

Bridging Neural Representations across Individuals via Deep Neural Network Representations



Haibao Wang

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 2025

I would like to dedicate this thesis to my loving parents ...

Declaration

I, Haibao Wang, declare that this thesis, entitled “Bridging Neural Representations across Individuals via Deep Neural Network Representations” 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.

Haibao Wang
August 2025

Acknowledgements

I would like to express my deepest gratitude to Prof. Yukiyasu Kamitani for not only welcoming me into his exceptional research group but also for his unwavering support and guidance throughout this transformative journey. His mentorship has been instrumental in shaping my growth as a researcher and as an individual. His passion for science and self-discipline have deeply inspired me throughout my journey. Working alongside him and his brilliant team has been an incredible privilege, and I look forward to a lifelong and fruitful collaboration.

To my amazing lab members, I cannot thank you enough for the camaraderie, support, and stimulating discussions that have made this journey so enriching and unforgettable. A special thanks to Dr. Fan Cheng, Dr. Jun Kai Ho, Ken Shirakawa, Dr. Shuntaro Aoki, Dr. Misato Tanaka, Dr. Yoshihiro Nagano, Dr. Jong-Yun Park, Shangfeng Jin, Hideki Izumi, Shiyun Yang, Xiaoxuan Liu from the Kyoto University lab, as well as Dr. Eizaburo Doi at ATR.

I would also like to extend my sincere appreciation to our incredible research administrators, Yukari Kado and Yukiko Masuda, for their meticulous handling of the intricacies of our research projects and their unwavering support.

Finally, to my beloved family and friends, words cannot express how grateful I am for your unwavering belief in me and your constant encouragement. You have been the bedrock of my academic journey, and I share this achievement with you as I proudly graduate from the illustrious Kyoto University.

Abstract

The emergence of large-scale brain data has expanded our capacity to investigate how the brain represents perception across individuals. However, substantial inter-individual variability in the brain anatomy and function poses a major challenge for brain data analysis and generalizing models across individuals. While anatomical alignment can mitigate the individual difference in the brain anatomy, it is insufficient for addressing differences in fine-grained functional topographies. Functional alignment aim to mitigate these variations by aligning brain activity patterns across individuals, but they typically require paired brain data or shared stimuli, which are often unavailable. This thesis mitigate this challenge by proposing a neural code conversion method that overcomes this constraint by optimizing conversion parameters based on the discrepancy between the stimulus contents represented by original and converted brain activity patterns. This approach, combined with hierarchical features of deep neural networks (DNNs) as latent content representations, achieves conversion accuracy comparable to methods using shared stimuli.

Using this method, we demonstrated that reliable inter-individual decoding and image reconstruction of natural visual stimuli is feasible, even in the absence of shared stimuli. The reconstruction quality is robust across various architectures of converters. An analysis of converters trained using different levels of image representations as visual content supports the notion that the use of multiple hierarchical visual features contributes to accurate neural code conversion. Moreover, the converted brain activity could be decoded into the features of another DNN model not used for converter training, to enable successful image reconstruction, indicating that the converter captures generalizable representations beyond the specific readout. Even with a limited amount of training data, images with recognizable object silhouettes could be reconstructed from the converted brain activity.

We evaluated the broad applicability of our proposed method by testing it across diverse experimental conditions. First, we applied the neural code conversion approach to datasets acquired from different sites with diverse scanning protocols, stimuli, and participant populations. Despite these variations, the converted brain activity preserved fine-grained information to support accurate decoding and image reconstruction. Furthermore, we investigated whether the method could generalize to internal, subjective visual experiences,

including visual illusions, attention, and mental imagery. Reconstructions from converted brain activity successfully captured illusory contours, color spreading, attended objects, and imagined shapes, indicating that some subjective perceptual content may share a representational structure across individuals. These results reveal the flexibility and scalability of the proposed framework.

Finally, we explore the extension of neural code conversion beyond vision by applying our method to auditory cortical response. We demonstrate that auditory content can also be converted and reliably decoded across individuals using the same content loss-based approach. The successful conversion of both visual and auditory cortical responses suggests the potential versatility of our approach for studying neural representations across different sensory and cognitive domains.

This thesis contributes a flexible method for functional alignment and inter-individual brain decoding. By bridging the neural representations across individuals via deep neural networks representations, this work advances our understanding of how object is represented in the brain, and offers a promising framework for scalable neural data analysis and modeling and a foundation for brain-to-brain communication.

Table of contents

List of figures	xvii
1 Introduction	1
1.1 The challenge of individual differences	2
1.1.1 Barrier to brain big data	2
1.1.2 Barrier to brain decoding	5
1.2 The source of individual differences	6
1.2.1 Anatomy	7
1.2.2 Function	9
1.3 Brain alignment	12
1.3.1 Anatomical alignment	12
1.3.2 Functional alignment	14
1.4 Proposed approach	16
1.5 The thesis organization	17
2 Functional brain activity data	19
2.1 Introduction	19
2.2 Fundamentals of fMRI	20
2.3 fMRI datasets	22
2.3.1 Deeprecon dataset	23
2.3.2 THINGS dataset	24
2.3.3 Natural Scenes Dataset	25
2.3.4 Illusion dataset	27
2.3.5 Attention dataset	28
2.3.6 DeepSoundRecon dataset	29
2.3.7 Data availability	31
2.4 Data for analysis	31

3	Functional alignment methods	33
3.1	Introduction	33
3.2	Hyperalignment	34
3.3	Neural code converter	37
3.4	Other alignment methods	39
3.5	Application of functional alignment	42
3.6	Summary	47
4	Within-individual brain decoding	49
4.1	Introduction	49
4.2	DNN architectures and content representations	50
4.3	Brain decoding and visual image reconstruction	55
4.3.1	Brain decoding	55
4.3.2	Visual image reconstruction	58
4.4	Setup of within-individual brain decoding	62
4.5	Summary	65
5	Neural code conversion using content representations	67
5.1	Introduction	67
5.2	Methods	68
5.2.1	Content loss-based neural code converter	68
5.2.2	Brain loss-based neural code converter	69
5.2.3	Other brain loss-based functional alignment methods	70
5.2.4	Conversion accuracy	71
5.2.5	Visual subarea-wise conversion	72
5.2.6	Conversion without shared stimuli	73
5.2.7	Group statistics	74
5.3	Results	75
5.3.1	Visual neural code conversion	75
5.3.2	Hierarchical correspondence of neural representations	76
5.3.3	Conversion without shared stimuli	79
5.4	Discussion	81
6	Inter-individual decoding and image reconstruction	83
6.1	Introduction	83
6.2	Methods	84
6.2.1	Deep neural networks	84

6.2.2	DNN feature decoding analysis	85
6.2.3	Visual image reconstruction	86
6.2.4	Identification analysis	87
6.2.5	Group statistics	87
6.3	Results	89
6.3.1	Inter-individual decoding	89
6.3.2	Inter-individual image reconstruction	89
6.3.3	Inter-individual decoding and image reconstruction using different alignment methods	97
6.3.4	The effect of overlapping (shared) stimuli between converter and decoder trainings	99
6.4	Discussion	100
7	Hierarchical levels of content representations	103
7.1	Introduction	103
7.2	Methods	104
7.2.1	Converter training using the loss from different DNN layers	104
7.2.2	Robustness across content-loss based converter architecture	105
7.2.3	Group statistics	105
7.3	Results	106
7.3.1	Converter training using the loss from different layers	106
7.3.2	Robustness of converter architecture	111
7.4	Discussion	111
8	Generalizability to different decoding tasks	115
8.1	Introduction	115
8.2	Methods	116
8.2.1	Decoding the converted brain activity using different schemes	116
8.2.2	Varying the number of training data for conversion	116
8.2.3	Group statistics	117
8.3	Results	117
8.3.1	Decoding the converted brain activity using different schemes	117
8.3.2	Varying the number of training data for conversion	121
8.4	Discussion	124
9	Inter-site neural code conversion	127
9.1	Introduction	127

9.2	Methods	128
9.2.1	Inter-site neural code conversion	128
9.2.2	Group statistics	129
9.3	Results	129
9.3.1	Inter-site DNN feature decoding	129
9.3.2	Inter-site image reconstruction	132
9.4	Discussion	137
10	Neural code conversion with subjective experience	139
10.1	Introduction	139
10.2	Methods	142
10.2.1	DNN feature decoding	142
10.2.2	Visual image reconstruction	142
10.2.3	Evaluation of reconstruction	143
10.2.4	Group statistics	145
10.3	Results	145
10.3.1	Neural code conversion for line illusion	145
10.3.2	Neural code conversion for color illusion	147
10.3.3	Neural code conversion during selective attention	151
10.3.4	Neural code conversion for visual imagery	154
10.4	Discussion	156
11	Auditory neural code conversion	159
11.1	Introduction	159
11.2	Methods	160
11.2.1	Auditory fMRI data	160
11.2.2	Auditory content representations	161
11.2.3	Auditory neural code conversion methods	161
11.2.4	Conversion accuracy for auditory stimuli	162
11.2.5	DNN feature decoding analysis	163
11.2.6	Group statistics	163
11.3	Results	163
11.3.1	Auditory neural code conversion	163
11.3.2	Auditory inter-individual DNN feature decoding	164
11.4	Discussion	166

12 General discussion	167
12.1 Summary	167
12.2 Revisit the methodology	168
12.3 Future directions	171
References	173
Appendix A Publications	191
Appendix B Code availability	193

List of figures

1.1	Trade-off between number of subjects and amount of data per subject . . .	4
1.2	Decoding orientation from human brain activity	6
1.3	The individual variability in V1 on the inflated left hemisphere of the brain for two individuals	8
1.4	The individual difference in the functional topography	11
1.5	Align brains via deep neural network representations	17
2.1	Deeprecon dataset	24
2.2	THINGS fMRI dataset	25
2.3	Natural Scenes Dataset	26
2.4	DeepSoundRecon dataset	30
3.1	Hyperalignment	35
3.2	Piecewise and searchlight alignment	36
3.3	Neural code converter	38
3.4	The overview of the MindEye2	42
3.5	The inter-individual reconstructed artificial images from the V1 area	43
3.6	The inter-individual reconstructed natural images from the whole visual cortex	45
4.1	Hierarchical processing in the human visual system and CNN	53
4.2	The schematic of generic object decoding using DNN representations . . .	58
4.3	Visual image reconstruction by multiscale local image decoders	59
4.4	Image reconstruction of perceptual and subjective experiences	62
4.5	The schematic of within-individual decoding and image reconstruction . . .	63
5.1	The schematic of the content-loss-based neural code converter	68
5.2	The comparison of converter training using content loss-based optimization and brain loss-based optimization	70
5.3	Evaluation of conversion accuracy	72

5.4	The schematic of subarea-wise conversion	73
5.5	Non-overlapping stimuli for source and target subjects	74
5.6	Conversion accuracy	75
5.7	Conversion accuracies of individual pairs	76
5.8	Correlation in conversion accuracy between brain loss and content loss . . .	77
5.9	Comparison of conversion accuracies across functional alignment methods .	78
5.10	Comparison of conversion accuracies between whole VC and subarea-wise approaches	79
5.11	Conversion performance between different visual subareas	80
5.12	The effect of overlapping (shared) stimuli between converter and decoder trainings	81
6.1	Inter-individual decoding and image reconstruction	84
6.2	The schematic of the reconstruction method	88
6.3	Inter-individual decoding accuracy	90
6.4	Visual image reconstruction	91
6.5	Reconstructed natural images of all pairs	92
6.6	Reconstructed artificial images of all pairs	93
6.7	Reconstructed natural images from different ROIs	94
6.8	Reconstructed artificial images from different ROIs	95
6.9	Identification accuracy	96
6.10	Evaluation of reconstructions based on feature correlations	97
6.11	Inter-individual decoding using different functional alignment methods . . .	98
6.12	Inter-individual image reconstruction using different functional alignment methods	98
6.13	Inter-individual decoding without overlapping stimuli	99
6.14	Inter-individual reconstruction without overlapping stimuli	100
7.1	Converter training using loss from different DNN layers	104
7.2	Converters trained using the loss from different DNN layers	107
7.3	DNN feature decoding performance (pattern) with converters trained using the loss from different DNN layers	108
7.4	DNN feature decoding performance (profile) with converters trained using the loss from different DNN layers	109
7.5	Reconstructions with converters trained using the loss from different DNN layers	110

