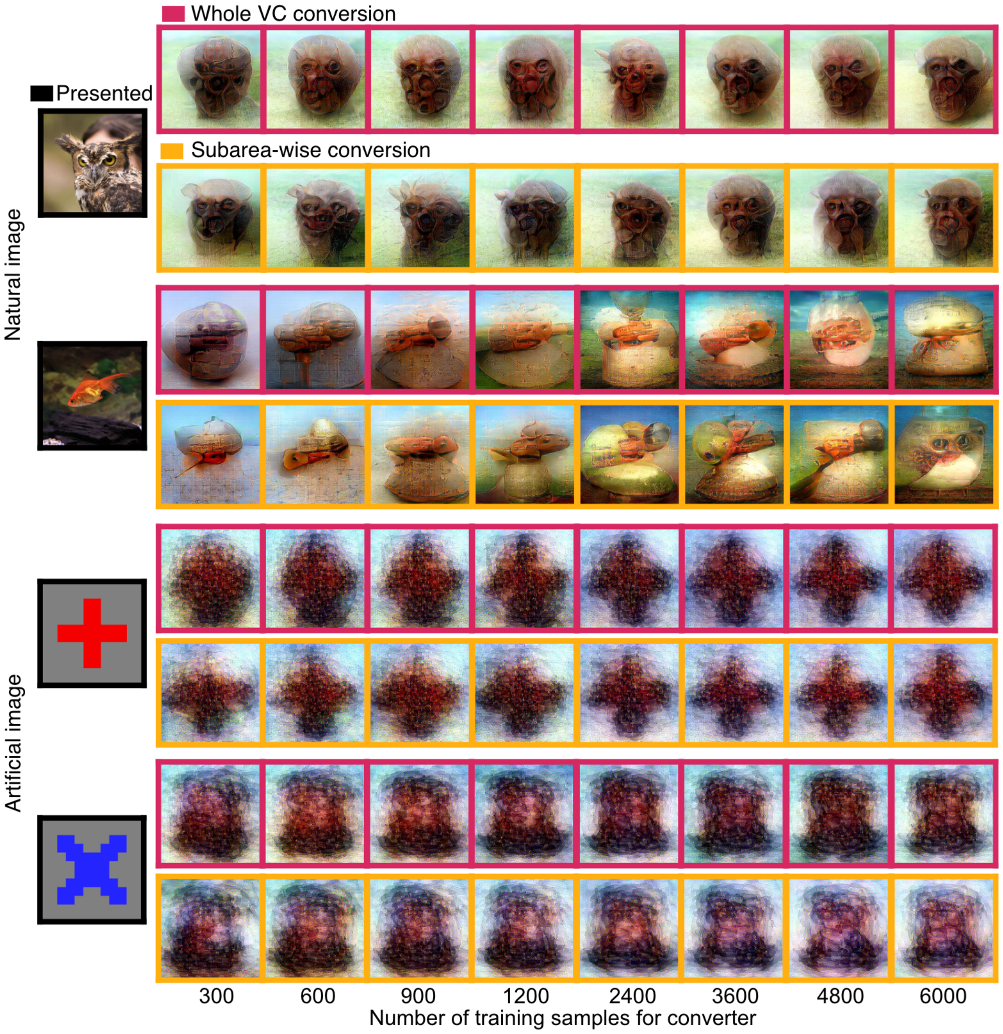

Figure 5.9: Reconstructed images via neural converters trained with a varying amount of data. Reconstructed natural images and artificial images were produced from the same subject pair respectively (natural image: from Subject 2 to Subject 3; artificial image: from Subject 3 to Subject 1

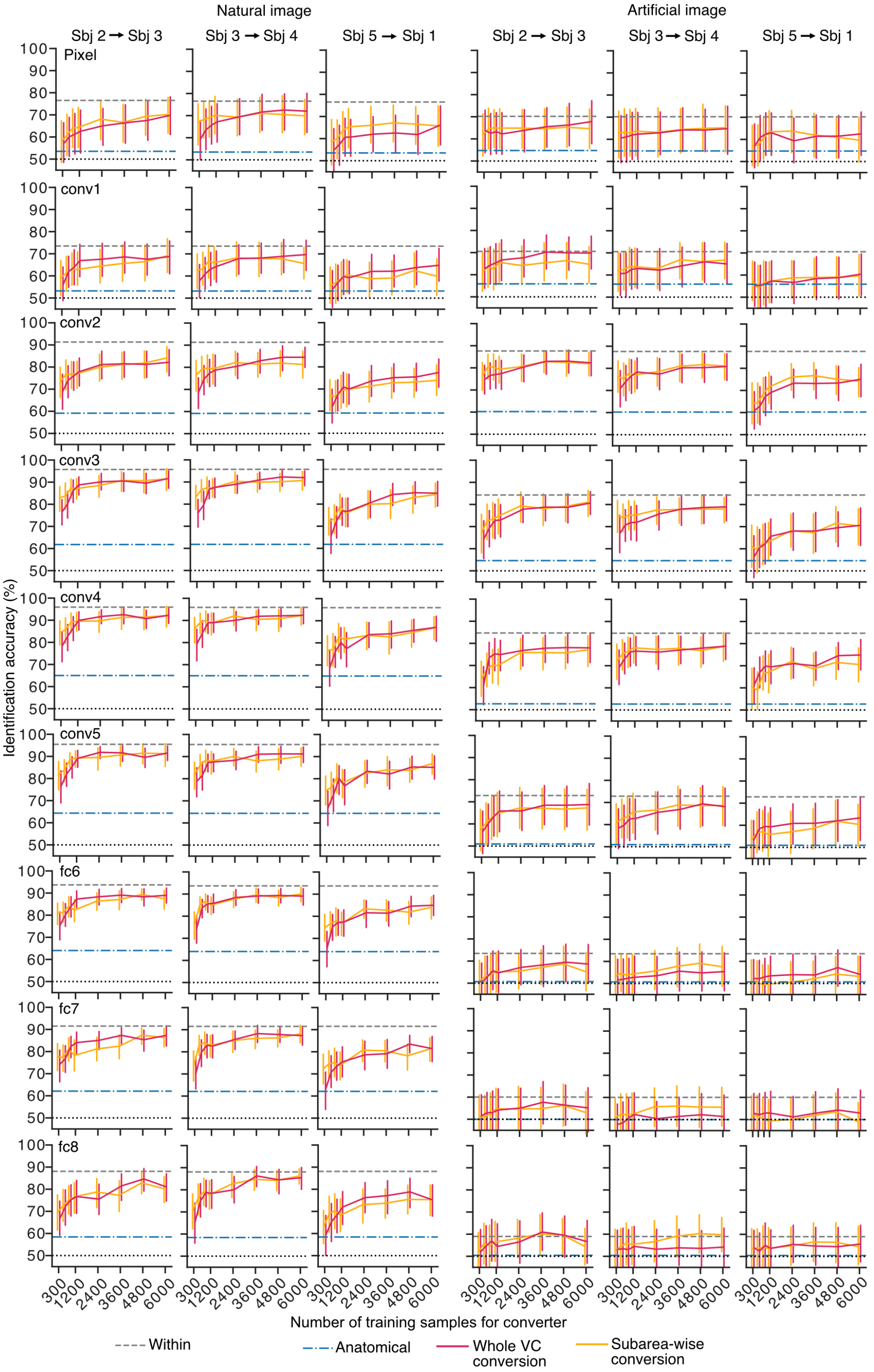

Figure 5.10: Identification accuracies of representative individual pairs with different numbers of training samples for converters. Identification accuracies were calculated with the pixel values and the extracted DNN feature values (AlexNet) from the reconstructed images. For each feature and each training sample condition, the identification accuracies for 50/40 reconstructed images were used to calculate a mean identification accuracy and its 95% confidence interval. The results are shown together with those from the within-individual condition (Within) and the anatomical alignment (Across-anatomical) (dotted lines, chance level = 50%).

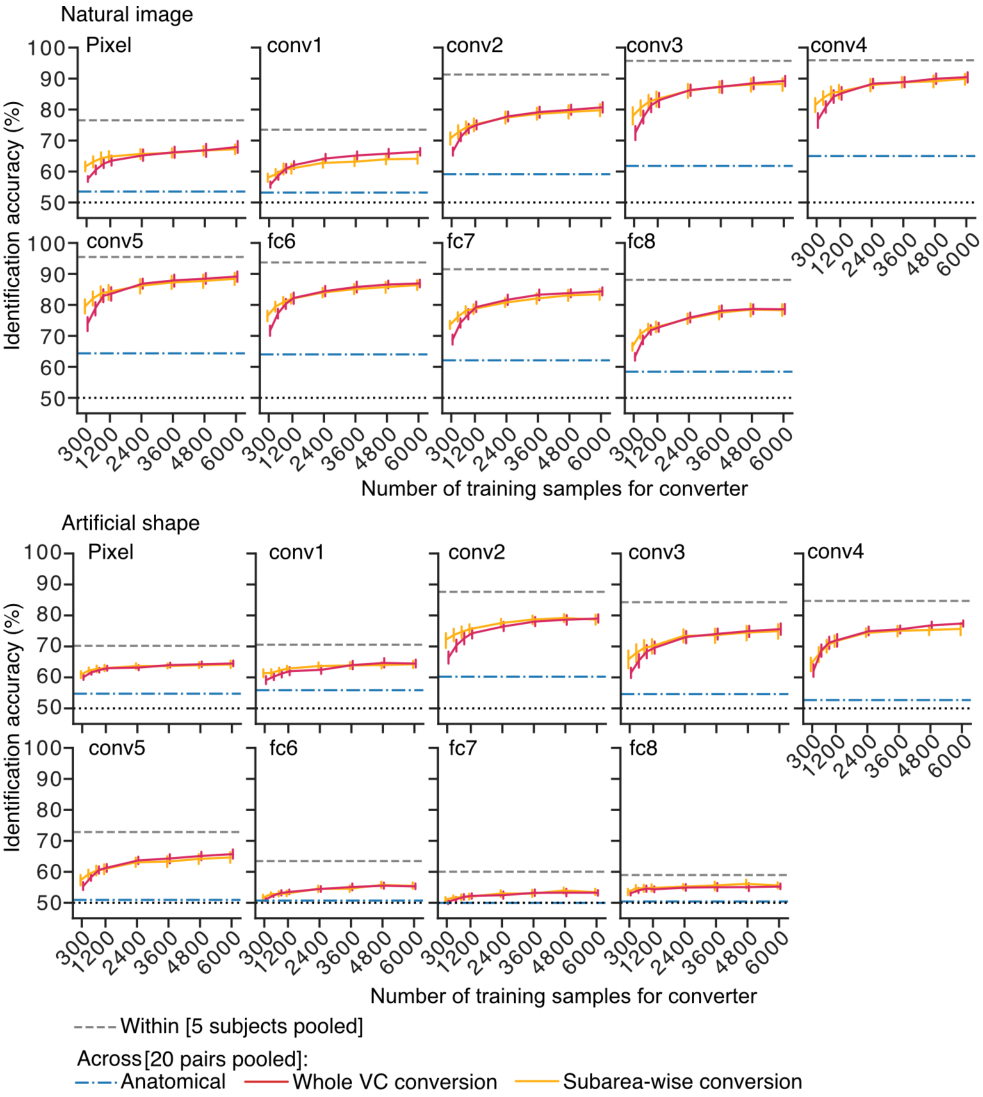

Figure 5.11: Mean identification accuracies across all individual pairs with varying numbers of training data for the whole VC and subarea-wise conversions. Identification accuracies were calculated with the pixel values and the extracted DNN feature values (AlexNet) from the reconstructed images. The results are shown together with those from the within-individual condition (Within) and the anatomical alignment (Across-anatomical) (error bars, 95% C.I. from 20 individual pairs for whole VC and subarea-wise conversions; dotted lines, chance level = 50%).

5.4 Discussions

This chapter is dedicated to comparing the effectiveness of two neural code conversion methods: whole VC conversion and subarea-wise conversion. The subarea-wise conversion marginally underperformed the whole VC conversion with sufficient training data in the inter-individual visual image reconstruction (Figures 5.4, 5.5, 5.7, and 5.8), with the whole VC conversions achieving slightly higher DNN feature decoding accuracy and marginally higher identification accuracy in the reconstruction. However, the subarea-wise conversion demonstrates better performance than the whole VC conversion when data is scarce (Figures 5.10 and 5.11). This result shows that when sufficient training data are available, the whole VC conversion can implicitly learn the information about explicit labels of visual subareas, without the need to explicitly impose the hierarchy constraint. The content of this section is based on the section 3: *Discussion* of Ho et al. (2023).

Training a full visual image reconstruction model requires an fMRI dataset that is costly and takes a long time to collect. In this study, the training of DNN feature decoders involved a dataset comprising 6,000 data samples, corresponding to an approximate data collection duration of 800 minutes. In fMRI studies, this long data collection time is impractical for most people. Nevertheless, it is feasible to achieve a reduction in the number of required data samples, such as employing 300 samples, for training a neural code converter and conducting inter-individual visual image reconstruction, albeit at the cost of sacrificing a certain degree of visual fidelity in the reconstructed images. In particular, the neural code converter is engineered to capture the relationships between individuals' voxels across a diverse array of visual scenes and holds promise for combination with other decoding models. The inter-individual decoding method with the neural code converter has the potential to reduce the time and costs of fMRI data collection.