7.6	Identification accuracy of reconstructions with converters trained using the loss from different DNN layers	110
7.7	Conversion accuracy using converters consisting of different architectures .	112
7.8	Reconstruction using converters consisting of different architectures	112
8.1	DNN feature decoding performance using the CLIP-ViT decoder	118
8.2	Reconstruction by converter and decoder trained with different DNNs . . .	119
8.3	Reconstructed images with the CLIP-ViT decoder for all pairs	120
8.4	Reconstructed images using the VGG19 decoder	122
8.5	Identification accuracy of reconstructed images using the VGG19 decoder .	122
8.6	The effect of training sample size for conversion on generalization	123
8.7	The effect of training sample size for conversion on identification accuracy .	123
9.1	Inter-site neural code conversion and image reconstruction	128
9.2	Feature decoding accuracy for Deeprecon test data	130
9.3	Feature decoding accuracy for THINGS test data	130
9.4	Feature decoding accuracy for NSD test data	131
9.5	Inter-site reconstruction for Deeprecon test data	132
9.6	Inter-site image reconstructions using Deeprecon test data for all pairs . . .	133
9.7	Inter-site reconstruction for THINGS test data	134
9.8	Inter-site reconstructions for NSD test data	134
9.9	Inter-site image reconstructions using THINGS test data for all pairs	135
9.10	Inter-site image reconstructions using NSD test data for all pairs	136
10.1	Neural code conversion with subjective experience	141
10.2	Reconstructed illusory line images	146
10.3	Quantitative evaluation of reconstructed illusory line images	148
10.4	Reconstructed neon color spreading images	149
10.5	Quantitative evaluation of reconstructed neon color spreading images . . .	150
10.6	Inter-individual reconstruction for attended images and single images . . .	152
10.7	Quantitative evaluation of reconstructed attended images and single images	153
10.8	Inter-individual reconstruction for imagined images	155
10.9	Quantitative evaluation of reconstructed imagined images	156
11.1	Align brains via latent representations derived from DNNs for the task of interest	160
11.2	Auditory stimuli and ROIs	161

11.3 The overview of auditory converter training using content loss-based optimization and brain loss-based optimization	162
11.4 Auditory conversion accuracy	164
11.5 Auditory decoding accuracy	165
12.1 Toward scalable neural data analysis	172

Chapter 1

Introduction

Neuroscience is being transformed by the growing availability of large-scale brain datasets that span individuals, cognitive tasks, experimental contexts, and imaging technologies. These data provide unprecedented opportunities for investigating the structure and functional organization of the brain. On the one hand, extensive individual-level data enable the construction of models that capture fine-grained information in the brain. On the other hand, diverse datasets across individuals help uncover general principles underlying human perception and cognition. With increasing scale and diversity, the field is moving beyond small-sample studies toward large-scale approaches aimed at identifying fine-grained information across individuals.

This shift marks an important step toward a more comprehensive understanding of the human brain at the population level. However, it also amplifies a central challenge: the substantial variability between individual brains. Unlike many other forms of data, brain data are inherently individual, shaped by each person's unique combination of biology, experience, and context. As a result, data from one brain may not easily generalize to others, making it difficult to combine datasets, identify consistent patterns, or build models that perform reliably across individuals. These individual differences limit the effectiveness of traditional group-level analyses and underscore the need to develop models that can generalize robustly across individuals, tasks, and datasets.

The central aim of this thesis is to understand and mitigate the impact of individual differences in brain data analysis and in developing models that generalize across individuals. By aligning brains while preserving fine-grained information, we can gain a deeper understanding of both shared and individual-specific neural representations. This not only advances fundamental neuroscience by revealing how the brain encodes fine-grained information across individuals, but also supports practical applications such as developing brain-computer interfaces that generalize across individuals.

In this chapter, we begin by introducing the key challenges posed by individual differences in large-scale brain data and brain decoding models. We then explore the origins of these differences in brain anatomy and functional organization. Next, we introduce brain alignment methods as a promising approach to mitigating inter-individual difference by aligning brain data across subjects at both anatomical and functional levels. We subsequently describe our research objectives and methodological framework. Finally, we provide an overview of the subsequent chapters to outline the overall structure of the thesis.

1.1 The challenge of individual differences

Neuroscience seeks to reveal how the brain encodes and processes information by identifying common principles that emerge across individuals. This section introduces the challenges posed by inter-individual differences in the brain big data and brain decoding.

1.1.1 Barrier to brain big data

The human brain, with its approximately 86 billion neurons and over 100 trillion synapses, presents one of the most complex and data-rich systems. Recent advances in neuroimaging, electrophysiology, and computational modeling have ushered in a new era characterized by the exponential growth of neuroscientific data. This transformation has given rise to the concept of "brain big data," referring to the massive, multimodal, and high-dimensional datasets now available for analysis. The scale of brain big data is not defined solely by volume, but also by how data are distributed along two essential dimensions: depth, which reflects the amount and resolution of data collected per individual, and width, which reflects the diversity and number of individuals studied. These dimensions fundamentally shape how brain data are acquired, processed, and interpreted in modern neuroscience. However, brain big data also present a fundamental challenge: significant differences between individual brains can make it difficult to compare, integrate, and analyze data across individuals.

The pursuit of the depth of brain big data has been made possible by major advancements in neuroimaging and electrophysiology technologies. Techniques such as electroencephalography (EEG), magnetic resonance imaging (MRI), functional MRI (fMRI), positron emission tomography (PET), and magnetoencephalography (MEG) have evolved rapidly since their introduction in the 20th century. Modern fMRI, for instance, can now acquire multiple whole-brain volumes per second, a significant improvement over the one-volume-every-four-seconds standard in early studies. Similarly, diffusion imaging has progressed from basic diffusion tensor imaging (DTI) to high-angular resolution diffusion imaging, allowing for

detailed mapping of neural fiber pathways (Van Horn & Toga, 2014). In parallel, large-scale electrophysiological tools such as Neuropixels probes enable the simultaneous recording of thousands of neurons with single-spike precision over extended periods in animals (Steinmetz et al., 2021) and humans (Paulk et al., 2022). Long-term two-photon imaging has made it possible to monitor the activity of genetically labeled neurons in awake macaques and freely moving mice with cellular and even dendritic resolution (Li et al., 2017; Zong et al., 2022). These technologies bridge cellular and systems-level observations, dramatically increasing the continuity, resolution, and volume of data obtained from an individual brain (Van Horn, 2021). However, while such depth enhances the ability to construct computational models for individual brains, the substantial variability across individuals can make it difficult to directly compare or analyze data at the group level.

On the other hand, expanding the width of brain big data has been a major focus of neuroscience. Large-scale initiatives such as the Human Connectome Project (Benson et al., 2018; Van Essen et al., 2012) and the UK Biobank (Alfaro-Almagro et al., 2018) have collected multimodal brain imaging and behavioral data from thousands of participants. These efforts enable researchers to explore inter-individual variability, identify population-level biomarkers, and build generalizable machine learning models. To support these studies, platforms such as NeuroData and brainlife.io provide scalable solutions for distributed storage, standardized processing, and reproducible analysis across institutions (Hayashi et al., 2024; Vogelstein et al., 2018). High-performance computing infrastructures further facilitate the large-scale extraction of imaging-derived phenotypes, which serve as key features for modeling brain-behavior relationships across diverse individuals (Chen et al., 2019; Gillis et al., 2025). These advances supported the broad, population-level investigations. However, the brain varies substantially across individuals, it is often difficult to directly integrate data across individuals.

A fundamental trade-off in brain big data lies between collecting limited data from many individuals (width) and collecting extensive data from a few individuals (depth). Naselaris et al. (2021) propose a two-dimensional conceptual framework defined by two orthogonal axes: the number of subjects (width) and the volume of data per subject (depth). Figure 1.1 illustrates how different experimental strategies distribute resources within this depth-width space. Traditional fMRI studies are typically positioned in the upper-left region of this space, characterized by large participant samples but limited data collected from each individual, and are primarily designed to support group-level generalization. In contrast, emerging datasets such as the Natural Scenes Dataset (NSD; Allen et al., 2022) are located in the lower-right region, where extensive data are collected from a small number of individuals, enabling the construction of complete computational models for single brains. They argue

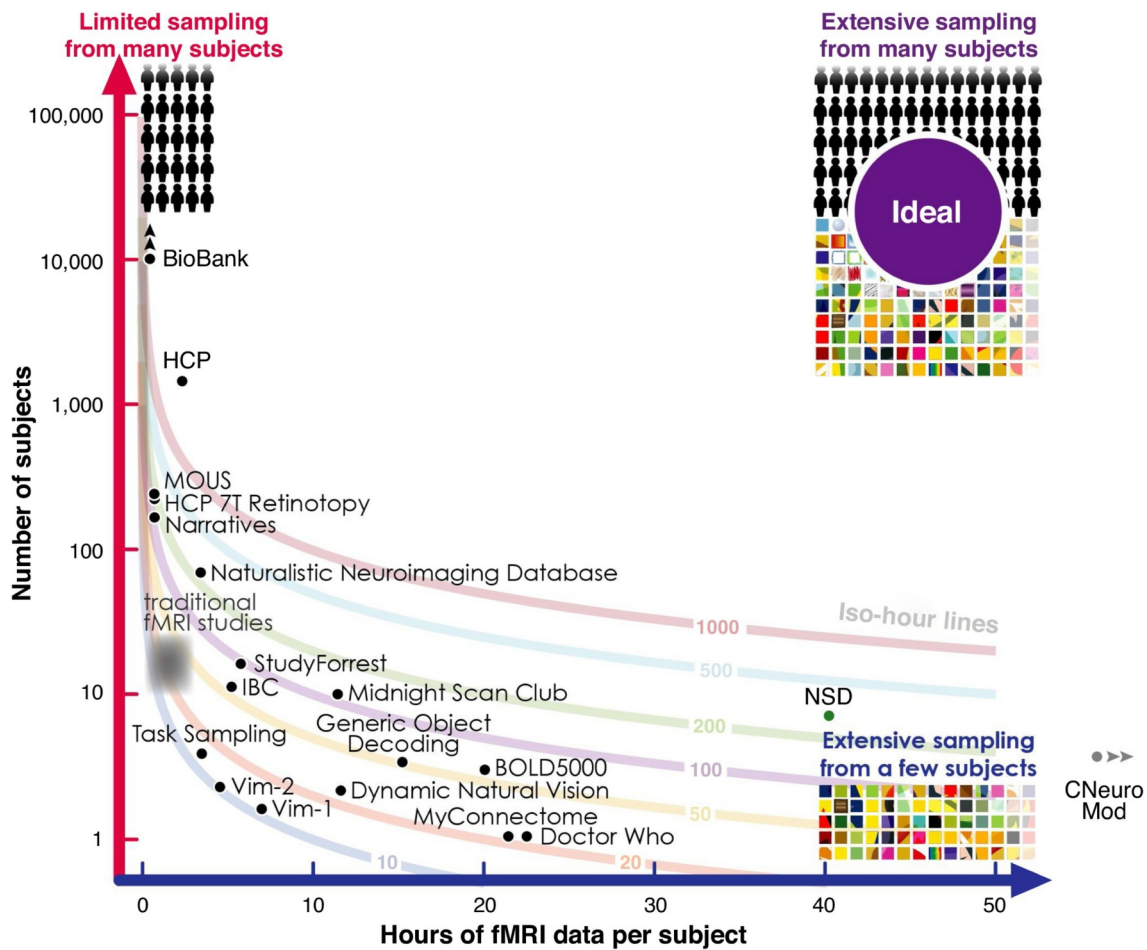


Fig. 1.1 Trade-off between number of subjects and amount of data per subject. In the context of basic-science fMRI studies with publicly available data, a few representative publications, biasing towards recent and/ or influential studies, were selected. For each publication, the number of subjects scanned and the number of hours of task-based and resting-state fMRI conducted for each subject (estimated based on information provided in the publication) are plotted. The thick 'iso-hour' lines indicate points that share the same number of total hours of fMRI data. Adapted from the Madan (2022), used under CC BY 4.0 license.

that, under limited experimental resources, sampling many individuals provides relatively few benefits, and intensive within-subject sampling of a limited number of subjects is often more effective for uncovering general principles of brain function. Building on this foundation of depth, expanding to include more diverse individuals becomes essential for developing models. Ultimately, the ideal scenario involves extensive sampling from many individuals, combining the advantages of depth and width to construct brain models that are robust and scalable.

1.1.2 Barrier to brain decoding

With extensive sampling of brain data within individuals, it becomes possible to build brain decoding models that reveal fine-grained perceptual or cognitive contents in the brain. Brain decoding involves classifying, identifying, or reconstructing what an individual perceives based on brain activity patterns, such as those recorded via functional magnetic resonance imaging (fMRI). The feasibility of decoding perceptual content from distributed brain activity patterns was first convincingly demonstrated by Kamitani and Tong (2005), who showed that simple visual features like grating orientation could be decoded from activity in early visual areas (Figure 1.2). Visual image reconstruction has emerged as one of the most technically demanding and conceptually important branches, aiming to reproduce perceived or imagined visual content directly from brain signals. Early studies focused on reconstructing simple stimuli such as binary patterns (Miyawaki et al., 2008). Since then, decoding and reconstruction techniques have progressed rapidly, contributing to both theoretical neuroscience and real-world applications. From a scientific perspective, they provide insights into how sensory information is represented and processed in the brain. From a practical perspective, they support technologies such as brain-computer interfaces.

Deep neural networks (DNNs), particularly convolutional neural networks (CNNs), have become indispensable tools for brain decoding. Their hierarchical architecture allows them to extract features ranging from low-level edges to high-level semantic categories, mirroring the progression from early to higher-order visual areas in the human brain. Studies using representational similarity analysis (RSA) have shown that early layers of deep neural networks (DNNs) align with activity in early visual areas such as V1 and V2, while deeper layers correspond to higher-order regions like V4 and the inferior temporal cortex (Cichy et al., 2016; Kriegeskorte, 2008). Similarly, Horikawa and Kamitani (2017) demonstrated that DNN-derived features could be reliably predicted from brain activity across visual cortical areas. Their results showed that early visual regions (e.g., V1/V2) achieved higher prediction accuracy for lower-layer features, while higher-order areas (e.g., the lateral occipital complex) better predicted higher-layer features associated with object-level representations. The hierarchical correspondence reflects the similarity between the feature representations learned by DNNs and the functional organization of the human visual system. These findings have laid the foundation for using DNNs as computational proxies for visual representations in brain decoding and reconstruction tasks. Moreover, by leveraging DNN-derived representations, recent visual image reconstruction studies have made significant progress in decoding complex visual experiences from brain activity. These include reconstructing externally presented stimuli (Shen et al., 2019b, 2019a) and subjective experiences (Cheng et al., 2023; Horikawa & Kamitani, 2022), and have achieved increasingly high levels of fidelity, demonstrating how

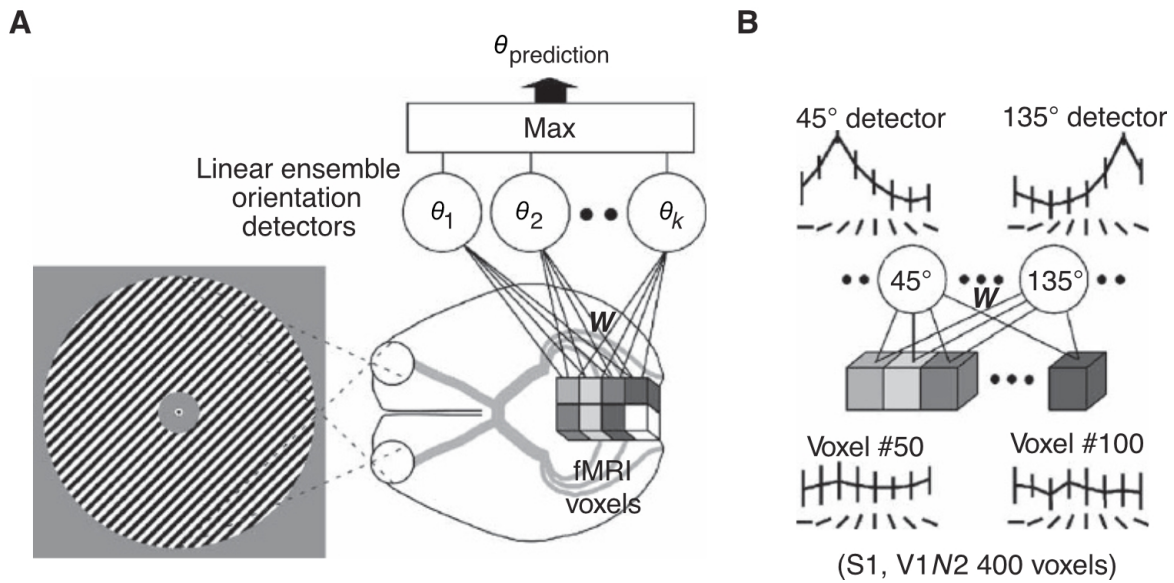


Fig. 1.2 Decoding orientation from human brain activity. Adapted from Kamitani and Tong (2005), copyright 2005 with permission from Springer.

deeply sampled brain data can reveal fine-grained neural representations. These decoding and reconstruction studies will be further discussed in Chapter 4.

Despite these advances in decoding and reconstruction studies, a major barrier remains: the substantial variability across individual brains. Unlike image or text data, which can be freely aggregated from diverse sources, brain data cannot be directly pooled across individuals due to individual differences. As a result, most existing models are trained and evaluated within individuals, limiting their ability to generalize across brains.

1.2 The source of individual differences

Individual differences in the brain originate from two primary sources: anatomical structure and functional organization. These foundational sources of variability not only distinguish one brain from another but also influence how individuals perceive, think, and behave. By reviewing how anatomy and functional organization vary across individuals, this section highlights the implications of such variability for neuroscience. Recognizing and understanding these differences is critical for developing brain models and analytical methods that reflect the both the commonalities and differences of each individual.

1.2.1 Anatomy

The human brain is a remarkably complex network of neural connections that defines our individuality and underlies higher-order functions such as cognition, language, memory, and sensory processing. Among its components, the cerebrum is the largest and most intricate part, divided into left and right hemispheres that are connected by the corpus callosum, a dense bundle of nerve fibers enabling efficient communication and integration between the two sides. While the hemispheres share a largely consistent structural organization, certain functions exhibit lateralization. For example, language is typically processed in the left hemisphere, while spatial awareness and face recognition are more associated with the right. The degree of such lateralization, however, can vary across individuals due to factors such as handedness and cultural background. Each hemisphere is further divided into four primary lobes: the frontal lobe, responsible for decision-making, planning, and motor control; the parietal lobe, which processes sensory input and spatial orientation; the temporal lobe, involved in auditory processing and memory; and the occipital lobe, dedicated to visual processing. Although the general structure of these lobes is similar across individuals, there are substantial inter-individual variability in brain anatomy. These include differences in total brain volume, as well as regional variations in the volume of gray matter, which contains neuronal cell bodies, and white matter, which contains axonal tracts (Alexander-Bloch et al., 2013; Llera et al., 2019). The thickness of the cerebral cortex, the outer layer involved in many higher cognitive functions, also varies among individuals (Riccelli et al., 2017). These anatomical differences may lead to the wide variability in human cognition and behavior, influencing how each individual perceives and engages with the world.

One important finding is the large variation in the surface area of early visual cortices among individuals. Recent high-resolution mapping studies have shown that the functionally-defined size of areas V1, V2, and V3 can differ by more than a factor of three between healthy adults (Benson et al., 2022), consistent with earlier findings (Andrews et al., 1997). This threefold difference in V1 surface area persists even after normalizing for overall brain or cortical size, indicating genuine variability in visual cortex morphology. Figure 1.3 in Himmelberg et al. (2022) shows polar angle and eccentricity maps on the inflated left hemisphere of the brain for two individuals: one with the largest and one with the smallest V1 surface areas. Both shows have complete representations of the right visual hemifield within their V1. This indicates that individual variability in V1 size does not reflect mapping errors, but rather a true anatomical difference in how much cortical surface is allocated for visual processing. These anatomical differences tend to be coordinated across the visual pathway. For example, a brain with an unusually large V1 often also has a larger lateral geniculate nucleus and optic tract, whereas a small V1 comes with smaller subcortical structures. Such

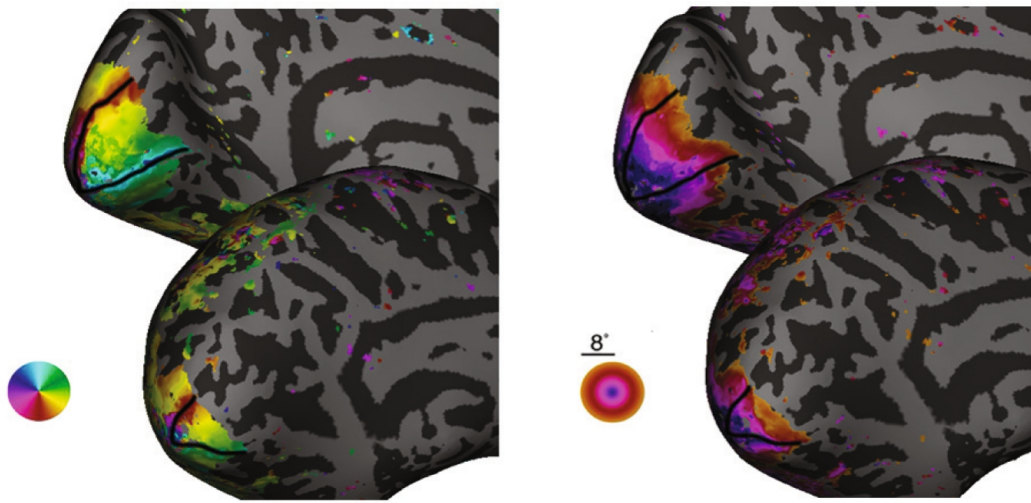


Fig. 1.3 The individual variability in V1 on the inflated left hemisphere of the brain for two individuals. Adapted from Himmelberg et al. (2022), used under CC BY 4.0 license.

differences suggest interdependent development of visual structures and imply that variability in anatomy may lead to meaningful differences in visual capabilities (Andrews et al., 1997).

Variability in cortical folding and microstructural features further underscores individual differences. The calcarine sulcus, which contains most of V1, varies in depth and extent across individuals, leading to differences in the precise shape and boundaries of V1. Similarly, the cortical magnification factor (defined as the amount of cortical area devoted to a given portion of the visual field) shows considerable inter-individual variability. In some individuals, a greater proportion of cortical area is devoted to central (foveal) vision, while in others, more is allocated to the periphery, with the proportion for central vision differing by over a factor of two between individuals (Benson et al., 2022). Genetic and developmental factors have been shown to contribute to these structural differences. For instance, twin studies have reported high intra-pair correlations in the surface areas of visual regions, consistent with a heritable component for the surface area of V1 through V3 (Benson et al., 2022). Overall, while the anatomical organization of the visual system is consistent, the precise morphometry of the visual cortex (including area size, cortical thickness, folding patterns, and internal microarchitecture) is unique to each individual.