Despite the promising performance of inter-individual image reconstruction, it did not surpass within-individual image reconstruction, regardless of the amount of converter training data used (Figures 5.10 and 5.11). This discrepancy may be attributed to the linear constraint applied during conversion, which might be too restrictive to capture complex statistical relationships like nonlinearity in brain activity patterns. Additionally, a brain's response to a stimulus comprises a consistent stimulus-evoked response across individuals, an idiosyncratic stimulus-evoked response and a noise component (Nastase et al., 2019). The brain decoders might leverage the idiosyncratic responses that could not be converted across subjects, as well as noise components. As a result, the inter-individual visual image reconstruction thus underperformed the within-individual visual image reconstruction.

Chapter 6

Application of neural code converters to pooling data analysis

6.1 Introduction

Brain decoding studies, which aim to understand and predict the complex processes within our brains, are continuously faced with a significant challenge—limited availability of data. This becomes especially problematic when the decoding model is unable to adequately account for individual differences among subjects. Chapters 3 and 4 shed light on how the application of neural code conversions could mitigate this issue. The process of neural code conversion aims to preserve the complex, hierarchical structure of visual information, thereby enabling the inter-individual reconstruction of visual images. This implies that even if the brain's workings are unique to each individual, there is still a level of commonality or universal "code" that can be shared and applied across different individuals.

Chapter 5 further demonstrated the efficacy of this approach, revealing successful application even with a relatively modest set of training samples—merely 300. The study underscored the potential of neural code conversions in reconstructing discernible shapes and colors from brain activity.

However, there's a critical limitation to inter-individual visual image reconstruction between a pair of subjects—the paucity of data that can be collected from each subject. Specifically, the process of visual image reconstruction necessitates an extensive sampling for each individual, aiming to capture a wide range of brain responses to diverse visual scenarios. However, factors such as the financial cost of data collection, physical limitations of the subjects

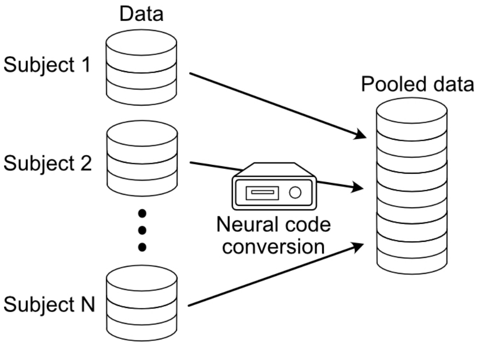

Figure 6.1: Basic idea of pooling data through neural code conversion. Multiple subjects' data can be combined into one dataset through the neural code conversion. The pooled data can be further used for other analyses.

(including time or comfort constraints during neuroimaging scans), typically restrict the total hours of fMRI scanning time to just a few ten hours. This, in turn, translates into only a few tens of thousands of training samples per subject.

Neural code conversion presents a potential solution to this challenge by pooling data from different subjects (Figure 6.1). In this approach, brain data from several subjects is consolidated into a single dataset, thereby substantially increasing the quantity of available data. Following this, a new decoding model can be trained utilizing this pooled dataset, which inherently incorporates a broader variety of brain activity.

Thus, the primary goal of this chapter is to delve deeper into the impact of data pooling on the performance of inter-individual visual image reconstruction.

6.2 Methods

6.2.1 Pooling data into a target subject

In the context of a source and target subject pair, I pooled data from the remaining three subjects, employing their respective converters (trained with 6,000 samples), and transformed this collective data into the target brain space. This procedure resulted in an aggregate dataset comprising a total of 24,000 samples (with each of the four subjects contributing 6,000 samples). This pooling process was facilitated through whole VC conversion, as depicted in Figure 6.2. Following this, I proceeded to train decoders on the pooled dataset, referring to these specific decoders as "multiple-subject feature decoders." This nomenclature serves as a point of distinction from the "single-subject feature decoders," which were exclusively trained on the target subject within its original brain space. In the process of training the

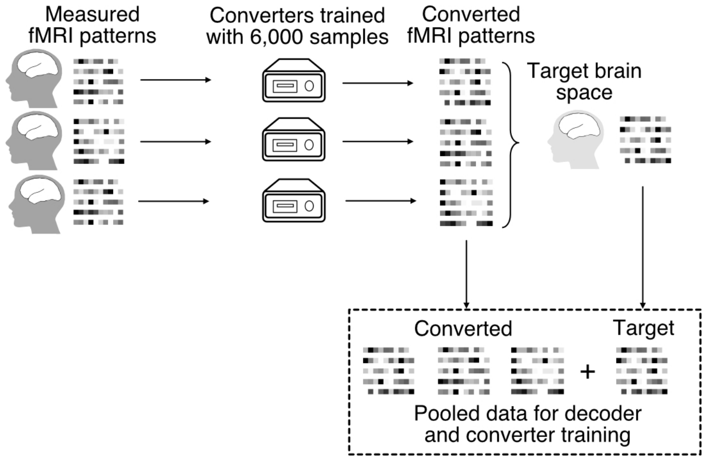
Measured fMRI patterns
Converters trained with 6,000 samples
Converted fMRI patterns
Target brain space
Converted
Target
+
Pooled data for decoder and converter training

Figure 6.2: Illustration of the pooling procedure. For a pair of a source (not shown) and a target subject, the training data of the other three subjects were converted into the target subject's brain space. DNN feature decoders were then retrained using the converted data from these three subjects in combination with the original data from the target subject (24,000 samples).

neural code converter between the source subject's data and the combined dataset, each set of 2,400 samples from the source subject was paired with an equivalent set of 2,400 samples drawn from the pooled data of the four pooled subjects. The converted brain activity from the source subject underwent DNN feature decoding with the multiple-subject feature decoders and then visual image reconstruction. The results were compared with those generated from the single-subject feature decoders.

6.2.2 Statistics

In comparing the "multiple-subject feature decoders" to the "single-subject feature decoders", a comprehensive analysis was conducted through ANOVA, wherein DNN feature decoding accuracies and identification accuracies were subjected to scrutiny. In this analysis, the conversion type assumed the role of a repeated measure factor, while the DNN layer functioned as a between-subject factor. Because encompassing millions of DNN units and their corresponding decoding accuracies (profile correlations) always lead to statistical significance, the focus of group-level ANOVA was directed solely towards the DNN feature decoding accuracies, culminating in the computation of F scores, p values, and effect sizes η^2. This computation was predicated on the mean DNN feature decoding accuracies sourced from 20 individual pairs.

Regarding the identification accuracies, ANOVA analysis was conducted using accuracies of individual reconstructed image as data points. This facilitated the computation of F scores, p

values, and effect sizes η^2 within each individual pair. At the group level, mean identification accuracies derived from the aggregation of images across 20 individual pairs were harnessed as data points, further contributing to the computation of F scores, p values, and the effect size η^2.

6.3 Results

This section presents the results of pooling data analysis and is based on the section 2.7: *Pooling data from multiple subjects* of Ho et al. (2023).

6.3.1 Pooling data into a target subject

Building upon the pooling procedure, the multiple-subject feature decoders were evaluated against single-subject feature decoders via DNN feature decoding. Inter-individual DNN feature decoding analysis on the natural images showed a small improvement in accuracy across all layers in the multi-subject condition as compared with the single-subject condition. The multiple-subject condition yielded better performance than the single-subject condition both in individual pairs (Figure 6.3 left) and at the group level (Figure 6.4 left; ANOVA, effect of decoder type, $F(1, 361) = 1968$, $p < .001$, η^2= .85; see section 6.2.2: Statistics). Similar results were obtained for artificial images, with the multiple-subject condition showing higher accuracies (see Figure 6.3 right for individual pair results and Figure 6.4 right for group results; ANOVA, effect of decoder type, $F(1, 361) = 172$, $p < .001$, η^2= .32).

Reconstructed images derived from both the single- and multiple-subject decoders exhibited distinguishable visual quality; however, notable distinctions in visual attributes between the two conditions were not prominently evident (Figure 6.5). In the identification analysis of the reconstructed natural images, the multiple-subject condition showed slightly higher accuracies than the single-subject condition (Figure 6.6 left for individual pairs; ANOVA, effect of decoder type, $p < .05$ in 10/20 individual pairs; Figure 6.7 left for group results; effect of decoder type, $F(1, 171) = 75.6$, $p < .001$, η^2= .31). Similar results were obtained for artificial images, with the multiple-subject condition showing slightly higher identification accuracies than the single-subject condition (Figure 6.6 right for individual pairs; ANOVA, effect of decoder type, $p < .05$ in 6/20 individual pairs; Figure 6.7 right for group results; effect of decoder type, $F(1, 171) = 35.9$, $p < .001$, η^2= .17).

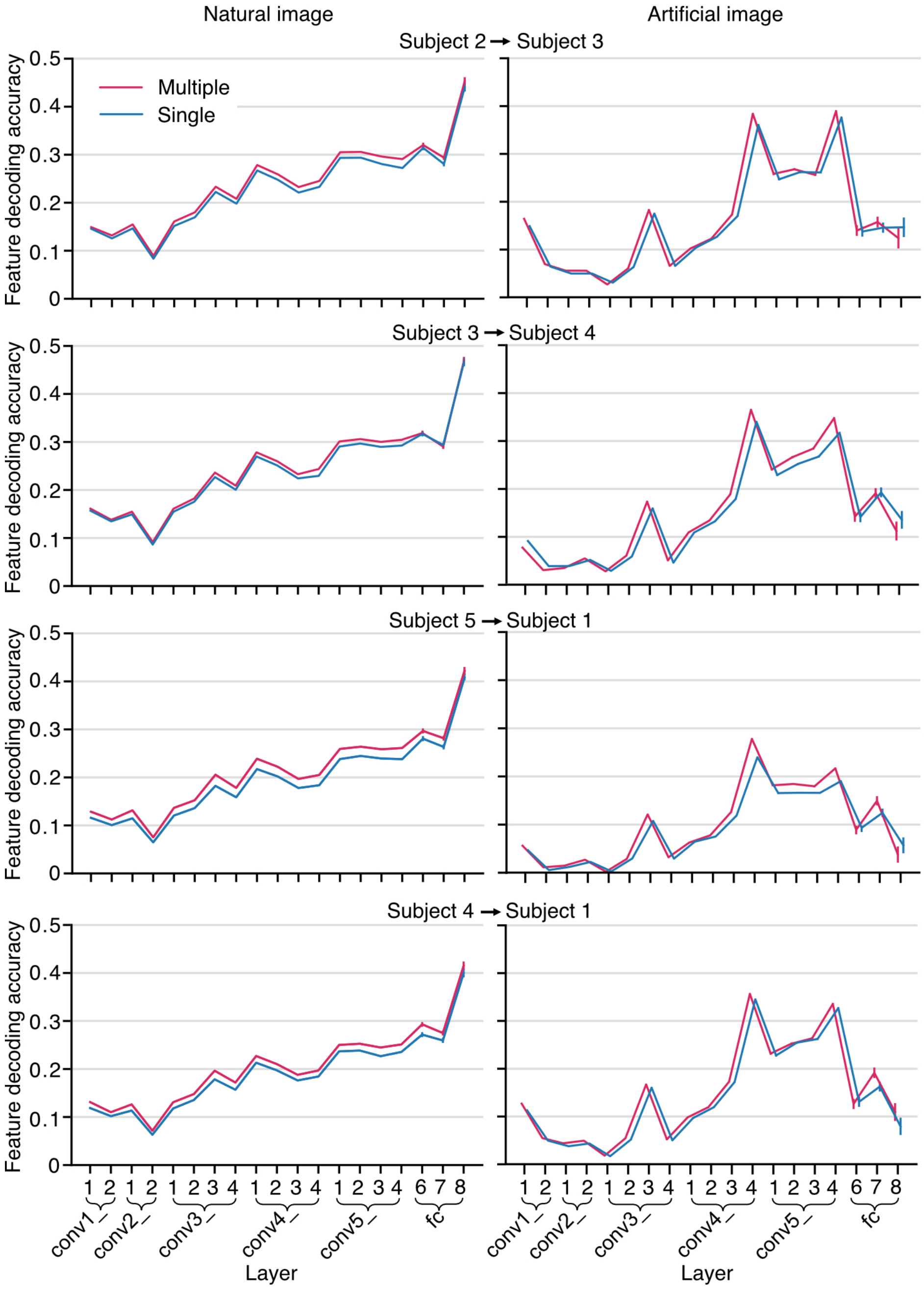

Figure 6.3: Inter-individual DNN Feature decoding accuracies of representative individual pairs with multiple- and single-subject feature decoders. The multiple-subject feature decoders were trained on pooled data, while the single-subject feature decoders were trained on a single subject's data. The decoding accuracies for all DNN units in each layer were used to calculate a mean decoding accuracy and its 95% confidence interval (error bars; 95% C.I. across DNN units).