Such anatomical differences in the visual cortex have been consistently linked to variations in visual perception and behavior. For instance, research indicates that individuals with larger primary visual cortex (V1) areas typically exhibit enhanced visual acuity and increased sensitivity in specific visual tasks. Recent empirical evidence has demonstrated that both the overall size of V1 and the cortical magnification factor within this region correlate positively with individual differences in visual contrast sensitivity (Himmelberg et al., 2022).

Specifically, individuals possessing larger V1 cortical areas, characterized by more detailed cortical mappings, demonstrate superior detection abilities for low-contrast visual stimuli compared to those with smaller V1 areas. These findings highlight the functional implications of anatomical variability, reinforcing those differences in cortical morphology, including surface area and folding patterns, directly influence how effectively visual information is processed and ultimately perceived by individuals.

Like the visual system, the human auditory system has a broadly consistent anatomical organization across individuals, especially at the macroscopic level of primary sensory areas. The primary auditory cortex is located on the transverse temporal gyrus, also known as Heschl's gyrus, on the superior surface of the temporal lobe in each hemisphere. This area, often referred to as the core auditory cortex and including fields such as A1, is tonotopically organized in all typical brains, meaning that neurons are arranged according to the frequency of sound they respond to, forming a gradient from low to high pitch. Surrounding the core are belt and parabelt regions that contribute to higher-level auditory processing, and beyond these lie multimodal association areas such as the superior temporal gyrus and the planum temporale, which are involved in integrating sound with language, music, or spatial information. However, there are significant individual differences in auditory cortex. One example is the variation in the morphology of Heschl's gyrus. While some individuals have a single gyrus in each hemisphere, others show a duplicated or segmented form, a common anatomical variant observed in up to 50 percent of healthy individuals (Takahashi et al., 2022). This variation has been linked to auditory abilities such as musical aptitude and language acquisition. Similarly, the surface area and volume of auditory cortex can differ substantially between individuals. Individuals with absolute pitch, for example, tend to have larger and more specialized auditory cortical regions (Leipold et al., 2019). The planum temporale also varies markedly across individuals, often showing asymmetry between the left and right hemispheres. These findings mirror those seen in the visual system and highlight the anatomical difference in the brain.

1.2.2 Function

The human brain supports high-level functions such as perception, memory, language, and thinking. The functional topography of the brain refers to the spatial mapping of these specialized functions onto distinct cortical areas. Among them, the visual system plays a central role in perceiving the external world. Its functional organization is hierarchical and broadly consistent across individuals, progressing from low-level feature encoding in early visual cortex to complex object and scene representations in higher-level areas. Visual input begins in the retina, is transmitted through the lateral geniculate nucleus (LGN) of

the thalamus, and reaches the primary visual cortex (V1) in the occipital lobe. V1 neurons respond to basic visual features such as edges, orientation, and contrast. From V1, visual information splits into two parallel processing streams: the ventral stream, also known as the “what” pathway, and the dorsal stream, also known as the “where” pathway. The ventral stream projects through V2 and V4 toward the inferior temporal (IT) cortex, including PIT and AIT, and is specialized for recognizing objects and visual form. The dorsal stream passes from V1 through V2 and MT (V5), continuing to MST, LIP, and VIP, and ultimately reaches the posterior parietal cortex (PP), supporting motion analysis, spatial awareness, and visually guided action (Felleman & Van Essen, 1991; Kruger et al., 2013; Mishkin et al., 1983).

Within the hierarchical architecture, different cortical areas are functionally specialized, each contributing uniquely to the processing of visual information. Across these levels, neuronal selectivity becomes progressively more complex. Neurons in V1 are most responsive to basic features such as edge orientation and spatial frequency (Kruger et al., 2013). V2 encodes more integrated properties, including illusory contours and texture-defined boundaries (Hegdé & Van Essen, 2000). V4 is involved in processing surface attributes, curvature, and color, and plays a central role in texture perception (Ungerleider & G., 2000). At the top of the ventral stream, the IT cortex integrates complex visual features and responds selectively to object identity and category (Kanwisher et al., 1997; Kourtzi & Kanwisher, 2000; Kruger et al., 2013). Texture, as a mid-level feature, is constructed through integration across multiple regions including V1, V2, and V4. Object recognition is further supported by a distributed set of high-level regions: the lateral occipital complex (LOC) is involved in general shape processing; the fusiform face area (FFA) is selectively engaged in face recognition; the parahippocampal place area (PPA) processes scenes and environmental layouts (Epstein & Kanwisher, 1998; Kanwisher et al., 1997).

This principle of hierarchical organization and functional topography is not unique to the visual system, and similar patterns are observed in other sensory modalities, such as audition. Sound enters through the ears and is relayed via subcortical structures to the primary auditory cortex (A1) on Heschl’s gyrus, where neurons are organized tonotopically and respond to specific frequencies (Formisano et al., 2003). Surrounding areas in the superior temporal plane and superior temporal gyrus integrate these elementary features into more complex auditory representations, such as speech, music, and spatial location. For example, the planum temporale contributes to sound localization through interaural cue processing (Griffiths & Warren, 2002), while anterior temporal areas are involved in pitch and timbre analysis (Zatorre et al., 2002), and dorsal pathways support auditory-motor interactions such as speech production (Hickok & Poeppel, 2007).

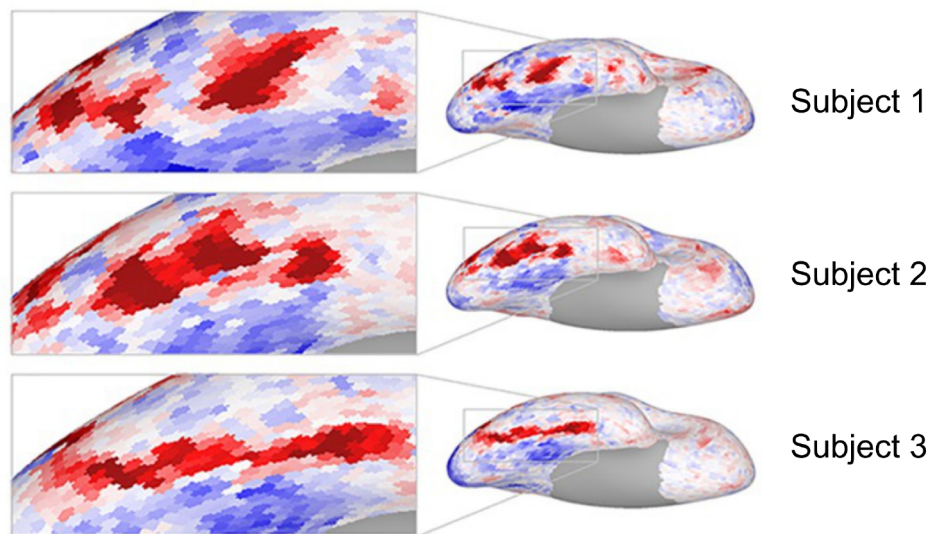


Fig. 1.4 The individual difference in the functional topography. Adapted from Haxby et al. (2020), used under CC BY 4.0 license.

Although the human brain exhibits a broadly similar functional organization across individuals, considerable variability exists in the location of functional areas. Even within anatomically comparable cortical regions, the precise position, size, and boundaries of functionally defined areas can differ from person to person. For instance, the positioning of the visual motion area (known as area V5 or MT) can differ as much as approximately 20 mm across different individuals, even when anatomical alignment is taken into account (Watson et al., 1993). Likewise, the fusiform face area (FFA) exhibits considerable inter-individual variability in both its spatial extent and anatomical location (Figure 1.4). While FFA is typically localized to the lateral fusiform gyrus, its precise position can vary by over 10-15 mm across individuals, even when aligned by anatomical landmarks such as the mid-fusiform sulcus (Kanwisher et al., 1997; Weiner et al., 2014). These findings indicate that anatomical landmarks, such as sulci or gyri, do not consistently predict functional boundaries, making it difficult to define functional regions solely based on anatomical coordinates.

In addition to differences in the spatial location of functional areas, individuals also exhibit variability in the neural activity patterns. Even when performing the same task or viewing the same stimulus, such as a picture of a cat, the specific pattern of neural activity within functionally similar regions, such as the visual cortex, can differ significantly across individuals. These individualized activation patterns, sometimes referred to as “neural fingerprints,” reflect idiosyncratic neural coding. However, while absolute activation patterns vary, the relative structure of representations (e.g., distinctions between categories like faces and places) often remains consistent across individuals. This points to a shared

representational geometry that is preserved despite inter-individual variability. In the case of fMRI, these patterns are reflected in voxel-wise activation maps that capture how different brain regions respond to specific tasks or stimuli. The observed variability underscores the importance of considering individual differences and presents challenges for analyzing brain activity patterns at the group level and developing decoding models that perform reliably across individuals.

1.3 Brain alignment

Brain alignment provides a potential solution to mitigating substantial differences in brain anatomy and function across individuals by mapping individual brain data into a common space. This section introduces two main approaches: anatomical alignment and functional alignment. It explains how each method helps reduce individual variability, allowing for more reliable cross-subject analyses and more generalizable neuroscience models.

1.3.1 Anatomical alignment

Anatomical alignment, also referred to as anatomical normalization or structural registration, is a fundamental preprocessing step in neuroimaging designed to reduce inter-individual variability in brain anatomy. The primary goal is to transform individual brain scans into a common coordinate system, thereby enabling meaningful cross-subject comparisons in group-level analyses. Several approaches have been developed to achieve anatomical alignment, including landmark-based, volume-based, and surface-based registration techniques.

One of the earliest methods, landmark-based registration, was introduced by Talairach and Tournoux (1988). It relies on the manual identification of a small set of anatomical landmarks, including the anterior and posterior commissures, the midsagittal plane, and the outer boundary of the brain. These landmarks are used to perform piecewise affine transformations that align an individual brain to a standard coordinate space. While this approach provided a standardized spatial framework enriched with histological annotations such as Brodmann areas (Strotzer, 2009), it was based on a single postmortem brain and therefore introduced substantial uncertainty when used as a stereotaxic reference for broader populations. In addition, its reliance on manually localized anatomical features introduced subjectivity and limited both spatial accuracy and reproducibility.

To address these limitations, modern neuroimaging has adopted template-based normalization using population-derived brain templates. Among these, the MNI152 template (Mazziotta et al., 2001), constructed by averaging T1-weighted MRI scans from 152 individ-

uals, is the most widely used. Compared to the Talairach atlas, it provides greater anatomical representativeness and reflects population-level structural variability more accurately. Nevertheless, substantial variability in cortical folding patterns remains, particularly in association cortices, and a single standard template cannot fully capture this diversity. For example, Heschl's gyrus, a key structure in the auditory cortex, exhibits duplicated or variable folding patterns in a substantial portion of the population (Marie et al., 2015), complicating one-to-one correspondence in anatomical alignment. To improve anatomical alignment under such variability, researchers have developed population-specific templates tailored to different age groups (Fonov et al., 2011), cultural or ethnic backgrounds (Liang et al., 2015), and clinical populations. Many of these resources are now accessible through centralized platforms such as TemplateFlow (Ciric et al., 2022), which facilitates standardized and reproducible template selection.

Once a template is selected, anatomical alignment is performed using either volume-based or surface-based registration techniques, depending on the data representation. In volume-based registration, alignment is performed directly on 3D anatomical images by comparing voxel intensity distributions. The process typically involves an initial affine transformation followed by nonlinear warping. Several widely used algorithms have been developed, including SyN in ANTs (Avants et al., 2008), which uses symmetric diffeomorphic mapping, discrete cosine transforms in SPM5 (Penny et al., 2011), and cubic B-spline interpolation in FSL's FNIRT (Andersson et al., 2007). A variety of similarity metrics can be used to guide registration, such as least squares, normalized correlation, correlation ratio and mutual information (Poldrack et al., 2011).

In contrast, surface-based registration aligns brains based on cortical surface geometry, focusing on features such as sulcal depth and gyral curvature. This approach requires reconstructing the cortical sheet from structural scans, a process widely automated in software like FreeSurfer (Dale et al., 1999). The surface is then projected onto a spherical representation, and alignment is achieved by minimizing geometric feature differences across subjects (Fischl et al., 1999). The fsaverage template, generated from 40 individual scans, is commonly used as a reference for this method. More advanced strategies, such as MSM-All (Glasser et al., 2016), integrate both anatomical and functional features to improve inter-individual correspondence.

While anatomical alignment focuses on normalizing structural images across individuals, anatomical-functional coregistration is typically performed beforehand when processing functional data, to align functional volumes to the subject's anatomical scan. This process aligns a subject's functional image, typically acquired using echo planar imaging (EPI), to their high-resolution anatomical scan, establishing a voxel-wise mapping between functional

activity and structural landmarks (Dohmatob et al., 2018). Accurate coregistration is critical for later transforming the functional data into standard space via the anatomical alignment. However, functional images often suffer from susceptibility-induced distortions, particularly in orbitofrontal and anterior temporal regions, due to magnetic field inhomogeneities at air-tissue interfaces. To address this issue, distortion correction techniques such as fieldmap-based unwarping are commonly applied prior to coregistration (Hutton et al., 2002).

Among the various alignment algorithms, boundary-based registration (Greve & Fischl, 2009) has become one of the most widely adopted methods, especially in software packages such as FSL and FreeSurfer. It uses tissue boundaries extracted from anatomical images to guide cross-modal alignment with functional volumes. In addition, nonlinear approaches such as the symmetric diffeomorphic transformation (Avants et al., 2008) have gained popularity, particularly in cases where fieldmap information is unavailable. Despite their effectiveness, these methods remain sensitive to image quality and require careful visual inspection to avoid misregistration artifacts. While anatomical-functional coregistration ensures that functional data are aligned to each subject's anatomy, it does not resolve inter-individual variability in functional topography. This motivates the development of functional alignment techniques, which aim to align functional responses directly across individuals.

1.3.2 Functional alignment

While anatomical alignment standardizes brain structure across individuals, it does not account for inter-individual variability in functional organization. In typical preprocessing pipelines, functional data are first coregistered to each subject's anatomical scan, and then indirectly transformed into standard space using the anatomical alignment. However, this indirect way assumes that functional regions are reliably positioned relative to anatomical landmarks, which is often not the case. This limitation is particularly problematic in higher-order association cortices, where the location, shape, and size of functional regions can vary substantially across individuals, even when their anatomy appears well aligned (Frost & Goebel, 2012; Sabuncu et al., 2010). As a result, spatial normalization based solely on anatomy may misalign functionally homologous regions, reducing the sensitivity and interpretability of group-level analyses.

This challenge has led to the development of functional alignment methods, which aim to directly align functional data across individuals by matching patterns of brain activity instead of anatomical features. This approach is based on the assumption that cognitive representations are reflected in distributed neural activity that systematically maps stimulus features to brain responses. These representations are often studied using decoding approaches (Cheng et al., 2023; Kamitani & Tong, 2005; Miyawaki et al., 2008; Shen et al., 2019b), which rely

on multivariate patterns of activity across voxels to distinguish between different stimulus conditions. However, because of differences in cortical folding and the way voxels sample neural populations, anatomical alignment alone cannot guarantee that corresponding voxels reflect the same functional signal across individuals. As a result, group analyses may miss shared representations if they rely solely on anatomical correspondence. Functional alignment addresses this challenge by learning transformations that bring individual-specific activity patterns into a shared representation space.

This idea, that distributed activation patterns represent meaningful cognitive content and can be compared across individuals, has its origins in connectionist models of representation. To address the challenge of comparing functionally distributed representations across different neural architectures, Paul Churchland proposed a view known as state-space semantics (Laakso & Cottrell, 2005). Within this framework, neural representations are conceptualized as points in a high-dimensional activation space, where each dimension corresponds to the activity of a specific unit or neuron. An activation pattern across three units can be interpreted as a histogram of activation values, a vector in a three-dimensional space, or a single point within that space. Even when the physical arrangement of neurons differs between brains, the relative similarity among representations may still be preserved (Churchland, 1998). For example, activation patterns evoked by the concepts “tiger” and “cat” may be closer to each other than to the pattern evoked by “tree.” Functional alignment techniques aim to identify transformations between individual activation spaces that maintain this underlying representational structure, thereby allowing for direct comparison of cognitive content across individuals.

Several methods have been developed to implement functional alignment. Hyperalignment (Haxby et al., 2011, 2020) estimates subject-specific linear transformations that project individual voxelwise response patterns into a shared high-dimensional representational space. This method preserves fine-grained topographic information and has demonstrated strong performance in decoding accuracy and cross-subject classification. In contrast, the neural code converter (Ho et al., 2023; Yamada et al., 2015) directly learns pairwise transformations between brain activity patterns from different individuals. Rather than aligning all subjects into a single common space, it learns direct transformations between individual brains, which helps preserve fine-grained visual information useful for tasks like inter-individual decoding and image reconstruction. A more detailed comparison of these methods, their extensions, and other functional alignment approaches will be presented in Chapter 3. When aligning neural representations, many of these methods often require participants to view the same stimulus to obtain paired data for training the functional alignment model. This require-

ment limits the scalability of such methods for large-scale or varied datasets and practical applications.

1.4 Proposed approach

While recent decoding and visual reconstruction studies have successfully revealed fine-grained information in the brain through deep (extensive) collection of individual data, these approaches are typically limited to the specific individuals. In this thesis, we aim to extend these deep approaches across individuals. Our central question is: how to align brains while preserving fine-grained information?

Although these functional alignment techniques have shown promising results in aligning fine-grained information across individuals, these methods often depend on paired brain data collected while subjects view the same stimuli. This requirement poses a significant limitation for applying these methods to increasingly diverse and large-scale datasets, where stimuli, recording conditions, and experimental protocols often differ across subjects. In such datasets, although brain responses are available, the lack of shared stimuli or paired brain responses makes it difficult to train conventional alignment models.

While shared stimuli in functional alignment ensure that brain activity patterns across subjects reflect the same stimulus content, the content can be more flexibly represented by a combination of elemental, latent features, such as image bases or deep neural network (DNN) features (Krizhevsky et al., 2012; Radford et al., 2021; Simonyan & Zisserman, 2014). By decomposing a single stimulus into these elemental features, functional alignment may encompass a broader range of stimuli beyond those that are explicitly shared. Recent advancements in brain decoding research have revealed that the contrasts of local image bases and the latent features from DNNs can be decoded from the brain activity patterns measured by fMRI (Horikawa & Kamitani, 2017; Miyawaki et al., 2008). Furthermore, these decoded features can be transformed into images by linear or deep generative networks (Cheng et al., 2023; Miyawaki et al., 2008; Shen et al., 2019b, 2019a), reconstructing perceptual experiences. These results pave the way for using latent features instead of shared stimuli to ensure content similarity in functional alignment.

In this thesis, we propose a functional alignment framework that leverages latent representations derived from DNNs to investigate how brain activity can be aligned while preserving fine-grained information (Figure 1.5). In our approach, we used latent representations derived from DNNs serve as proxies for perceptual content to perform functional alignment. Since these content representations can be inferred from individual’s brain activity or predicted brain activity, they enable content-based functional alignment even without the need for

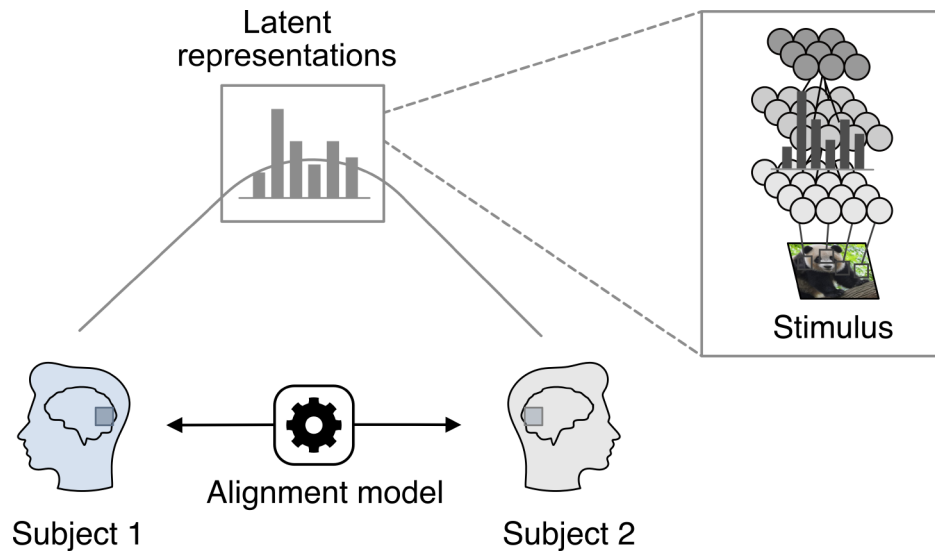


Fig. 1.5 Align brains via deep neural network representations

paired brain data, offering a more flexible alternative to traditional alignment methods. We use brain decoding and visual image reconstruction to evaluate this framework by examining whether fine-grained information is preserved after functional alignment. In addition, we assess the generalizability of the approach across individuals, datasets, cognitive tasks, and sensory modalities to investigate its potential for scalable applications in neuroscience.

1.5 The thesis organization

This thesis is organized into 11 chapters, each focusing on a specific aspect of our investigation into individual differences and the proposed functional alignment framework.

Chapter 1 provides a general introduction to the concept of individual differences. It outlines the challenges these differences pose to brain big data and decoding, explores their anatomical and functional sources, and introduces common strategies for addressing them. It also outlines our proposed approach, which bridges the neural representations across subjects via DNN representations.

Chapter 2 introduces the functional neuroimaging data used in this thesis. It describes the methods of data acquisition, including the visual stimuli employed, the selection of regions of interest, experimental procedures, and data preprocessing. Several representative datasets are introduced, as well as their roles in the thesis.

Chapter 3 focuses on functional alignment methods designed to mitigate individual differences in functional topography. It reviews previous approaches, discusses their strengths

and limitations, and provides context for the methodological improvements proposed in this thesis.

Chapter 4 presents methods for examining neural representations and externalizing individual differences. It introduces within-subject brain decoding and reconstruction techniques, which serve as baseline models in subsequent analyses, and explains the rationale for using reconstruction as a tool to probe neural representational across individuals.

Chapter 5 details the neural code conversion method proposed in this thesis. This chapter introduces a flexible training framework based on DNN latent representations, which removes the need for identical stimuli across individuals. This chapter also evaluates the conversion performance and explores the hierarchical correspondence of neural representations across subjects.

Chapter 6 investigates whether fine-grained visual information is preserved after functional alignment. By applying decoding and reconstruction methods to converted brain activity, the extent to which the fine-grained representation is preserved across individuals is reflected.

Chapter 7 examines how hierarchical features from DNNs influence the performance of neural code conversion models trained using our content loss-based approach.

Chapter 8 explores the generalizability of content loss-based neural code converters. It assesses whether the converted activity retains meaningful information that can be decoded by different schemes, even with limited training data.

Chapter 9 extends the scope of analysis to inter-site neural code conversion. It evaluates the feasibility of performing neural code conversion across different datasets, demonstrating the broader utility of the proposed approach.

Chapter 10 investigates the feasibility of converting neural representations of subjective visual experiences across individuals. It examines visual illusions, selective attention, and visual imagery to assess whether internally driven perceptual content can be aligned and reconstructed across different individuals.

Chapter 11 presents a preliminary extension of the proposed method to the auditory domain. By using auditory latent DNN representations for neural code conversion, this chapter demonstrates the potential application of the proposed method beyond visual domain.

Chapter 12 concludes the thesis by revisiting and discussing the proposed methods. It outlines potential implications, current limitations, and suggests future research directions.