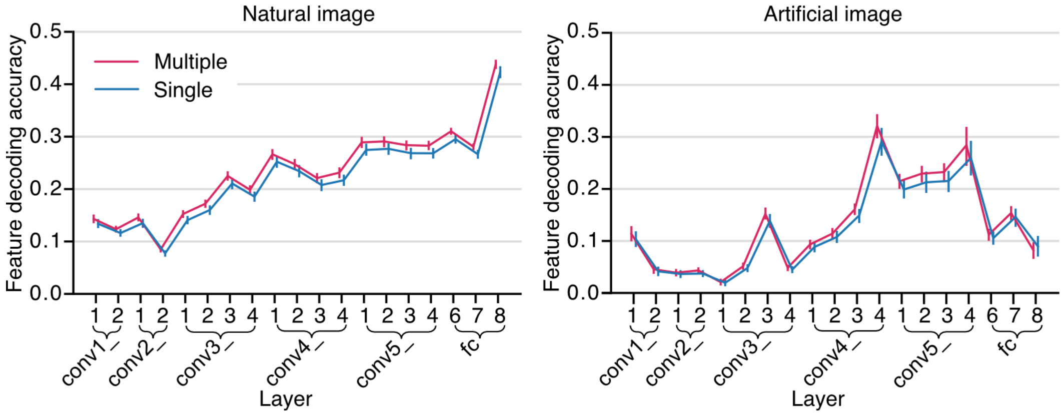

Figure 6.4: Mean DNN feature decoding accuracy obtained via multiple- and single-subject feature decoders across all individual pairs. The multiple-subject feature decoders were trained on pooled data, while the single-subject feature decoders were trained on a single subject's data. The accuracies were obtained from the source subjects' test dataset (error bars, 95% C.I. from 20 individual pairs).

6.3.2 Pooling data analysis with limited data availability

To examine the impact of limited data availability on the benefits of pooling multiple-subject data, I conducted a similar analysis using only 300 training samples for the source subject, reflecting situations where data collection is restricted due to cost constraints.

The DNN feature decoding analysis performed on natural images revealed a small enhancement in accuracy across certain layers within the multi-subject condition, when contrasted with the single-subject condition. The multiple-subject condition yielded better performance than the single-subject condition both in individual pairs (Figure 6.8 left) and at the group level (Figure 6.9 left; ANOVA, effect of decoder type, $F(1, 361) = 1300$, $p < .001$, η^2= .78). Similar results were obtained for artificial images, with the multiple-subject condition showing higher accuracies (see Figure 6.8 right for individual pair results and Figure 6.9 right for group results; ANOVA, effect of decoder type, $F(1, 361) = 30.2$, $p < .001$, η^2= .33).

The reconstructed images derived via both single- and multiple-subject decoders under restricted data availability yielded recognisable, albeit lower-quality visuals as anticipated, compared to the images in Figure 6.5, but the visual qualities were not substantially different between the two conditions (Figure 6.10). There was also a slight improvement for the identification accuracies of the reconstructed natural images in some of the pairs and at the group level when using the multiple-subject feature decoders (Figure 6.11 left for individual pairs; ANOVA, effect of decoder type, $p < .05$ in 5/20 individual pairs; Figure 6.12 left

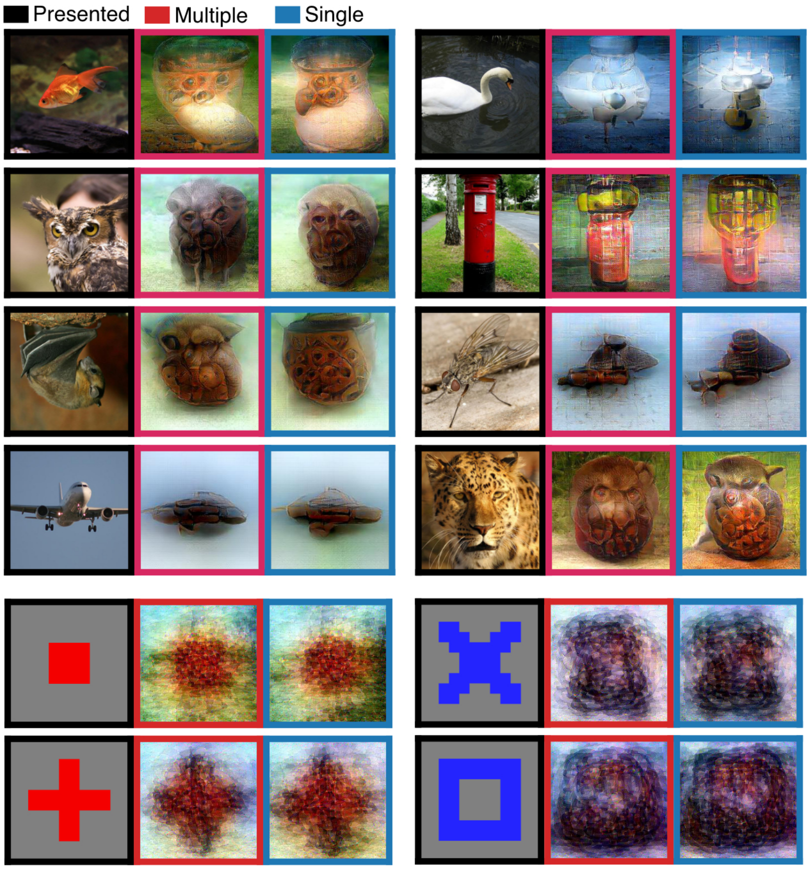

Figure 6.5: Reconstructed natural images for the multiple- and single-subject conditions. The reconstructions shown under the two analytical conditions for each stimulus image were all from the same source subject. The results for different stimulus images are from different source subjects.

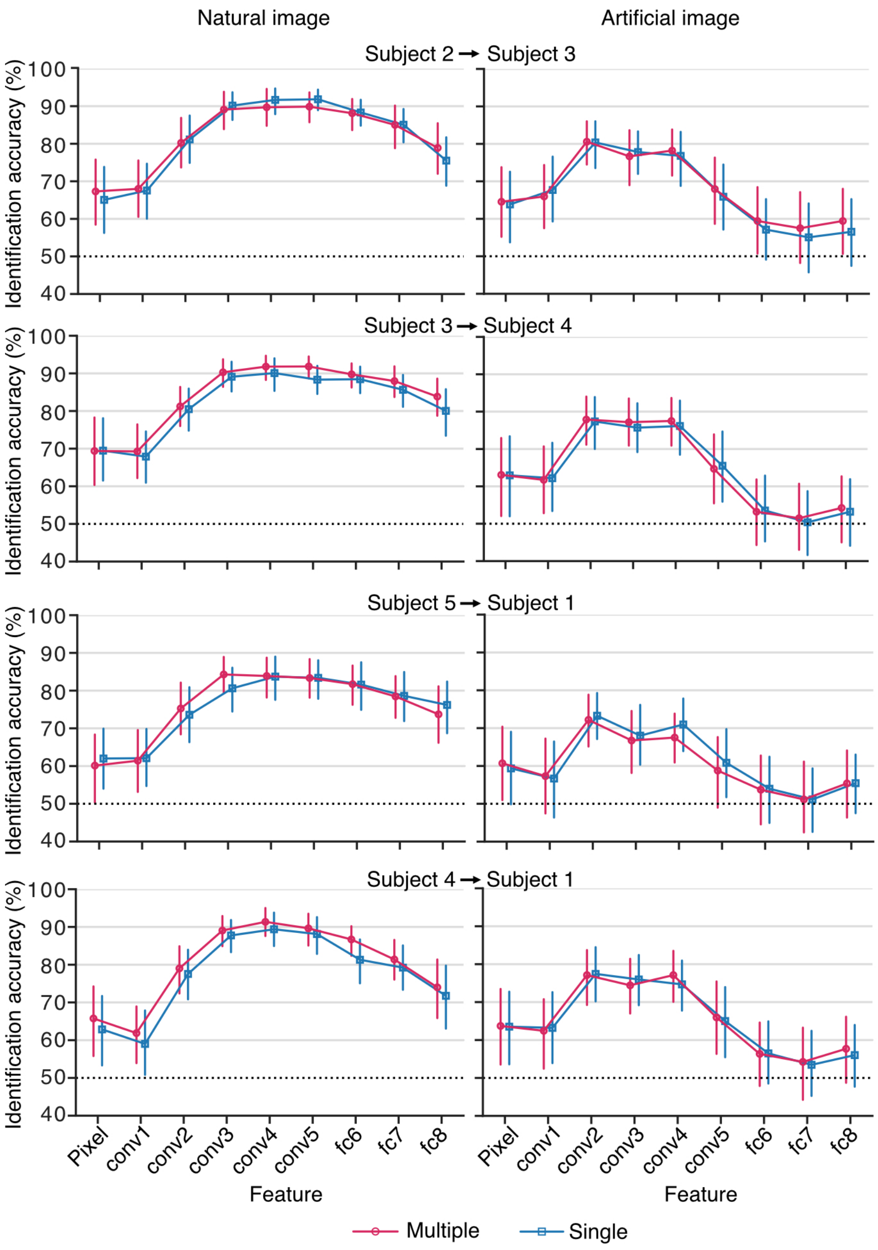

Figure 6.6: Identification analyses of representative individual pairs with multiple- and single-subject feature decoders. The identification accuracies for reconstructed images were used to calculate a mean identification accuracy and its 95% confidence interval. The multiple-subject feature decoders outperformed the single-subject feature decoders (dotted line, chance level = 50%).

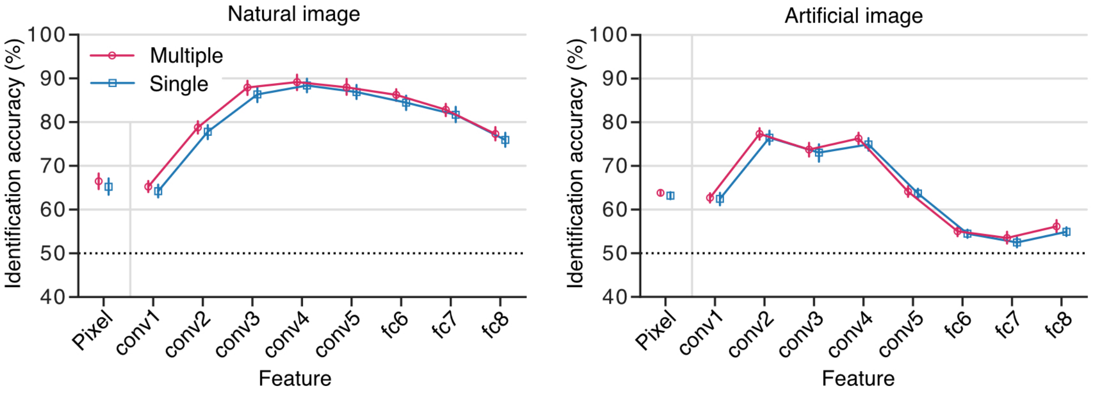

Figure 6.7: Mean identification accuracies obtained via multiple- and single-subject feature decoders across all individual pairs. The identification analysis was performed using the pixel values and the extracted DNN feature values of the reconstructions obtained via multiple- and single-subject feature decoders (error bars, 95% C.I. from 20 individual pairs; dotted lines, chance level = 50%).

for group results; effect of decoder type, $F(1, 171) = 28.3$, $p < .001$, $\eta^2 = .14$). Similar observations were made with artificial images, with the multiple-subject condition showing marginally higher identification accuracies than the single-subject condition (Figure 6.11 right for individual pairs; ANOVA, effect of decoder type, $p < .05$ in 4/20 individual pairs; Figure 6.12 right for group results; effect of decoder type, $F(1, 171) = 4.6$, $p < .05$, $\eta^2 = .03$). These results indicate that pooling multiple-subject data is somewhat beneficial for improving the accuracy of inter-individual decoding and reconstruction, even when data availability on the source subject is constrained. Nevertheless, it is noteworthy that this pooling did not result in a substantial improvement in the visual quality of the reconstructed images.

6.4 Discussions

This chapter investigates the effect of data pooling on the performance of inter-individual visual image reconstruction. The findings indicate that data pooling from multiple subjects did not yield a substantial enhancement in visual image reconstruction performance. One plausible explanation for this observation could be attributed to the inherent variability in data quality, wherein data originating from certain subjects contributed to relatively poor visual quality of the reconstructed images. Poor quality data limited the capability of the decoders to leverage the pooled data and resulted in a limited improvement in visual image reconstruction performance. Furthermore, it is noteworthy that linear regression models may exhibit limitations in effectively addressing the inherent feature mismatch present in brain

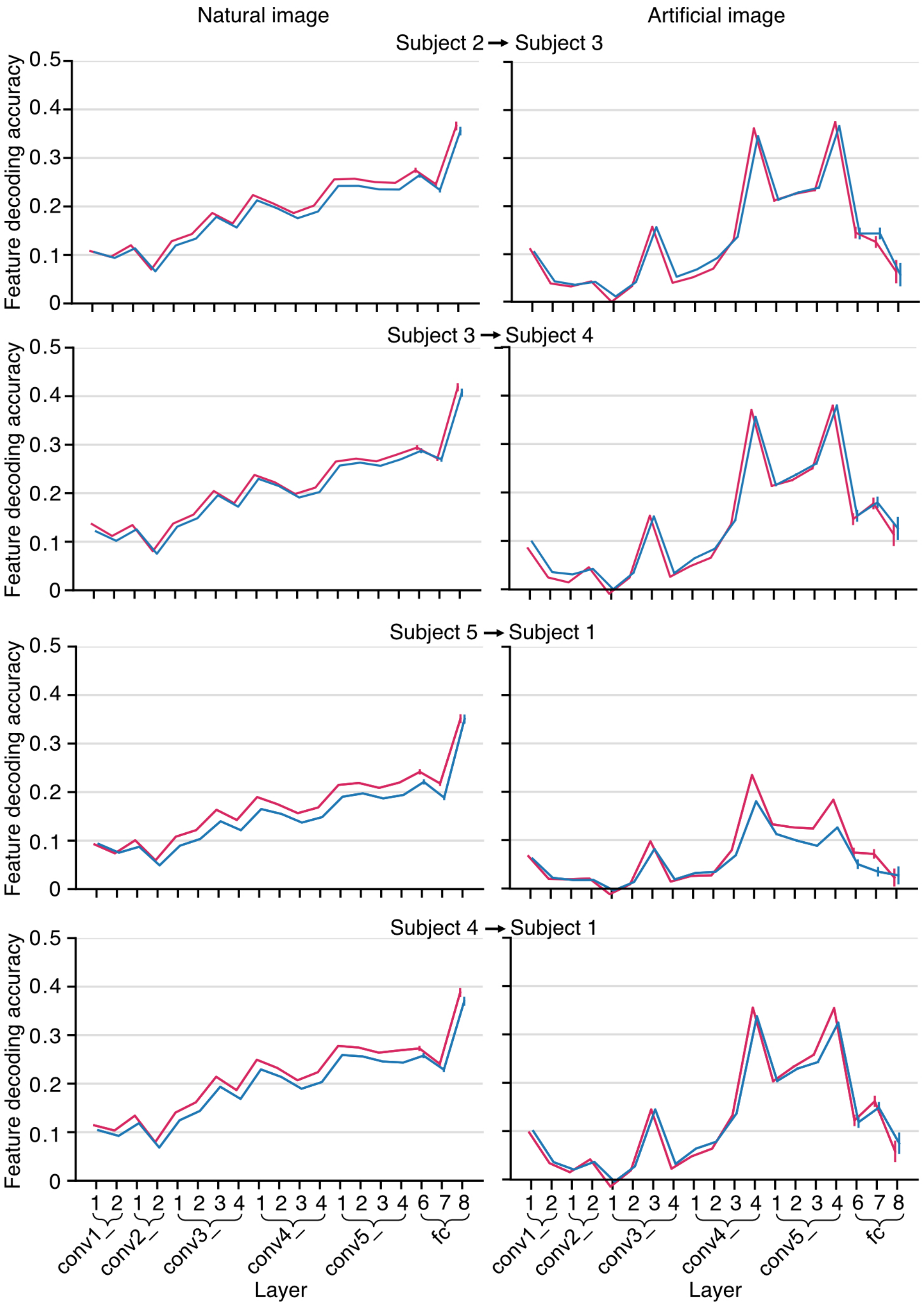

Figure 6.8: Inter-individual DNN Feature decoding accuracies of representative individual pairs with multiple- and single-subject feature decoders under limited data availability. The multiple-subject feature decoders were trained on pooled data, while the single-subject feature decoders were trained on a single subject's data. The decoding accuracies for all DNN units in each layer were used to calculate a mean decoding accuracy and its 95% confidence interval (error bars; 95% C.I. across DNN units).

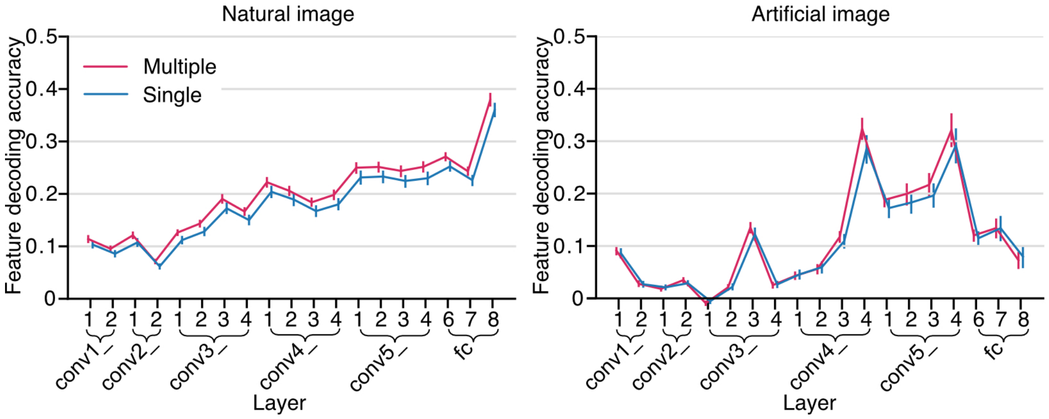

Figure 6.9: Mean DNN feature decoding accuracy obtained via multiple- and single-subject feature decoders across all individual pairs under limited data availability. The multiple-subject feature decoders were trained on pooled data, while the single-subject feature decoders were trained on a single subject's data. The accuracies were obtained from the source subjects' test dataset (error bars, 95% C.I. from 20 individual pairs).

activity patterns across individuals. Consequently, the resolution of this challenge might necessitate the application of more advanced methodologies (Li et al., 2021).

From a pragmatic standpoint, the pooling of additional training data from other subjects expands the room for hyperparameter refinement. The number of voxels encompassed within the visual cortex in this study generally varies between 10,000 and 15,000, a figure that significantly surpasses the number of samples for decoder training. As a result, strategies such as regularization or voxel selection become essential. Nonetheless, the inclusion of a greater volume of data could eliminate this necessity, potentially leading to enhanced performance outcomes.

Despite our findings indicating that pooling data by a linear method did not lead to great improvements in visual image reconstruction quality, it is still a promising direction for future fMRI research.

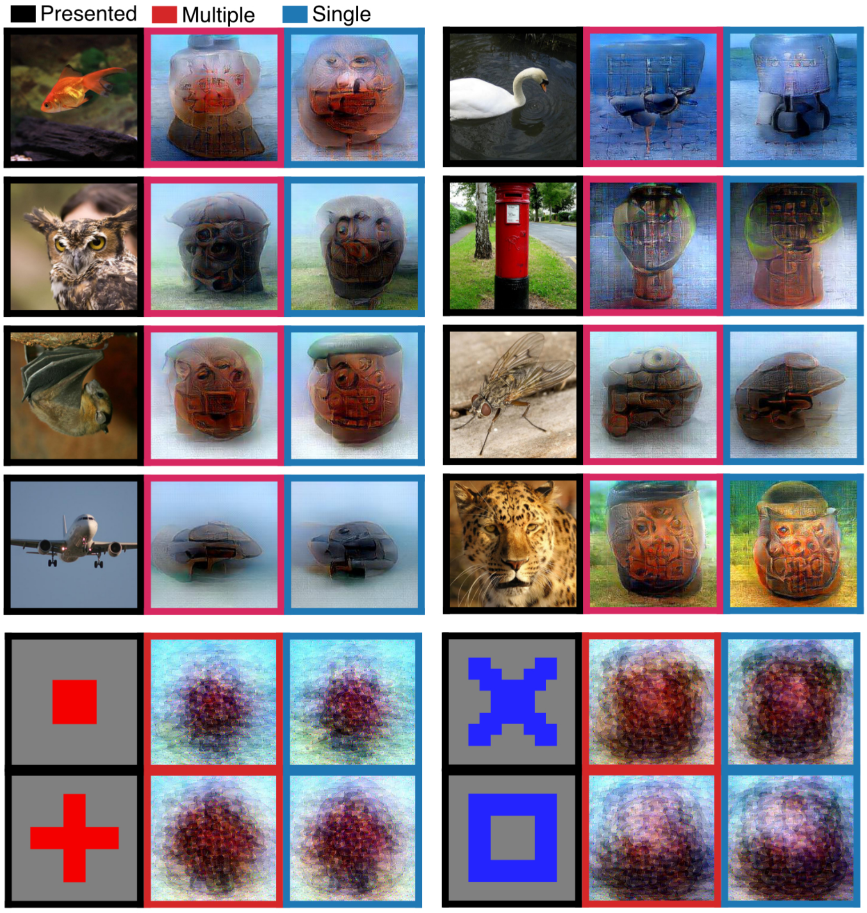

Figure 6.10: Reconstructed natural images for the multiple- and single-subject conditions under limited data availability. The reconstructions shown under the four analytical conditions for each stimulus image were all from the same source subject. The results for different stimulus images are from different source subjects.

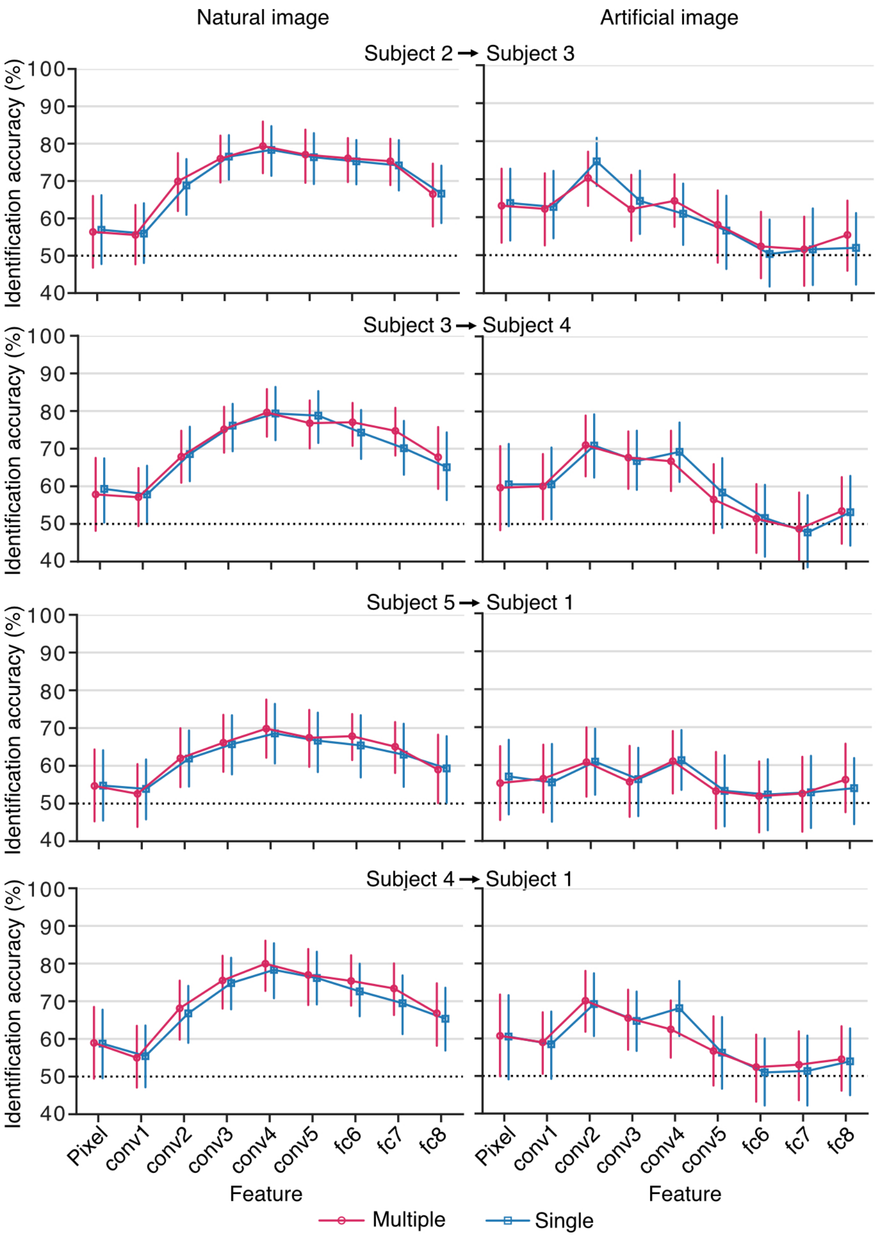

Figure 6.11: Identification analyses of representative individual pairs with multiple- and single-subject feature decoders under limited data availability. The identification accuracies for reconstructed images were used to calculate a mean identification accuracy and its 95% confidence interval. The multiple-subject feature decoders outperformed the single-subject feature decoders (dotted line, chance level = 50%).

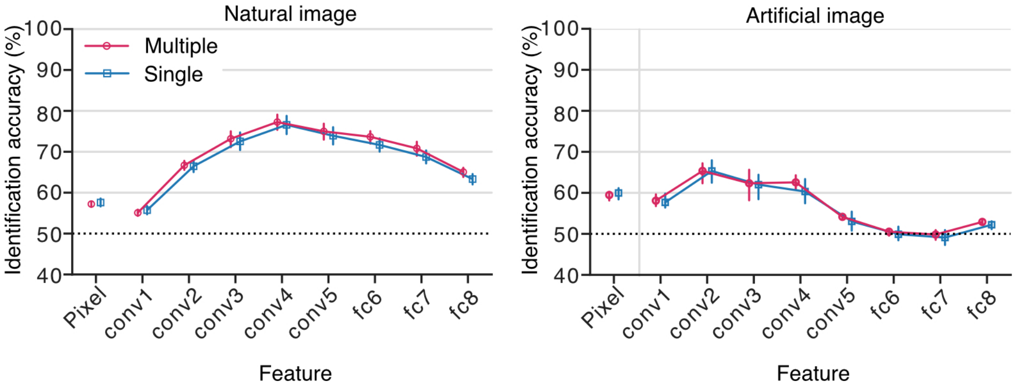

Figure 6.12: Mean identification accuracies obtained via multiple- and single-subject feature decoders across all individual pairs under limited data availability. The identification analysis was performed using the pixel values and the extracted DNN feature values of the reconstructions obtained via multiple- and single-subject feature decoders (error bars, 95% C.I. from 20 individual pairs; dotted lines, chance level = 50%).