Chapter 2

Functional brain activity data

2.1 Introduction

Understanding how the brain enables complex cognitive functions hinges on the availability of accurate and reliable brain activity data. Neuroimaging technologies have played an important role in this pursuit, providing powerful tools to visualize and quantify brain activity. Over the past decades, a broad range of functional brain imaging modalities has been developed, including Positron Emission Tomography (PET), functional Magnetic Resonance Imaging (fMRI), Near-Infrared Spectroscopy (NIRS), Diffuse Optical Imaging (DOI), Electroencephalography (EEG), Stereo electroencephalography (sEEG), Electrocorticography (ECoG), and Magnetoencephalography (MEG). The rapid advancement of these technologies has significantly deepened our understanding of the human brain and its function.

Among these modalities, EEG, MEG, and fMRI have been particularly influential in the study of stimulus-evoked brain activity data. EEG, with its millisecond-level temporal resolution, enables researchers to capture the brain's electrical responses to external stimuli. By recording voltage fluctuations on the scalp generated by neuronal firing, EEG provides detailed temporal information, offering valuable insights into the timing of cognitive processes. Its relatively low cost, portability, and ease of use make EEG a widely adopted tool in both basic and clinical research. However, due to the distortion of electrical signals as they pass through the skull and scalp, EEG has limited spatial resolution. MEG measures the magnetic fields produced by neural activity, offering temporal resolution comparable to EEG but with improved spatial precision. Because magnetic fields are less affected by the skull and surrounding tissues, MEG is more effective at localizing the sources of brain activity. This makes it useful for studying rapid cognitive processes such as auditory and visual perception, as well as language processing. However, MEG systems are expensive, non-portable, and require highly controlled environments for accurate data acquisition.

In contrast to EEG and MEG, which directly record neural activity through electrical or magnetic signals, fMRI provides an indirect measure of neural response. While standard MRI generates static images of brain anatomy, fMRI captures dynamic fluctuations in brain activity over time with relatively high spatial resolution, allowing researchers to examine how different brain regions respond to and interact during specific cognitive or sensory events. It enables the investigation of neural responses or representations during task-based paradigms, such as viewing images or movies, or listening to sounds, as well as during resting-state conditions. The fMRI data collected under these conditions offer critical insights into how the brain processes a wide range of external stimuli, forming the foundation of the present thesis.

This chapter provides an overview of functional brain activity data, with a particular focus on fMRI. It begins by introducing the fundamental principles of fMRI, followed by a discussion of the criteria for selecting appropriate fMRI datasets. It then presents a review of several widely used datasets involving visual or auditory stimuli. Finally, it explains how these datasets are used for the analyses conducted in the subsequent chapters of this thesis.

2.2 Fundamentals of fMRI

Magnetic Resonance Imaging (MRI) is a widely used non-invasive imaging technique that provides high-resolution structural images of the brain. It is based on the physical principles of nuclear magnetic resonance, which exploits the magnetic properties of certain atomic nuclei, particularly hydrogen protons. These protons are abundant in biological tissues due to the high water content in the human body.

In an MRI scanner, the subject is placed within a strong static magnetic field (typically ranging from 1.5 to 7 Tesla), which causes the magnetic moments of hydrogen protons to align with the external field. A brief radiofrequency (RF) pulse is then applied to disrupt this alignment, tipping the protons away from equilibrium. Once the RF pulse is turned off, the protons relax back to their original state, emitting RF signals in the process. These emitted signals are detected by receiver coils and processed to generate detailed images of internal tissue structures. The signal intensity in MRI is influenced by factors such as proton density and the tissue-specific relaxation times known as T1 (longitudinal relaxation) and T2 (transverse relaxation). By adjusting imaging parameters like repetition time (TR) and echo time (TE), different tissue contrasts can be emphasized, resulting in T1-weighted or T2-weighted images. These contrast variations enable researchers to visualize anatomical structures with high precision. Structural MRI typically provides submillimeter spatial

resolution, making it essential for analyzing brain morphology, measuring cortical thickness, detecting lesions, and serving as an anatomical reference for functional imaging studies.

fMRI extends the capabilities of structural MRI, enabling the measurement of brain activity by detecting changes in blood flow (Belliveau et al., 1991). The core principle behind fMRI is that neural activity is coupled with local changes in cerebral blood flow and oxygenation. When a brain region becomes more active, it consumes more oxygen, prompting an increase in blood flow to replenish the oxygen supply. This results in a localized increase in oxygenated hemoglobin and a decrease in deoxyhemoglobin.

The signal measured by fMRI is the Blood-Oxygen-Level-Dependent (BOLD) signal. Deoxyhemoglobin is paramagnetic and distorts the surrounding magnetic field, whereas oxygenated hemoglobin is diamagnetic and has a much weaker effect (Ogawa et al., 1990). Increased levels of oxygenated hemoglobin in active brain areas lead to a more uniform magnetic field, producing a stronger MRI signal. This contrast in signal intensity between regions with varying levels of oxygenated hemoglobin forms the basis of the BOLD signal used to capture brain activity.

It is important to note that the BOLD signal is an indirect measure of neural activity. It reflects the hemodynamic response that follows neural activation, typically delayed by several seconds. The relationship between the BOLD signal and the underlying neuronal firing is complex and remains an active area of research. The hemodynamic response involves a dynamic interplay between neurons, glial cells, and local vasculature. Following neural activity, a characteristic temporal profile known as the Hemodynamic Response Function (HRF) occurs. The BOLD signal typically begins to rise approximately 1–2 seconds after neural activity occurs, reaches its peak around 4–6 seconds, and returns to baseline over the following several seconds. Understanding these characteristics of the BOLD signal is crucial for accurately interpreting fMRI data and designing effective experimental paradigms.

Building upon these physiological foundations, fMRI has emerged as a powerful tool for investigating human brain function in a wide range of cognitive and clinical contexts. fMRI offers several advantages over other neuroimaging techniques. As a non-invasive and safe method, it does not require the use of radioactive tracers or contrast agents, relying instead on natural changes in blood oxygenation. By tracking changes in BOLD signals over time, researchers can infer which brain regions are engaged during specific tasks, sensory experiences, or mental states. This characteristic makes fMRI suitable for repeated use and applicable across diverse populations, including children and clinical patients. Furthermore, fMRI provides high spatial resolution, typically in the range of millimeters, which enables precise localization of brain activity within specific cortical and subcortical regions. Unlike surface-based techniques such as EEG, fMRI can capture functional signals

from the entire brain. Another important advantage is its flexibility in experimental design, allowing researchers to tailor scanning parameters and employ various task paradigms, such as block designs and event-related designs, to suit specific research questions. In addition, the stability and reproducibility of fMRI measurements make it particularly well-suited for longitudinal studies, such as tracking brain development, cognitive aging, or responses to therapeutic interventions.

Despite its strengths, fMRI also has several limitations. One of the primary drawbacks is its low temporal resolution. The BOLD signal reflects changes in blood flow and oxygenation that occur several seconds after neural activity, with a peak typically observed around 4-6 seconds post-stimulus. This delay limits the technique's ability to capture fast, transient neural events, making it less suitable for studies requiring precise timing, such as those investigating sensory processing or rapid decision-making. Another limitation lies in the fact that the BOLD signal is an indirect measure of neural activity, as it is modulated by vascular and metabolic factors that can vary across individuals and brain regions. fMRI is also highly susceptible to motion artifacts; even small head movements can degrade data quality and complicate interpretation, particularly in studies involving young children, elderly participants, or patients with neurological conditions. Additionally, fMRI systems are expensive to acquire and maintain, and the operational costs are high. Finally, the analysis of fMRI data is complex and computationally intensive, requiring advanced preprocessing pipelines and statistical modeling techniques.

2.3 fMRI datasets

Grounded in the foundational principles and advantages of fMRI, an increasing number of publicly available datasets have been developed to support the investigation of brain function across diverse tasks and experimental conditions. These datasets have become essential resources for neuroscience studies.

To ensure that the datasets used in this thesis are suitable for studying neural representations, several key criteria were considered. First, datasets with naturalistic stimuli were preferred, as they evoke brain responses that more closely reflect real-world perception and behavior. It was also important that each participant was exposed to a large and varied set of stimuli, which allows for more accurate estimation of stimulus-response relationships. In addition, datasets with larger participant samples were favored, as they enable more reliable evaluation of whether the neural representations are shared across individuals. Datasets acquired with high-field MRI scanners (such as 3T or 7T) were selected due to their higher

signal-to-noise ratio, which enables improved resolution. Finally, all chosen datasets are publicly available, ensuring transparency and reproducibility.

The following sections describe several fMRI datasets selected based on the above criteria. These datasets are closely relevant to the goal of this thesis and provide reliable data for studying visual and auditory brain responses.

2.3.1 Deeprecon dataset

The Deeprecon dataset (Figure 2.1) is a 3T visual fMRI dataset developed for brain decoding and visual image reconstruction studies. It was first introduced by Shen et al. (2019b) and later expanded in subsequent studies (Ho et al., 2023; Horikawa & Kamitani, 2022). The dataset includes fMRI recordings from five healthy subjects during image viewing and visual imagery involving natural and artificial stimuli, with a particular focus on decoding visual neural representation in the brain. The training and test stimulus sessions do not share any overlapping categories of images. This design provides a standard paradigm for studying neural representation and assessing the generalization performance of brain decoding models.

Subjects. Five healthy subjects (four males and one female, aged 25-36 years) with normal or corrected-to-normal vision participated in the experiment. Subjects 1-3 were originally included in the study by Shen et al. (2019b), while Subjects 4 and 5 were sourced from Horikawa and Kamitani (2022) and Ho et al. (2023).

Stimuli. The image stimuli consisted of both natural and artificial images. The natural images were selected from 200 representative categories in the ImageNet dataset (2011 fall release; Deng et al., 2009). The training set included 1,200 images from 150 categories, and the test set contained 50 images from the remaining 50 categories. In addition, artificial images were constructed by combining five basic shapes (square, small frame, large frame, plus sign, cross sign) with eight colors (red, green, blue, cyan, magenta, yellow, white, black), resulting in 40 distinct stimuli. For the imagery task, 15 artificial shapes from the artificial test images and 10 natural images from natural test images were used as target images.

Experimental paradigm and data acquisition. In the image viewing task, subjects were instructed to fixate on a central point on the screen and respond with a button press when they saw the same image appear in two consecutive blocks. Each image was presented for 8 seconds in a stimulus block. In the imagery task, subjects imagined the 25 target images, guided by a word cue that had been associated with each image prior to the experiment. Each trial consisted of a 4-second cue period, followed by an 8-second imagery period, a 3-second evaluation period, and a 1-second rest period. Over the course of 15 to 20 scanning sessions, each subject viewed a total of 1,290 images and imagined 25 images across approximately

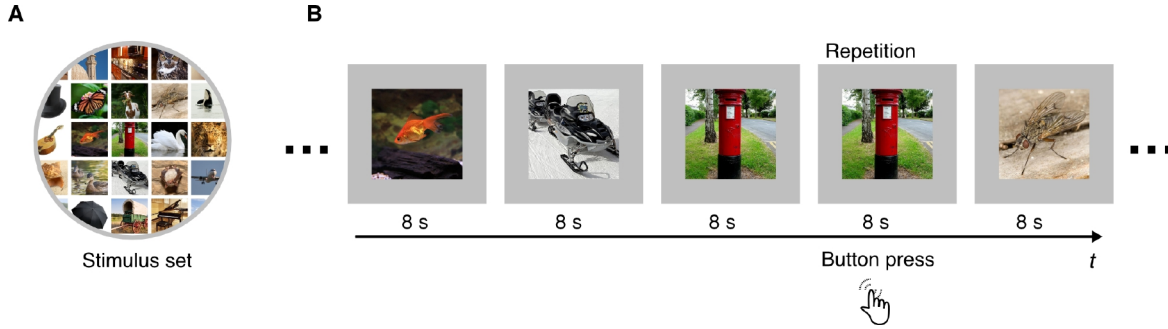


Fig. 2.1 Deeprecon dataset. (A) The colored natural image stimuli. (B) Schematics of image presentation experiment. Each image display lasted for 8 seconds in a stimulus block. Subjects were instructed to fixate on a central point on the screen and respond with a button press when they saw the same image appear in two consecutive blocks.