Chapter 7

General discussion

7.1 Summary of findings and contributions

In this study, I investigated the potential of neural code converters for inter-individual visual image reconstruction. In Chapter 3, I used functional magnetic resonance imaging (fMRI) data to train the neural code converters, which were then used to predict brain activity patterns of a target subject from a source subject. The accuracy of the conversion was moderate, but it was sufficient for inter-individual visual image reconstruction. One of the key findings of the study was that the neural code converters learned the hierarchical correspondence of visual areas without imposing the cortical hierarchy constraint. This allowed for the preservation of fine-grained visual features, which are important for capturing the richness of visual perception.

In Chapter 4, the converted brain activity patterns were then decoded into hierarchical deep neural network (DNN) features to reconstruct visual images that showed recognizable shapes and colors of the objects in the presented images. The results demonstrated that the hierarchical and fine-grained DNN features can be converted across individuals while retaining sufficient encoded perceptual content to reconstruct visual images.

In Chapter 5, I compared the performance of whole visual cortex (VC) conversion versus subarea-wise conversion in inter-individual visual image reconstruction. The results showed that the whole VC conversion slightly outperformed the subarea-wise conversion with sufficient training data. However, the subarea-wise conversion performed better with minimal data. These findings suggest that the whole VC conversion, a purely data-driven approach, preserves the hierarchical structure that is explicitly assumed in the subarea-wise conversion.

In Chapter 6, an interesting finding of the study was that pooling data from multiple subjects just slightly enhanced visual image reconstruction performance. The reconstruction quality was not greatly improved, probably because of the variability of data quality and the limitation of the linear-based neural code converter model.

Overall, this study demonstrated the potential of neural code converters for inter-individual visual image reconstruction. The converters can capture the hierarchical and fine-grained visual features of the brain activity patterns and decode them into visual images.

7.2 Fine-scaled voxel mapping

The present research employed neural code converters that were trained on brain activity patterns corresponding to natural images, which are hypothesized to emulate typical visual experiences encountered in daily life. Naturalistic stimuli are presumed to encompass a comprehensive range of visual experiences, thereby furnishing the neural code converters with the capacity to transform a variety of visual representations across distinct individuals.

Indeed, the findings presented in this thesis underscore the efficacy of these neural code converters in converting the hierarchical visual representation across individuals, which are constituted by an array of visual features. However, it is noteworthy that some researchers might exhibit a more pointed interest in the commonality of particular visual features. Chapter 1 has introduced several of them, including image contrast, object identity, retinotopic organization, and semantic contents (Yamada et al., 2011 & 2015; Haxby et al., 2011; Bilenko and Gallant, 2016 and Guntupalli et al., 2016; Van Uden et al., 2018). Yet there are still many other unexplored visual features, such as visual illusion. Such features may be overlooked by DNN representations, yet they might be effectively apprehended by the neural code converters. The potential for these converters to capture such specific features underlines their versatile functionality and indicates a promising avenue for future research in visual perception studies.

Moreover, the transformation matrix of functional alignment encodes the fine-scaled mapping rule of how voxels of a subject is mapped to another (Haxby et al., 2020). To see how it works, the neural code conversion can be expressed in the form of:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m1} & w_{m2} & \cdots & w_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

where $\mathbf{W}$ is a $m \times n$ converter matrix, and $\mathbf{x}$ and $\mathbf{y}$ are brain activity patterns of the source and target subjects, with number of voxels n and m respectively. Here I omit the bias vector for simplicity. This expression can be rewrote into the form of:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = x_1 \begin{bmatrix} w_{11} \\ w_{21} \\ \vdots \\ w_{m1} \end{bmatrix} + x_2 \begin{bmatrix} w_{12} \\ w_{22} \\ \vdots \\ w_{m2} \end{bmatrix} + \cdots + x_n \begin{bmatrix} w_{1n} \\ w_{2n} \\ \vdots \\ w_{mn} \end{bmatrix}$$

or

$$\mathbf{y} = x_1\mathbf{w}_1 + x_2\mathbf{w}_2 + \cdots + x_n\mathbf{w}_n.$$

The columns vector $\mathbf{w}_i$ maps a voxel of the source subject to the target brain space, or in other words, it represent the fine-scaled topography of the voxel in the target brain space.

With the fine-scaled mapping of voxels between subjects, neural code converters could be employed to scrutinize the conservation of specific local visual representations, rather than converting the entirety of the hierarchical visual representation. To illustrate, consider a unique type of stimulus that incites activation within a localized region of the visual cortex, identifiable by voxel activity. The neural code converter can be tactically utilized to investigate whether analogous voxels within a similar region of the visual cortex of a different subject are likewise activated. If such a parallel activation is observed, it serves as an indication that the visual features encoded within this specific region of the visual cortex may indeed be universally represented across different individuals. Such focused use of neural code converters can therefore shed light on the inter-individual consistency of local visual representation.

Moreover, the fine-scaled voxel mapping approach represents a valuable method for estimating known topographies (Haxby et al., 2020), such as the functional V1 visual area. This technique is particularly advantageous in situations where fMRI data from a retinotopic experiment is unavailable to delineate the functional visual area in the target subject. Instead, by leveraging the functional visual area data obtained from a source subject, it becomes feasible to infer and estimate corresponding regions within the target brain space.

7.3 The Black-Box of DNN

The work presented in this thesis involves the utilization of DNNs for the analysis of hierarchical visual representation. As underscored in Chapter 1, DNNs present a significant advantage in terms of their capacity for automatic feature extraction, obviating the need for hand-crafted features. However, it is noteworthy that the features discerned by a DNN are considerably influenced by the training dataset and the specific architecture of the network. Indeed, a bias towards particular features has been observed in DNNs, with ImageNet-trained DNNs exhibiting a known tendency towards textures (Geirhos et al., 2022). Such biases could potentially lead to an inadvertent neglect of other salient features.

In addition, it is important to acknowledge that the VGG19 DNN model, an integral part of this study, is a purely feedforward network without any feedback connections. Contrastingly, it is a well-established fact that the visual cortex harbors numerous feedback connections from higher visual areas to lower visual areas, facilitating a top-down modulation. Consequently, the VGG19 DNN model may fall short in effectively modeling such top-down modulation effects, which may be fundamental to comprehending the visual system. However, an increase in network complexity and superior performance in object recognition tasks does not necessarily assure a more brain-like representation, as evidenced by Nonaka et al. (2021). This finding introduces a level of uncertainty regarding the suitability of non-brain-like DNNs in modeling the intricacies of the brain.

While the VGG19 DNN model undeniably encodes a wealth of information pertaining to image features, it must be conceded that we lack total control over what DNNs will learn. This thesis does not posit the superiority of any specific DNN; each network can serve as an effective tool for brain modeling as long as its DNN features correlate with neural data. The point of emphasis, rather, is that there may be certain features overlooked by DNNs, which are nonetheless of interest to researchers. To build on the findings of the present study, a promising future direction may involve the use of alternative DNNs that are capable of learning a richer set of features compared to the VGG19 DNN model.

7.4 Potential of neural code converters in visual restoration

The research delineated in this thesis entails the implementation of functional alignment on brain activity patterns within the visual cortex. This presented approach necessitates the training data while subjects are involved in visual tasks. However, for visually impaired individuals, this data acquired through visual tasks may not be viable. Nevertheless, evidence indicates that visual cortical activity is observed when blind individuals engage in nonvisual tasks such as hearing words or Braille reading (Burton, 2003). This suggests a potential repurposing of the visual cortex for language functions, implying that its activity may signify something distinct from that of sighted individuals.

However, the potential for residual visual functions in the visual cortex of late-blind individuals remains a compelling proposition. It raises the possibility of converting brain activity patterns from visually impaired subjects to sighted subjects' brain spaces. This scenario invites a slight relaxation of the functional alignment assumption: that all subjects are viewing a predetermined sequence of images or a movie. Instead, an experiment could be designed where a sighted individual views images or a movie while a blind individual performs Braille reading with content matching the visual stimuli. A neural code converter could be trained on the brain data from both the blind (source) and sighted (target) subjects. The brain activity patterns of the blind individual could then be converted into the target brain space and analyzed through the decoding of DNN features and visual image reconstruction, as delineated in this thesis.

While it is plausible to anticipate reduced performance or even failure from this conversion, even a modestly successful conversion would hold considerable significance. Such a result would imply the presence of shared visual features between sighted and blind individuals. This finding could be particularly beneficial in the realm of visual prosthetics, which aim to restore vision via brain stimulation (Lewis et al., 2015). The fine-scale voxel mapping rule, as described in section 7.2, might offer an initial guide regarding the appropriate placement of electrodes to optimally replicate normal vision through visual restoration.

In essence, this innovative approach may not only further our understanding of the human visual system, but also pave the way for transformative solutions in the field of visual restoration, thus creating a brighter future for those living with visual impairments.

7.5 Saving data collection time

The process of training a decoding model necessitates extensive human brain data sampling to encompass the comprehensive variation of stimuli. In the context of fMRI decoding research, this amounts to several tens of hours of scanning time - a duration that is typically prohibitive due to physical constraints and financial considerations. However, as inter-individual analyses have demonstrated the feasibility of employing decoding models from other subjects when brain activity patterns are functionally aligned, there emerges a compelling motivation to explore inter-individual visual image reconstruction from a practical perspective.

As illustrated in Chapter 5, the neural code converter, trained with a modest 300 samples (approximating one hour of data collection time), was capable of generating discernible reconstructed images in the inter-individual visual image reconstruction analysis. This implies that a minimal one or two hours of data collection can suffice to facilitate visual image reconstruction, assuming the availability of a well trained decoding model from other subjects.

This finding is of considerable practical significance. For instance, if artists desire to use visual image reconstruction techniques to manifest their internal visions, the neural code converter presents a valuable tool. Looking ahead, brain-machine interfaces may become increasingly prevalent, and the calibration process for such interfaces with individual users could be both time-consuming and impractical. A more viable approach may involve calibrating the user's brain activity patterns with those of other users who have undergone extensive training with the brain-machine interface. Consequently, the progress in functional alignment methods could potentially propel the adoption and effectiveness of brain-machine interfaces.

7.6 Inter-individual neural code conversion without paired stimuli

The technique of neural code conversion relies on an assumption - it requires training data where all subjects are exposed to an identical sequence of stimuli. This requirement inevitably imposes limitations on its generalizability and real-world applications. Earlier research has proposed an approach using fMRI data with partially unpaired stimuli (Li et al., 2020), which showed encouraging results when up to half of the data involved unpaired stimuli. However, this method is not applicable when the data are completely unpaired.

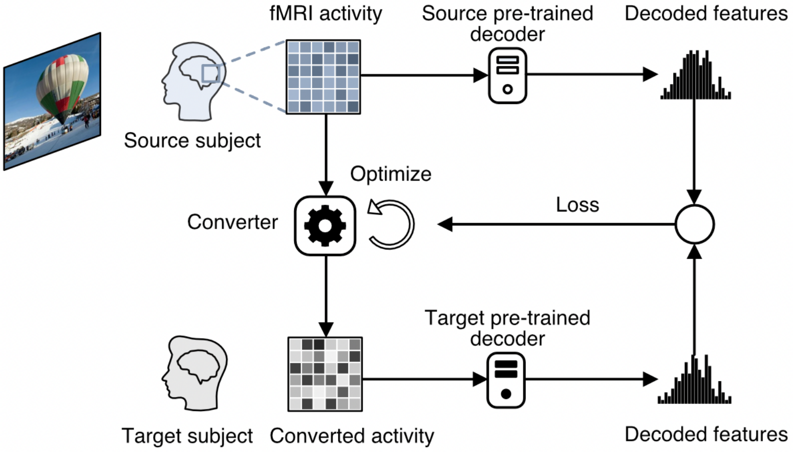

Figure 7.1: Schematic of the training process for the neural code converter the DNN feature space. The optimization goal of the converter is to ensure the similarity between DNN features decoded from the source brain activity patterns and those decoded from the converted brain activity patterns. Image courtesy of Haibao Wang, used with permission.

In an ongoing collaborative work with my colleagues, we are developing an innovative neural code conversion method that does not require paired stimuli, drawing upon the concept of DNN feature decoding. In the training phase, a pair of source and target subjects is identified, along with their respective DNN feature decoders. The brain activity patterns of the source subject are decoded into DNN features using the source subject's decoder, while simultaneously being converted into the target brain space. These converted brain activity patterns are subsequently decoded into DNN features using the target subject's decoder. The converter is trained to minimize the loss function between these two sets of decoded DNN features (Figure 7.1). Notably, since the converter is trained in the DNN feature space rather than voxel space, paired-stimuli fMRI data are not necessary. Additionally, the converter employs a DNN rather than linear regression, thus leveraging the capacity of DNNs to handle nonlinearity.

Preliminary results from this method are promising in the context of inter-individual visual image reconstruction. However, perhaps the most significant aspect of this technique is that it does not require paired-stimuli fMRI data. This feature has the potential to facilitate the pooling of data from entirely different fMRI datasets, thereby enhancing its generalizability and enabling the pooling of data from diverse sources.

7.7 Other future directions

The present study provides significant insights into the hierarchical visual information and their individual differences, but there are still several relevant directions that could not be addressed within the scope of this thesis. I summarized them in the following subsections.

7.7.1 Other visual features

This section enumerates two visual features that are potentially of significant interest in the context of visual image reconstruction. The extent to which the neural representation of these features is shared across individuals is a topic that has not yet been thoroughly explored in the existing literature. Nonetheless, delving into this area of study may provide invaluable insights that could enhance our understanding of the human visual system.

Color

The extent to which visual attributes, such as color, are universally represented in the human brain is a compelling research question. Amongst the spectrum of visual features, color information is particularly relevant to visual image reconstruction. In contemporary approaches to visual image reconstruction, methodologies have evolved to a degree that allows the generation of colored images from brain activity. However, these techniques appear to demonstrate a biased proficiency towards the reconstruction of reddish hues, while other colors in the spectrum are often less reliably reproduced. This raises intriguing questions about the neural encoding and representation of different colors, and why certain color information might be more easily reconstructed than others.

The successful implementation of inter-individual visual image reconstruction has been demonstrated in this thesis, providing a promising avenue for further investigation into shared and unique aspects of visual perception. Nevertheless, the question as to whether the neural representations of color information are universally shared across individuals remains open.

Visual illusion

Visual illusions, phenomena that elicit perceptual experiences differing from physical reality, provide intriguing insights into the complex mechanisms of visual processing. While our perception is often reliable, illusions illustrate how it can be systematically misled, emphasizing the interpretative nature of vision. In essence, visual illusions manifest as a discrepancy between what we see and the objective attributes of the stimuli, demonstrating the intricate interplay of sensory processing, cognitive interpretation, and prior experiences. Illusions highlight the constructive nature of perception and illustrate how our brains make educated guesses about the world based on limited sensory information. It is suggested that these illusions result from the brain's attempt to interpret ambiguous or incomplete sensory information based on prior knowledge and expectations.

Investigating how these illusory percepts are represented in the brain can provide valuable insights into the mechanisms of visual processing. It is unclear whether the underlying representations of such illusions are shared across individuals. Cheng et al. (2023) successfully reconstructed visual illusions through visual image reconstruction techniques (Figure 7.2). It is compelling to consider applying the inter-individual framework presented in this thesis to explore whether neural representations of visual illusions can be converted across individuals, and even more fascinating, across individuals who lack the ability to perceive visual illusions.

7.7.2 Decoding task optimized neural code converters

The neural code converters developed in this thesis have been designed to be task-independent, aiming to produce brain activity patterns that accurately resemble those in the target brain space. While these converters have wide applicability in multiple decoding models for inter-individual brain decoding research, they may not be specifically fine-tuned for a given decoding task, which could result in less-than-optimal performance.

In visual image reconstruction, not all voxels from the whole visual cortex are equally important for DNN feature decoding and reconstruction. Careful voxel selection can potentially address this issue. For instance, one approach is to identify a subset of voxels that have higher weights in the decoding model for each DNN layer and train a neural code converter specifically for that DNN layer.

The methodology presented in section 7.6 represents another technique intended to optimize the performance of DNN feature decoding. It focuses on training the converter within the DNN feature space, rather than the voxel space. Given that the decoders carry vital information about the voxels crucial for DNN feature decoding, this strategy allows the integration of such knowledge into the converter. Ultimately, these strategies underscore the potential for task-specific enhancements to further refine the applicability and effectiveness of the neural code converters.

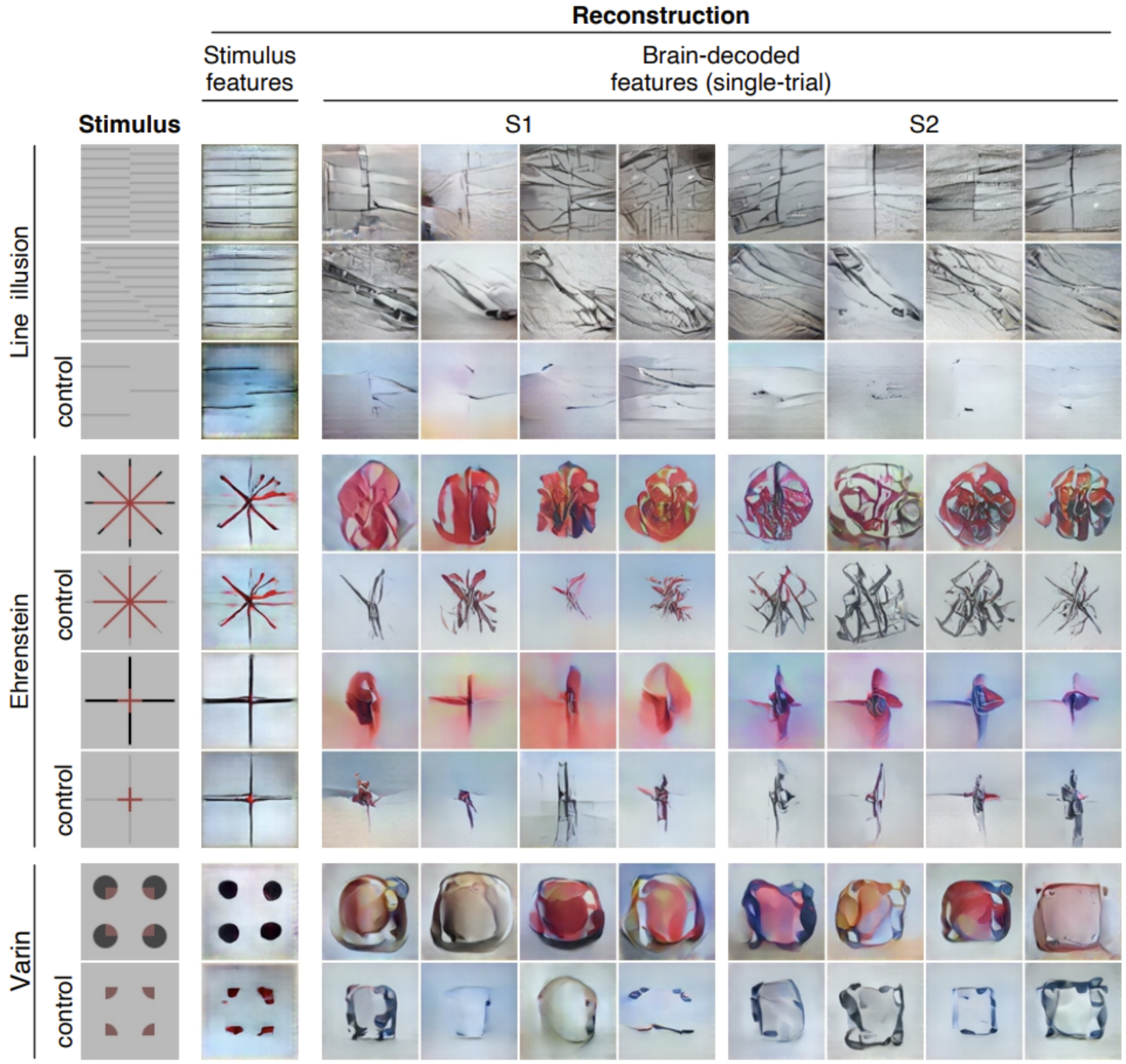

Figure 7.2: Reconstructions of illusory images from brain activity patterns. The figure showed three kinds of illusory images, including line illusion, Ehrenstein illusion, and Varin illusion. The reconstructions from two subjects S1 and S2 were shown together with the reconstructions from the stimulus DNN features of the control images. Reprinted from *Biorxiv*, 2023, Cheng et al., Reconstructing visual illusory experiences from human brain activity, Page No. 3, licensed under CC BY 4.0.

7.7.3 Pooling open data from other laboratories

In Chapter 6, I explored the possibility of pooling data from other subjects to enhance visual image reconstruction. The conclusion reached is that the potential for this approach to be effective remains neutral, but it is still a promising direction for future research if the amount of pooled data can be scaled up to several orders. Because of the cost of data collection, the amount of data that one laboratory can collect can be limited. Pooling data from other laboratories could potentially provide a solution to this problem by increasing the amount of data available for analysis. However, this approach is not without its challenges. The main challenge is that the data collected from different laboratories may not be directly comparable due to differences in experimental design, machines used for data collection, and data preprocessing techniques.

These differences in experimental design and data collection can make it difficult to achieve functional alignment across datasets. Functional alignment is the process of ensuring that the data collected from different sources is comparable and can be combined in a meaningful way. Without functional alignment, combining data from different sources may lead to unreliable or meaningless results.

Despite these challenges, pooling open data from other laboratories remains a promising direction for future research. In order to achieve success with this approach, researchers will need to develop new techniques for functional alignment (for example, the approach introduced in section 7.6) and standardize experimental designs and data collection methods across laboratories. This will require collaboration across different research groups and a willingness to share data openly and transparently.

7.7.4 Geometric-based alignment

Although fMRI responses are measured in 3D volumes, it's important to note that neural activity actually takes place on the cortical surface. Consequently, two proximate voxels within 3D space may not necessarily be adjacent on a flattened cortical surface. While many functional alignment techniques adhere to the fundamental principles without accounting for cortical structure, these methods often overlook the spatial organization of the cortical surface, which could furnish additional data to improve alignment precision.

A promising approach to integrating information about the cortical surface structure is through surface-based convolution networks, a component of geometric deep learning (for a comprehensive introduction, see Bronstein et al., 2021). These networks can perform convolution operations on vertices on a cortical surface, allowing the integration of information about the cortical surface across the stacked convolutional layers. Moreover, other types of convolution could be employed with more intricate connectivity structures, such as those observed in fMRI functional or structural connectivity. For example, Ribeiro et al. (2021) used geometric CNN to predict functional retinotopic organization based on anatomical attributes, including curvature, myelin values, and connectivity among vertices.

Geometric deep learning offers advantages because it integrates both anatomical structure and connectivity data, with each vertex on the surface carrying multiple attributes. This presents the potential to simultaneously perform anatomical and functional alignment.

7.7.5 Domain adaptation

Domain adaptation serves as a strategy within the broader framework of transfer learning, enabling the application of a model trained within one source domain to a distinct target domain, despite disparate statistics. This technique operates under the presumption that despite variances in statistics, the source and target domains possess a degree of commonality in their features. By capitalizing on this shared attribute, domain adaptation techniques facilitate the transferability of models across domains.

A classic illustration of domain adaptation could be the application of a GAN trained on real human faces to the domain of cartoon faces, given the shared features such as eyes and noses. In this scenario, the aim is to leverage the knowledge acquired from the source domain (realistic faces) and extend it to the target domain (cartoon faces), thereby obviating the need to train an entirely new generator. This technique acquires particular relevance when data availability is limited in the target domain (Noguchi and Harada, 2019; Ojha et al., 2021).

Neuroscientific research has harnessed the potential of transfer learning and domain adaptation (Koyamada et al., 2015; Valverde et al., 2021). However, the application of these techniques to inter-individual studies remains relatively uncharted territory and presents a more practical perspective than traditional functional alignment. As discussed earlier, while there are statistical differences in fMRI activity patterns across subjects, they also contain common features. It is plausible, therefore, to adapt a model trained on one individual's data to the data of other individuals.

For instance, a generator-based visual image reconstruction model (Shen et al., 2019b) that takes fMRI responses as inputs could potentially be adapted to the fMRI responses of other subjects, requiring less data for the adaptation. In this scenario, the fMRI responses of subjects might not necessarily be aligned; instead, the model is tailored for a specific subject. This application of domain adaptation could hold practical implications for the implementation of brain-machine interfaces, offering a potentially efficient solution to individual differences in neural activity.

7.8 Conclusions

The presented thesis makes a contribution to the field of neuroscience by demonstrating the possibility of inter-individual visual image reconstruction through functional alignment of fMRI data. The findings suggest that neural code converters can predict the brain activity patterns of a target subject from a source subject with moderate conversion accuracies. Additionally, the converted brain activity patterns can be decoded into hierarchical deep neural network features to reconstruct visual images, preserving the information of hierarchical fine-grained visual features.

This thesis underscores the potential of neural code converters to delve deeper into the shared properties of localized visual features beyond just the hierarchical structure, as facilitated by fine-scaled voxel mapping. When combined with DNNs, inter-individual analysis could enable the investigation of lesser-understood visual features. Crucially, identifying potential common visual features between sighted and visually impaired individuals can provide pivotal insights for advancing visual restoration efforts for the blind.

The thesis also highlights the potential advantages of the functional alignment approach in reducing the amount of data required for model training, making it an efficient and feasible approach for reconstructing visual images without the need for subject-specific models. This approach could have significant implications for the development of brain-machine interfaces and computer interfaces that communicate with our internal world. The ability to visualize the perceived stimulus through inter-individual visual image reconstruction could help to bridge the gap between human cognition and artificial intelligence, paving the way for more efficient and natural human-machine interactions.

In summary, the presented thesis opens up new avenues for future research in inter-individual visual image reconstruction and functional alignment of fMRI data. The findings have the potential to advance the development of brain-machine interfaces, ultimately leading to a better understanding of the human brain and its intricate relationship with technology.