8,000 trials. Functional brain activity was recorded using a 3T MRI scanner, with a resolution of 2 mm isotropic voxels and a repetition time (TR) of 2 seconds.

MRI data preprocessing. The fMRI data was preprocessed with the fMRIPrep pipeline (Esteban et al., 2019). The BOLD time series were temporally shifted by 4 s to account for hemodynamic delays and then regressed for nuisance variables. The data samples were finally despiked to reduce extreme values (beyond ± 3 SD for each run) in the time series and averaged within each 8-s trial (four volumes). The brain regions V1, V2, V3, and V4 were demarcated using a standard retinotopy experiment (Engel et al., 1994; Sereno et al., 1995) in each subject's native brain space. The higher visual cortex (HVC) was defined by conventional functional localizers (Epstein & Kanwisher, 1998; Kanwisher et al., 1997; Kourtzi & Kanwisher, 2000). The whole visual cortex (VC) was defined as the combination of the regions V1, V2, V3, V4, and HVC.

2.3.2 THINGS dataset

The THINGS dataset (Figure 2.2; Hebart et al., 2023) used in this thesis refers to the fMRI component of the THINGS-data collection, which is a large-scale, multimodal dataset designed to investigate object representations in the human brain and behavior. The fMRI data were collected using a 3T scanner while three participants viewed thousands of naturalistic object images drawn from the THINGS database (Hebart et al., 2019). The large-scale object stimulus set makes it valuable for studying neural representations of objects.

Subjects. Three healthy subjects (one male and two females; mean age: 25.33 years) with normal or corrected-to-normal vision participated in the experiment.

Stimuli. The image stimuli were taken from the THINGS object concept and image database (Hebart et al., 2019). The training images were 8,640 images taken from 720

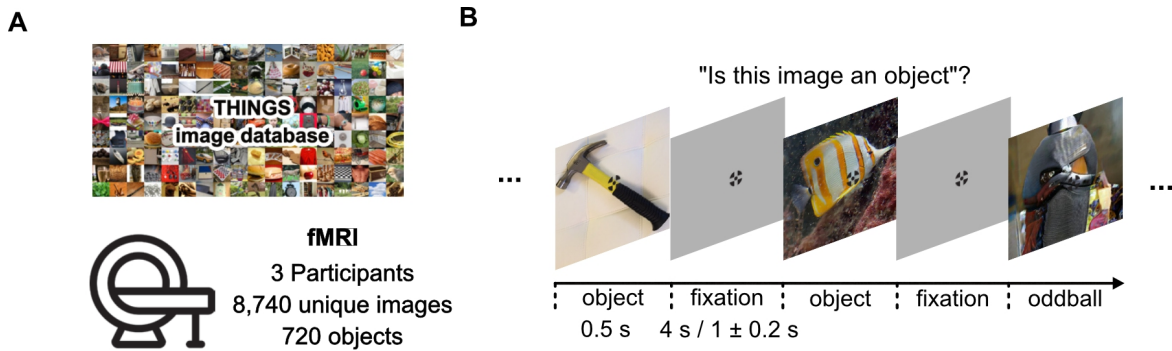


Fig. 2.2 THINGS fMRI dataset. (A) THINGS fMRI data comprises fMRI responses to large samples of object images taken from the THINGS database. (B) In the fMRI experiment, participants viewed object images while performing an oddball detection task (synthetic image). Adapted from Hebart et al. (2023), used under CC BY 4.0 license.

representative object concepts, with the first 12 examples per concept, and the test images were 100 separate images taken from the remaining THINGS images.

Experimental paradigm and data acquisition. The subjects were asked to keep their eyes on the fixation spot of the screen and report the presence of a catch image with a button press on a fiber-optic diamond-shaped button box. Each image was presented for 0.5 ms, followed by 4 s of eye fixation without image stimuli. A 3T scanner was used to collect whole-brain functional MRI data with 2-mm isotropic resolution and 1.5-s repetition time. fMRI signals were measured while subjects each viewed 8,740 unique visual images (11,040 trials) over the course of 15-16 scan sessions.

MRI data preprocessing. The fMRI data were preprocessed using fMRIPrep (Esteban et al., 2019), including slice timing correction, head motion correction, susceptibility distortion correction, alignment to T1-weighted images, tissue segmentation, and cortical surface reconstruction. Independent component analysis (ICA)-based denoising was performed to identify and regress out noise components (Hebart et al., 2023). Then, fMRI single-trial responses were estimated by fitting a general linear model (GLM) on the preprocessed fMRI time series. The region of interest (ROI) was defined based on retinotopic mapping (Engel et al., 1994; Sereno et al., 1995) and functional localizer experiments (Epstein & Kanwisher, 1998; Kanwisher et al., 1997; Kourtzi & Kanwisher, 2000), with VC as the combination of all visual areas.

2.3.3 Natural Scenes Dataset

The Natural Scenes Dataset (NSD; Figure 2.3; Allen et al., 2022) is one of the most extensive and high-resolution fMRI resources available for investigating visual perception. It includes

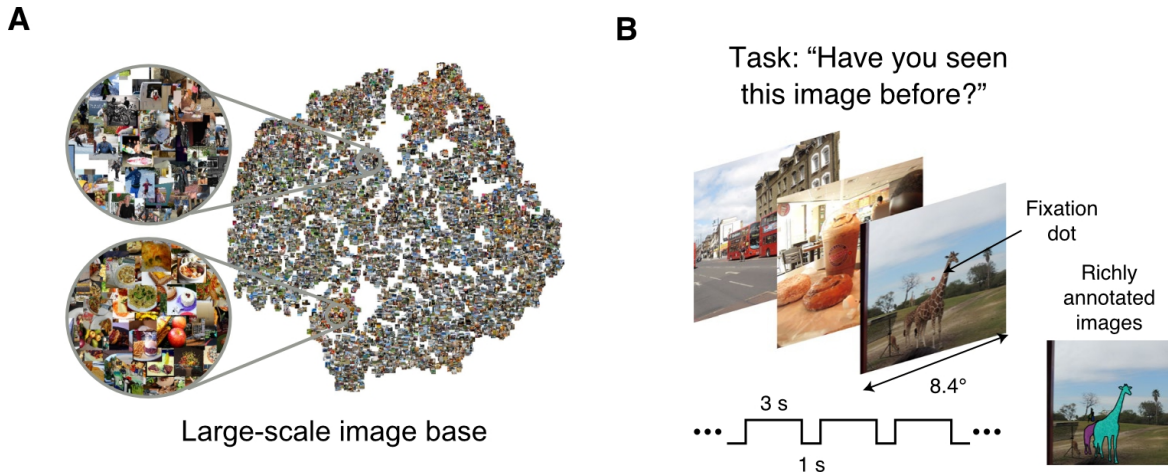


Fig. 2.3 Natural Scenes Dataset. (A) The large-scale colored natural image stimuli. (B) Schematics of image presentation experiment. While maintaining central fixation, participants viewed sequences of color natural scenes and judged whether each image had been previously shown at any point in the past. Adapted from Allen et al. (2021), used under CC BY 4.0 license.

ultra-high-field (7T) fMRI recordings from eight participants, each of whom viewed approximately 10,000 natural images across multiple sessions. With high signal-to-noise ratio (SNR) and extensive sampling of stimuli and neural responses, NSD enables detailed analysis of visual representations in the brain. Its scale, data quality, and experimental design make it a valuable benchmark for research at the intersection of cognitive neuroscience and artificial intelligence.

Subjects. Eight healthy subjects (two males and six females; age range, 19-32 years) with normal or corrected-to-normal vision took part in the experiment. Among them, Subjects 1, 2, 5, and 7 completed all scanning sessions.

Stimuli. The image stimuli were selected from the Microsoft Common Objects in Context (COCO) dataset (Lin et al., 2014). Each of the eight subjects viewed about 10,000 distinct natural images, comprising around 9,000 unique images per participant and a shared set of 1,000 images across all participants, resulting in a total of around 73,000 image stimuli. Within the shared set, a special subset of 100 images was selected to maximally span the semantic space. This sampling strategy was designed to maximize the number of distinct images in the dataset and enable cross-subject analyses of brain representations.

Experimental paradigm and data acquisition. Subjects fixated centrally and performed a long-term continuous recognition task on the images. Images were presented for 3 s with 1-s gaps in between images. Scanning was conducted at 7T using whole-brain gradient-echo EPI at 1.8-mm resolution and 1.6-s repetition time. fMRI signals were measured while

subjects each viewed 9,000-10,000 distinct natural scenes (22,000-30,000 trials) over the course of 30-40 scan sessions.

MRI data preprocessing. The fMRI data were pre-processed by performing one temporal interpolation (to correct for slice time differences) and one spatial interpolation (to correct for head motion). A general linear model was then used to estimate single-trial beta weights. Cortical surface reconstructions were generated using FreeSurfer, and both volume- and surface-based versions of the beta weights were created. The surface-derived ROIs and volume-derived ROIs are provided with the data. The VC used in this thesis was defined by the Human Connectome Project (HCP) multi-modal cortical parcellation (Glasser et al., 2016).

2.3.4 Illusion dataset

The Illusion dataset (Cheng et al., 2023) is a 3T visual fMRI dataset designed to investigate neural representations of subjective visual perception under visual illusions. It includes brain responses from seven subjects while viewing a natural images and illusory stimuli, such as illusory lines and neon color spreading patterns. The dataset enables the study of how illusory content is represented in the brain, distinct from the physical properties of the stimuli.

Subjects. Seven healthy subjects (four males and three females; age range, 25-36 years) with normal or corrected-to-normal vision participated in the experiment. Four subjects (Subjects 1-4) who were the same as those in previous studies (Horikawa & Kamitani, 2022; Shen et al., 2019b).

Stimuli. The natural image stimuli include a total of 3,200 images selected from multiple publicly available image databases. This includes 1,200 object images from 150 representative categories in ImageNet (Deng et al., 2009), with eight images per category, 1,000 material images from the Flickr Material Database (FMD; Sharan et al., 2014), and 1,000 object or scene images from the COCO dataset (Lin et al., 2014).

For illusory line images, six images were created using offset gratings to induce illusory lines. Each image was defined by three parameters: the orientation of the illusory line (0° , 45° , 90° , or 135°), the orientation of the inducer lines (0° or 90°), and the number of inducer lines (fixed at 19 for illusion images). Illusions occurred only when inducer and illusory orientations differed. Four additional control images were generated by reducing the number of inducer lines (to 9 or 3), which weakened or eliminated the illusion. Ten positive control images were also included, in which real lines were drawn at the perceived illusory locations.

For neon color spreading images, two classic configurations were used: Ehrenstein and Varin. For Ehrenstein, twelve images were prepared by varying line number (2 or 4) and disk size (small or large). Illusion strength was modulated via line width, and control images were

generated by thinning the black lines to disrupt contour continuity. For Varin, six images were created using four circular inducers with colored sectors and black Pacmans. Control images were made by removing the black inducers or altering disk composition. In both types, red was used for the colored components.

Experimental paradigm and data acquisition. The subjects fixated on a central fixation point while viewing each stimulus for 8 seconds as in the Deeprecon dataset. Functional brain activity was recorded using a 3T MRI scanner while each subject viewed 3200 training images, about 38 test images across approximately 8000-17000 trials.

MRI data preprocessing. The illusion fMRI data were preprocessed in the same way as in Deeprecon. The ROIs were defined following the same procedure used in Deeprecon.

2.3.5 Attention dataset

The Attention dataset (Horikawa & Kamitani, 2022) is a 3T visual fMRI dataset developed to study the neural mechanisms underlying selective attention in visual scenes. It contains brain responses from seven subjects who were instructed to attend to one of two superimposed images. Combined with standard image viewing trials, this dataset allows for the analysis of attention-modulated neural representations during overlapping visual input.