References

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., & Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, *8*. https://doi.org/10.3389/fninf.2014.00014

Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A*, *2*, 284–299. https://doi.org/10.1364/JOSAA.2.000284

Anzai, A., Peng, X., & Van Essen, D. C. (2007). Neurons in monkey visual area v2 encode combinations of orientations. *Nature Neuroscience*, *10*, 1313–1321. https://doi.org/10.1038/nn1975

Avants, B. B., Epstein, C. L., Grossman, M., & Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, *12*, 26–41. https://doi.org/10.1016/j.media.2007.06.004

Bazeille, T., Richard, H., Janati, H., & Thirion, B. (2019). Local optimal transport for functional brain template estimation. In *Information processing in medical imaging* (pp. 237–248). Springer International Publishing. https://doi.org/10.1007/978-3-030-20351-1_18

Bazeille, T., DuPre, E., Richard, H., Poline, J.-B., & Thirion, B. (2021). An empirical evaluation of functional alignment using inter-subject decoding. *NeuroImage*, *245*, 118683. https://doi.org/10.1016/j.neuroimage.2021.118683

Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage*, *37*, 90–101. https://doi.org/10.1016/j.neuroimage.2007.04.042

Belliveau, J. W., Kennedy, D. N., McKinstry, R. C., Buchbinder, B. R., Weisskoff, R. M., Cohen, M. S., Vevea, J. M., Brady, T. J., & Rosen, B. R. (1991). Functional mapping of the human visual cortex by magnetic resonance imaging. *Science*, *254*, 716–719. https://doi.org/10.1126/science.1948051

Bilenko, N. Y., & Gallant, J. L. (2016). Pyrcca: Regularized kernel canonical correlation analysis in python and its applications to neuroimaging. *Frontiers in Neuroinformatics*, *10*. https://doi.org/10.3389/fninf.2016.00049

Blasdel, G. G., & Salama, G. (1986). Voltage-sensitive dyes reveal a modular organization in monkey striate cortex. *Nature*, *321*, 579–585. https://doi.org/10.1038/321579a0

Blumensath, T., Jbabdi, S., Glasser, M. F., Van Essen, D. C., Ugurbil, K., Behrens, T. E. J., & Smith, S. M. (2013). Spatially constrained hierarchical parcellation of the brain with resting-state fMRI. *NeuroImage*, *76*, 313–324. https://doi.org/10.1016/j.neuroimage.2013.03.024

Brewer, A. A., & Barton, B. (2012). Visual field map organization in human visual cortex. In *Visual cortex*. IntechOpen. https://doi.org/10.5772/51914

Bronstein, M. M., Bruna, J., Cohen, T., & Veličković, P. (2021, May 2). Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. https://doi.org/10.48550/arXiv.2104.13478

Burton, H. (2003). Visual cortex activity in early and late blind people. *Journal of Neuroscience*, *23*, 4005–4011. https://doi.org/10.1523/JNEUROSCI.23-10-04005.2003

Busch, E. L., Slipski, L., Feilong, M., Guntupalli, J. S., Castello, M. V. d. O., Huckins, J. F., Nastase, S. A., Gobbini, M. I., Wager, T. D., & Haxby, J. V. (2021). Hybrid hyperalignment: A single high-dimensional model of shared information embedded in cortical patterns of response and functional connectivity. *NeuroImage*, *233*, 117975. https://doi.org/10.1016/j.neuroimage.2021.117975

Chen, P.-H. (, Chen, J., Yeshurun, Y., Hasson, U., Haxby, J., & Ramadge, P. J. (2015). A reduced-dimension fMRI shared response model. In *Advances in neural information processing systems* (Vol. 28). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2015/file/b3967a0e938dc2a6340e258630febd5a-Paper.pdf

Cheng, F., Horikawa, T., Majima, K., Tanaka, M., Abdelhack, M., Aoki, S. C., Hirano, J., & Kamitani, Y. (2023, June 15). Reconstructing visual illusory experiences from human brain activity. https://doi.org/10.1101/2023.06.15.545037

Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, *29*, 162–173. https://doi.org/10.1006/cbmr.1996.0014

Craddock, R. C., James, G., Holtzheimer III, P. E., Hu, X. P., & Mayberg, H. S. (2012). A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Human Brain Mapping*, *33*, 1914–1928. https://doi.org/10.1002/hbm.21333

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis: I. segmentation and surface reconstruction. *NeuroImage*, *9*, 179–194. https://doi.org/10.1006/nimg.1998.0395

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Desimone, R., Albright, T. D., Gross, C. G., & Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. *Journal of Neuroscience*, *4*, 2051–2062. https://doi.org/10.1523/JNEUROSCI.04-08-02051.1984

Dosovitskiy, A., & Brox, T. (2016). Generating images with perceptual similarity metrics based on deep networks. *Advances in Neural Information Processing Systems*, *29*. Retrieved June 7, 2023, from https://proceedings.neurips.cc/paper/2016/hash/371bce7dc83817b7893bcdeed13799b5-Abstract.html

Engel, S. A., Glover, G. H., & Wandell, B. A. (1997). Retinotopic organization in human visual cortex and the spatial precision of functional MRI. *Cerebral Cortex*, *7*, 181–192. https://doi.org/10.1093/cercor/7.2.181

Engel, S. A., Rumelhart, D. E., Wandell, B. A., Lee, A. T., Glover, G. H., Chichilnisky, E.-J., & Shadlen, M. N. (1994). fMRI of human visual cortex. *Nature*, *369*, 525–525. https://doi.org/10.1038/369525a0

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

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., & Gorgolewski, K. J. (2019). fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*, 111–116. https://doi.org/10.1038/s41592-018-0235-4

Fischl, B., Rajendran, N., Busa, E., Augustinack, J., Hinds, O., Yeo, B. T., Mohlberg, H., Amunts, K., & Zilles, K. (2008). Cortical folding patterns and predicting cytoarchitecture. *Cerebral Cortex*, *18*, 1973–1980. https://doi.org/10.1093/cercor/bhm225

Fischl, B., Sereno, M. I., & Dale, A. M. (1999). Cortical surface-based analysis: II: Inflation, flattening, and a surface-based coordinate system. *NeuroImage*, *9*, 195–207. https://doi.org/10.1006/nimg.1998.0396

Fonov, V. S., Evans, A. C., McKinstry, R. C., Almli, C. R., & Collins, D. L. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, *Supplement 1*, S102. https://doi.org/10.1016/S1053-8119(09)70884-5

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2414–2423. https://doi.org/10.1109/CVPR.2016.265

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2022, November 9). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. https://doi.org/10.48550/arXiv.1811.12231

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (Vol. 27). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf

Gordon, E. M., Chauvin, R. J., Van, A. N., Rajesh, A., Nielsen, A., Newbold, D. J., Lynch, C. J., Seider, N. A., Krimmel, S. R., Scheidter, K. M., Monk, J., Miller, R. L., Metoki, A., Montez, D. F., Zheng, A., Elbau, I., Madison, T., Nishino, T., Myers, M. J., … Dosenbach, N. U. F. (2023). A somato-cognitive action network alternates with effector regions in motor cortex. *Nature*, *617*, 351–359. https://doi.org/10.1038/s41586-023-05964-2

Gorgolewski, K., Burns, C., Madison, C., Clark, D., Halchenko, Y., Waskom, M., & Ghosh, S. (2011). Nipype: A flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in Neuroinformatics*, *5*. https://doi.org/10.3389/fninf.2011.00013

Gorgolewski, K. J., Esteban, O., Ellis, D. G., Notter, M. P., Ziegler, E., Johnson, H., Hamalainen, C., Yvernault, B., Burns, C., Manhães-Savio, A., Jarecka, D., Markiewicz, C. J., Salo, T., Clark, D., Waskom, M., Wong, J., Modat, M., Dewey, B. E., Clark, M. G., … Ghosh, S. (2017, May 21). *Nipype: A flexible, lightweight and extensible neuroimaging data processing framework in python. 0.13.1*. Zenodo. https://doi.org/10.5281/zenodo.581704

Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, *48*, 63–72. https://doi.org/10.1016/j.neuroimage.2009.06.060

Grodd, W., Hülsmann, E., Lotze, M., Wildgruber, D., & Erb, M. (2001). Sensorimotor mapping of the human cerebellum: fMRI evidence of somatotopic organization. *Human Brain Mapping*, *13*, 55–73. https://doi.org/10.1002/hbm.1025

Güçlü, U., & van Gerven, M. A. J. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *The Journal of Neuroscience*, *35*, 10005–10014. https://doi.org/10.1523/JNEUROSCI.5023-14.2015

Güçlü, U., & van Gerven, M. A. J. (2017). Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. *NeuroImage*, *145*, 329–336. https://doi.org/10.1016/j.neuroimage.2015.12.036

Güçlütürk, Y., Güçlü, U., Seeliger, K., Bosch, S., van Lier, R., & van Gerven, M. A. J. (2017). Reconstructing perceived faces from brain activations with deep adversarial neural decoding. In *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/efdf562ce2fb0ad460fd8e9d33e57f57-Paper.pdf

Guntupalli, J. S., Hanke, M., Halchenko, Y. O., Connolly, A. C., Ramadge, P. J., & Haxby, J. V. (2016). A model of representational spaces in human cortex. *Cerebral Cortex*, *26*, 2919–2934. https://doi.org/10.1093/cercor/bhw068

Han, K., Wen, H., Shi, J., Lu, K.-H., Zhang, Y., Fu, D., & Liu, Z. (2019). Variational autoencoder: An unsupervised model for encoding and decoding fMRI activity in visual cortex. *NeuroImage*, *198*, 125–136. https://doi.org/10.1016/j.neuroimage.2019.05.039

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *Science*, *303*, 1634–1640. https://doi.org/10.1126/science.1089506

Haxby, J. V., Guntupalli, J. S., Connolly, A. C., Halchenko, Y. O., Conroy, B. R., Gobbini, M. I., Hanke, M., & Ramadge, P. J. (2011). A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, *72*, 404–416. https://doi.org/10.1016/j.neuron.2011.08.026

Haxby, J. V., Guntupalli, J. S., Nastase, S. A., & Feilong, M. (2020). Hyperalignment: Modeling shared information encoded in idiosyncratic cortical topographies. *eLife*, *9*, e56601. https://doi.org/10.7554/eLife.56601

Hegdé, J., & Van Essen, D. C. (2000). Selectivity for complex shapes in primate visual area v2. *The Journal of Neuroscience*, *20*, RC61–RC61. https://doi.org/10.1523/JNEUROSCI.20-05-j0001.2000

Ho, J. K., Horikawa, T., Majima, K., Cheng, F., & Kamitani, Y. (2023). Inter-individual deep image reconstruction via hierarchical neural code conversion. *NeuroImage*, *271*, 120007. https://doi.org/10.1016/j.neuroimage.2023.120007

Horikawa, T., & Kamitani, Y. (2022). Attention modulates neural representation to render reconstructions according to subjective appearance. *Communications Biology*, *5*, 1–12. https://doi.org/10.1038/s42003-021-02975-5

Horikawa, T., & Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, *8*, 15037. https://doi.org/10.1038/ncomms15037

Hsu, A., Borst, A., & Theunissen, F. E. (2004). Quantifying variability in neural responses and its application for the validation of model predictions. *Network: Computation in Neural Systems*, *15*, 91–109. https://doi.org/10.1088/0954-898X_15_2_002

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, *148*, 574–591. https://doi.org/10.1113/jphysiol.1959.sp006308

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*, 106–154. https://doi.org/10.1113/jphysiol.1962.sp006837

Hubel, D. H., & Wiesel, T. N. (1965). Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *Journal of Neurophysiology*, *28*, 229–289. https://doi.org/10.1152/jn.1965.28.2.229

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*, 453–458. https://doi.org/10.1038/nature17637

Ince, R. A. A., Kay, J. W., & Schyns, P. G. (2022). Within-participant statistics for cognitive science. *Trends in Cognitive Sciences*, *26*, 626–630. https://doi.org/10.1016/j.tics.2022.05.008

Ito, M., & Komatsu, H. (2004). Representation of angles embedded within contour stimuli in area v2 of macaque monkeys. *Journal of Neuroscience*, *24*, 3313–3324. https://doi.org/10.1523/JNEUROSCI.4364-03.2004

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, *17*, 825–841. https://doi.org/10.1006/nimg.2002.1132

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., & Darrell, T. (2014, June 20). Caffe: Convolutional architecture for fast feature embedding. https://doi.org/10.48550/arXiv.1408.5093

Julesz, B. (1981). Textons, the elements of texture perception, and their interactions. *Nature*, *290*, 91–97. https://doi.org/10.1038/290091a0

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience*, *17*, 4302–4311. https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997

Kastner, S., De Weerd, P., & Ungerleider, L. G. (2000). Texture segregation in the human visual cortex: A functional MRI study. *Journal of Neurophysiology*, *83*, 2453–2457. https://doi.org/10.1152/jn.2000.83.4.2453

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, *452*, 352–355. https://doi.org/10.1038/nature06713

Klein, A., Ghosh, S. S., Bao, F. S., Giard, J., Häme, Y., Stavsky, E., Lee, N., Rossa, B., Reuter, M., Neto, E. C., & Keshavan, A. (2017). Mindboggling morphometry of human brains. *PLOS Computational Biology*, *13*, e1005350. https://doi.org/10.1371/journal.pcbi.1005350

Komatsu, H., Ideura, Y., Kaji, S., & Yamane, S. (1992). Color selectivity of neurons in the inferior temporal cortex of the awake macaque monkey. *Journal of Neuroscience*, *12*, 408–424. https://doi.org/10.1523/JNEUROSCI.12-02-00408.1992

Kourtzi, Z., & Kanwisher, N. (2000). Cortical regions involved in perceiving object shape. *The Journal of Neuroscience*, *20*, 3310–3318. https://doi.org/10.1523/JNEUROSCI.20-09-03310.2000

Koyamada, S., Shikauchi, Y., Nakae, K., Koyama, M., & Ishii, S. (2015, January 31). *Deep learning of fMRI big data: A novel approach to subject-transfer decoding* [arXiv.org]. Retrieved June 24, 2023, from https://arxiv.org/abs/1502.00093v1

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, *103*, 3863–3868. https://doi.org/10.1073/pnas.0600244103

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (Vol. 25). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

Laumann, T. O., Gordon, E. M., Adeyemo, B., Snyder, A. Z., Joo, S. J., Chen, M.-Y., Gilmore, A. W., McDermott, K. B., Nelson, S. M., Dosenbach, N. U. F., Schlaggar, B. L., Mumford, J. A., Poldrack, R. A., & Petersen, S. E. (2015). Functional system and areal organization of a highly sampled individual human brain. *Neuron*, *87*, 657–670. https://doi.org/10.1016/j.neuron.2015.06.037

Le, Q., Ngiam, J., Coates, A., Lahiri, A., Prochnow, B., & Ng, A. (2011, June). On optimization methods for deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)* (pp. 265–272). ACM.

Lescroart, M. D., & Gallant, J. L. (2019). Human scene-selective areas represent 3d configurations of surfaces. *Neuron*, *101*, 178–192.e7. https://doi.org/10.1016/j.neuron.2018.11.004

Lewis, P. M., Ackland, H. M., Lowery, A. J., & Rosenfeld, J. V. (2015). Restoration of vision in blind individuals using bionic devices: A review with a focus on cortical visual prostheses. *Brain Research*, *1595*, 51–73. https://doi.org/10.1016/j.brainres.2014.11.020

Li, D., Du, C., Wang, S., Wang, H., & He, H. (2021). Multi-subject data augmentation for target subject semantic decoding with deep multi-view adversarial learning. *Information Sciences*, *547*, 1025–1044. https://doi.org/10.1016/j.ins.2020.09.012

Li, W., Liu, M., Chen, F., & Zhang, D. (2020). Graph-based decoding model for functional alignment of unaligned fMRI data. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*, 2653–2660. https://doi.org/10.1609/aaai.v34i03.5650

Liu, D. C., & Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, *45*, 503–528. https://doi.org/10.1007/BF01589116

Livingstone, M. S., & Hubel, D. H. (1984). Anatomy and physiology of a color system in the primate visual cortex. *Journal of Neuroscience*, *4*, 309–356. https://doi.org/10.1523/JNEUROSCI.04-01-00309.1984

Mahendran, A., & Vedaldi, A. (2015). Understanding deep image representations by inverting them. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5188–5196. https://doi.org/10.1109/CVPR.2015.7299155

Mishkin, M., & Ungerleider, L. G. (1982). Contribution of striate inputs to the visuospatial functions of parieto-preoccipital cortex in monkeys. *Behavioural Brain Research*, *6*, 57–77. https://doi.org/10.1016/0166-4328(82)90081-X

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*, 915–929. https://doi.org/10.1016/j.neuron.2008.11.004

Monti, M. (2011). Statistical analysis of fMRI time-series: A critical review of the GLM approach. *Frontiers in Human Neuroscience*, *5*. https://doi.org/10.3389/fnhum.2011.00028

Naselaris, T., Allen, E., & Kay, K. (2021). Extensive sampling for complete models of individual brains. *Current Opinion in Behavioral Sciences*, *40*, 45–51. https://doi.org/10.1016/j.cobeha.2020.12.008

Nastase, S. A., Gazzola, V., Hasson, U., & Keysers, C. (2019). Measuring shared responses across subjects using intersubject correlation. *Social Cognitive and Affective Neuroscience*, *14*, 667–685. https://doi.org/10.1093/scan/nsz037

Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., & Clune, J. (2016). Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In *Advances in neural information processing systems* (Vol. 29). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2016/file/5d79099fcdf499f12b79770834c0164a-Paper.pdf

Noguchi, A., & Harada, T. (2019). Image generation from small datasets via batch statistics adaptation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2750–2758. https://doi.org/10.1109/ICCV.2019.00284

Nonaka, S., Majima, K., Aoki, S. C., & Kamitani, Y. (2021). Brain hierarchy score: Which deep neural networks are hierarchically brain-like? *iScience*, *24*, 103013. https://doi.org/10.1016/j.isci.2021.103013

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*, 9868–9872. https://doi.org/10.1073/pnas.87.24.9868

Ojha, U., Li, Y., Lu, J., Efros, A. A., Jae Lee, Y., Shechtman, E., & Zhang, R. (2021). Few-shot image generation via cross-domain correspondence. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10738–10747. https://doi.org/10.1109/CVPR46437.2021.01060

Oosterhof, N. N., Wiestler, T., Downing, P. E., & Diedrichsen, J. (2011). A comparison of volume-based and surface-based multi-voxel pattern analysis. *NeuroImage*, *56*, 593–600. https://doi.org/10.1016/j.neuroimage.2010.04.270

Parker, A. J. (2007). Binocular depth perception and the cerebral cortex. *Nature Reviews Neuroscience*, *8*, 379–391. https://doi.org/10.1038/nrn2131

Peirce, J. W. (2015). Understanding mid-level representations in visual processing. *Journal of Vision*, *15*, 5. https://doi.org/10.1167/15.7.5

Penfield, W., & Boldrey, E. (1937). Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. *Brain*, *60*, 389–443. https://doi.org/10.1093/brain/60.4.389

Poggio, G. F., Motter, B. C., Squatrito, S., & Trotter, Y. (1985). Responses of neurons in visual cortex (v1 and v2) of the alert macaque to dynamic random-dot stereograms. *Vision Research*, *25*, 397–406. https://doi.org/10.1016/0042-6989(85)90065-3

Poldrack, R. A., Mumford, J. A., & Nichols, T. (2011). *Handbook of functional MRI data analysis*. Cambridge University Press.

Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage*, *84*, 320–341. https://doi.org/10.1016/j.neuroimage.2013.08.048

Puce, A., Allison, T., Asgari, M., Gore, J. C., & McCarthy, G. (1996). Differential sensitivity of human visual cortex to faces, letterstrings, and textures: A functional magnetic resonance imaging study. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *16*, 5205–5215. https://doi.org/10.1523/JNEUROSCI.16-16-05205.1996

Rakhimberdina, Z., Jodelet, Q., Liu, X., & Murata, T. (2021). Natural image reconstruction from fMRI using deep learning: A survey. *Frontiers in Neuroscience*, *15*, 795488. https://doi.org/10.3389/fnins.2021.795488

Ribeiro, F. L., Bollmann, S., & Puckett, A. M. (2021). Predicting the retinotopic organization of human visual cortex from anatomy using geometric deep learning. *NeuroImage*, *244*, 118624. https://doi.org/10.1016/j.neuroimage.2021.118624

Roux, F.-E., Niare, M., Charni, S., Giussani, C., & Durand, J.-B. (2020). Functional architecture of the motor homunculus detected by electrostimulation. *The Journal of Physiology*, *598*, 5487–5504. https://doi.org/10.1113/JP280156

Schönemann, P. H. (1966). A generalized solution of the orthogonal procrustes problem. *Psychometrika*, *31*, 1–10. https://doi.org/10.1007/BF02289451

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. https://doi.org/10.1016/j.neuroimage.2018.07.043

Sereno, M. I., Dale, A. M., Reppas, J. B., Kwong, K. K., Belliveau, J. W., Brady, T. J., Rosen, B. R., & Tootell, R. B. H. (1995). Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science*, *268*, 889–893. https://doi.org/10.1126/science.7754376

Shen, G., Horikawa, T., Majima, K., & Kamitani, Y. (2019a). Deep image reconstruction from human brain activity. *PLOS Computational Biology*, *15*, e1006633. https://doi.org/10.1371/journal.pcbi.1006633

Shen, G., Dwivedi, K., Majima, K., Horikawa, T., & Kamitani, Y. (2019b). End-to-end deep image reconstruction from human brain activity. *Frontiers in Computational Neuroscience*, *13*, 21. https://doi.org/10.3389/fncom.2019.00021

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. https://doi.org/10.48550/arXiv.1409.1556

Smith, P. L., & Little, D. R. (2018). Small is beautiful: In defense of the small-n design. *Psychonomic Bulletin & Review*, *25*, 2083–2101. https://doi.org/10.3758/s13423-018-1451-8

Talairach, J., & Tournoux, P. (1988). *Co-planar stereotaxic atlas of the human brain: 3-dimensional proportional system : An approach to cerebral imaging*. Thieme Medical Publishers, Inc., New York. https://books.google.co.jp/books?id=pYFiQgAACAAJ

Tustison, N. J., Avants, B. B., Cook, P. A., Yuanjie Zheng, Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4itk: Improved n3 bias correction. *IEEE Transactions on Medical Imaging*, *29*, 1310–1320. https://doi.org/10.1109/TMI.2010.2046908

Vaina, L. (1987). Visual texture for recognition. In *Matters of intelligence: Conceptual structures in cognitive neuroscience* (pp. 89–114). Springer Netherlands. https://doi.org/10.1007/978-94-009-3833-5_4

Valverde, J. M., Imani, V., Abdollahzadeh, A., De Feo, R., Prakash, M., Ciszek, R., & Tohka, J. (2021). Transfer learning in magnetic resonance brain imaging: A systematic review. *Journal of Imaging*, *7*, 66. https://doi.org/10.3390/jimaging7040066

Van Essen, D. C. (2005). A population-average, landmark- and surface-based (PALS) atlas of human cerebral cortex. *NeuroImage*, *28*, 635–662. https://doi.org/10.1016/j.neuroimage.2005.06.058

Van Essen, D. C. (2004). Surface-based approaches to spatial localization and registration in primate cerebral cortex. *NeuroImage*, *23*, S97–S107. https://doi.org/10.1016/j.neuroimage.2004.07.024

Van Uden, C. E., Nastase, S. A., Connolly, A. C., Feilong, M., Hansen, I., Gobbini, M. I., & Haxby, J. V. (2018). Modeling semantic encoding in a common neural representational space. *Frontiers in Neuroscience*, *12*, 437. https://doi.org/10.3389/fnins.2018.00437

Watson, J. D. G., Myers, R., Frackowiak, R. S. J., Hajnal, J. V., Woods, R. P., Mazziotta, J. C., Shipp, S., & Zeki, S. (1993). Area v5 of the human brain: Evidence from a combined study using positron emission tomography and magnetic resonance imaging. *Cerebral Cortex*, *3*, 79–94. https://doi.org/10.1093/cercor/3.2.79

Wild, H. M., Butler, S. R., Carden, D., & Kulikowski, J. J. (1985). Primate cortical area v4 important for colour constancy but not wavelength discrimination. *Nature*, *313*, 133–135. https://doi.org/10.1038/313133a0

Yamada, K., Miyawaki, Y., & Kamitani, Y. (2015). Inter-subject neural code converter for visual image representation. *NeuroImage*, *113*, 289–297. https://doi.org/10.1016/j.neuroimage.2015.03.059

Yamada, K., Miyawaki, Y., & Kamitani, Y. (2011). Neural code converter for visual image representation. *2011 International Workshop on Pattern Recognition in NeuroImaging*, 37–40. https://doi.org/10.1109/PRNI.2011.13

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

Yen, S.-C., & Finkel, L. H. (1998). Extraction of perceptually salient contours by striate cortical networks. *Vision Research*, *38*, 719–741. https://doi.org/10.1016/S0042-6989(97)00197-1

Yousefnezhad, M., Selvitella, A., Han, L., & Zhang, D. (2021). Supervised hyperalignment for multisubject fMRI data alignment. *IEEE Transactions on Cognitive and Developmental Systems*, *13*, 475–490. https://doi.org/10.1109/TCDS.2020.2965981

Yousefnezhad, M., & Zhang, D. (2017). Deep hyperalignment. In *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/0768281a05da9f27df178b5c39a51263-Paper.pdf

Zeki, S. M. (1973). Colour coding in rhesus monkey prestriate cortex. *Brain Research*, *53*, 422–427. https://doi.org/10.1016/0006-8993(73)90227-8

Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, *20*, 45–57. https://doi.org/10.1109/42.906424

Zhuang, X., Yang, Z., & Cordes, D. (2020). A technical review of canonical correlation analysis for neuroscience applications. *Human Brain Mapping*, *41*, 3807–3833. https://doi.org/10.1002/hbm.25090

Appendix A

Publications

The current work has yielded the following publications and presentation:

A.1 Manuscript

- Ho, J. K., Horikawa, T., Majima, K., Cheng, F., & Kamitani, Y. (2023). Inter-individual deep image reconstruction via hierarchical neural code conversion. *NeuroImage*, *271*, 120007. https://doi.org/10.1016/j.neuroimage.2023.120007

A.2 Presentation

- Ho, J. K., Horikawa, T., Majima, K., & Kamitani, Y. (2021). Inter-individual deep image reconstruction. Flash talk presentation for *Neuromatch Conference*. https://www.youtube.com/watch?v=z-6LcSEd9H8
- Wang, H., Ho, J. K., Cheng, F., Aoki S. C., & Kamitani, Y. (2023). Inter-individual neural code conversion without paired stimuli. Poster presentation for *Conference on Cognitive Computational Neuroscience*, Oxford, UK.

Appendix B

Code availability

The experimental code that support the findings in this thesis is available from the repository:

- Code for inter-individual deep image reconstruction including neural code converter, Procrustes transformation, optimal transport and hyperalignment:
  https://github.com/KamitaniLab/InterIndividualDeepImageReconstruction
- Code for feature decoding:
  https://github.com/KamitaniLab/dnn-feature-decoding
- Code for image reconstruction:
  https://github.com/KamitaniLab/DeepImageReconstruction
- Code for BH score calculation:
  https://github.com/KamitaniLab/BHscore