Subjects. Seven healthy subjects (four males and three females; age range, 25-36 years) with normal or corrected-to-normal vision participated in the experiment. Five subjects (Subjects 1-5) who were the same as those in previous studies (Ho et al., 2023; Horikawa & Kamitani, 2022; Shen et al., 2019b).

Stimuli. The training image stimuli consist of 1,200 natural images from 150 representative categories in ImageNet (2011 fall release; Deng et al., 2009), with eight images per category. These images were the same as those in the Deeprecon dataset. For the attention task, ten natural images that were not included in the training stimuli were used to construct the test stimuli. From these ten images, 45 unique superimposed image pairs were generated by spatially overlaying two different images at a time (excluding reversed combinations). Each of the ten images was also presented individually as single image viewing trials for baseline comparison.

Experimental paradigm and data acquisition. The training sessions are the same as the those in the Deeprecon dataset. For the test sessions, in each attention trial, subjects were presented with a sequence of images, each of which consisted of two successive cue images (2 s, 1 s for each cue) and spatially superimposed images of the two cue images (6 s), and were asked to attend to one image of a superposition of two images while ignoring the other. In each single-image trial, images were presented in the same way as the training session. Functional brain activity was recorded using a 3T MRI scanner while each subject

viewed 1,200 training images, 90 superimposed test images, and 10 single test images across approximately 7000 trials.

MRI data preprocessing. The attention fMRI data were preprocessed in the same way as in Deeprecon. The ROIs were defined following the same procedure used in Deeprecon.

2.3.6 DeepSoundRecon dataset

The DeepSoundRecon dataset (Figure 2.4; Park et al., 2023) is a 3T auditory fMRI dataset developed for studying auditory neural representations and sound reconstruction. It includes fMRI recordings from four healthy subjects when they listened to a wide range of naturalistic sound stimuli. The training and test stimuli were designed with distinct category structures, with mixed-category clips in training and clearly defined single-category clips in testing. This dataset offers a valuable resource for studying auditory neural representations.

Subjects. Five healthy subjects (four males and one female; mean age: 27.6 years) with normal hearing abilities participated in the experiment.

Stimuli. The sound stimuli were selected from the VGGsound test dataset (Chen et al., 2020). The training stimuli consisted of 1,200 audio clips, each containing multiple sound categories, while the test stimuli comprised 50 audio clips, each containing a representative sound category.

Experimental paradigm and data acquisition. The subjects were asked to listen to auditory stimuli and to click a button when two sequential blocks presented the same sound. Each sound stimulus was presented for 8 s within a stimulus block. Whole-brain functional MRI data were collected using a 3T scanner with 2-mm isotropic resolution and a repetition time of 2 seconds. fMRI signals were measured while subjects each listened to 1,250 sound stimuli (5,200 trials) over the course of 13-17 scan sessions.

MRI data preprocessing. The fMRI data was preprocessed with the fMRIPrep pipeline (Esteban et al., 2019). The preprocessed functional data were adjusted by shifting them forward 2 s to account for the hemodynamic delay. To increase the number of available data samples, a 4 s time window was slid across the original 8 s stimulus at 2 s intervals. For each of these 4 s sound stimuli, an fMRI sample was created by averaging the three consecutive functional volumes following the stimulus onset. The auditory ROIs was defined by the HCP multi-modal cortical parcellation (Glasser et al., 2016). There are thirteen anatomical ROIs identified in the early auditory cortex (A1, LBelt, MBelt, PBelt, RI) and auditory association cortex (A4, A5, TA2, STGa, STSd anterior, STSv anterior, STSd posterior, and STSv posterior). The whole Auditory cortex (AC) was defined as the combination of the early auditory cortex and auditory association cortex.

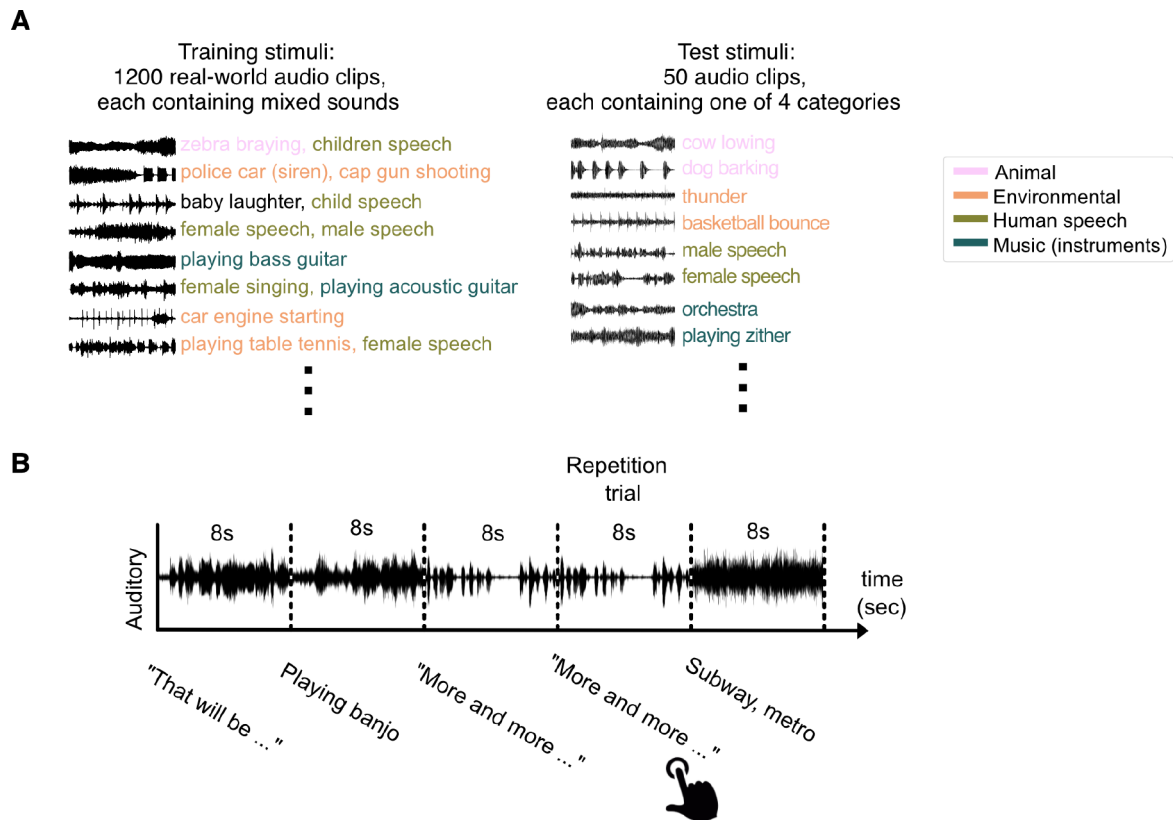


Fig. 2.4 DeepSoundRecon dataset. (A) Sound stimuli. The training dataset consisted of 1,200 clips (left panel), each containing a mixture of different sound categories that represent a wide array of natural auditory scenes. The test dataset consisted of 50 audio clips (right panel), each of which contained only one specific sound category. These test categories included human speech, animal sounds, musical instruments, and environmental sounds. (B) Schematic of the experimental design for the natural sound listening condition. Participants were continuously presented with 8-second sound stimuli without any intervening silent periods. To ensure participants' attention was uniformly distributed across all stimuli, they were instructed to perform a one-back repetition detection task. Participants were required to press a button whenever they detected consecutive trials featuring the same sound. Adapted from Park et al. (2023), used under CC BY 4.0 license.

2.3.7 Data availability

The above datasets are available publicly (Deeprcon dataset: <https://doi.org/10.6084/m9.figshare.17985578.v1>; THINGS dataset: <https://doi.org/10.18112/openneuro.ds004192.v1.0.5>; Natural Scene Dataset: <https://naturalscenesdataset.org>; DeepSoundRecon dataset: <https://doi.org/10.6084/m9.figshare.23633751.v9>; Illusion dataset: <https://doi.org/10.6084/m9.figshare.23590302.v2>; Attention dataset: <https://openneuro.org/datasets/ds004192/versions/1.0.5>).

2.4 Data for analysis

The fMRI datasets, including Deeprcon, NSD, THINGS, Illusion, Attention, and DeepSoundRecon, were subjected to analysis in the subsequent chapters. All available training samples from each subject were used for analysis, except where explicitly noted.

The primary analysis was performed using the Deeprcon dataset (Ho et al., 2023; Horikawa & Kamitani, 2022; Shen et al., 2019b). This dataset comprises five subjects, each with 6,000 training samples (1,200 natural images with five repetitions), 1,200 test samples (50 natural images with 24 repetitions), and an additional 800 test samples for generalization evaluation (40 artificial images with 20 repetitions).

The NSD and THINGS datasets were analyzed in Chapter 6. For the NSD, four subjects who completed all experimental sessions (Subjects 1, 2, 5, and 7) were included. Each subject was provided with 24,980 training samples (approximately 9,000 natural images with three repetitions) and 300 test samples (100 natural images with three repetitions). For the THINGS dataset, three subjects were included, each with 8,640 training samples (8,640 natural images with one repetition) and 1,200 test samples (100 natural images with 12 repetitions).

The Illusion and Attention datasets, along with the imagery data of the Deeprcon dataset, were analyzed in Chapter 10. For the Illusion dataset, four subjects (Subjects 1-4) who participated in previous studies (Horikawa & Kamitani, 2022; Shen et al., 2019b) and completed all experimental sessions were selected. Each subject had 16,000 training samples (3,200 natural images with 5 repetitions) and approximately 760 test samples, including 10 illusory line images, 10 real line images, 12 Ehrenstein images, and 6 Varin images, each repeated 20 times. For the Attention dataset, five subjects (Subjects 1-5) from previous studies (Ho et al., 2023; Horikawa & Kamitani, 2022; Shen et al., 2019b) were included. Each subject had 6,000 training samples (1,200 natural images with 5 repetitions) and 880 test samples, consisting of 10 natural images presented 16 times and 90 superimposed image pairs presented 8 times. For the imagery data, four subjects (Subjects 1, 2, 4, and 5) who

completed all imagery sessions were analyzed (Ho et al., 2023; Horikawa & Kamitani, 2022; Shen et al., 2019b). Each subject had 6,000 training samples (1,200 natural images with 5 repetitions) and 500 test samples, including 15 imagined artificial images and 10 imagined natural images, each presented 20 times.

The DeepSoundRecon dataset was used for the auditory analyses presented in Chapter 11. Four subjects (Subjects 2-5) were included, while Subject 1 (S1) was excluded due to the exploratory nature of their data (Park et al., 2023). Each subject was provided with 1,200 training audio stimuli and 50 test audio stimuli, with corresponding fMRI responses (four repetitions for each training clip and eight repetitions for each test clip), resulting 4800 training samples and 400 test samples.

Chapter 3

Functional alignment methods

3.1 Introduction

Revealing similarity and dissimilarity in brain data across individuals has long been a critical challenge in cognitive neuroscience. However, individual variability in the human brain poses a major challenge for comparing brain data across individuals. These differences are evident not only in brain anatomy but also in how functional regions are organized and activated (Cox & Savoy, 2003; Fischl et al., 2008; Guntupalli et al., 2016; Haxby et al., 2011; Van Essen, 2005, 2004). While anatomical alignment can mitigate anatomical difference, it cannot effectively handle fine-grained variability in functional topographies. Even when presented with identical stimuli, different individuals can exhibit significantly different brain activity patterns, severely hindering the identification of generalizable functional patterns at the group level.

To mitigate this individual difference, functional alignment (Chen et al., 2015; Guntupalli et al., 2016; Haxby et al., 2011, 2020; Ho et al., 2023; Yamada et al., 2015) methods have been proposed based on functional rather than structural information. These approaches directly align brain activity patterns or the neural representations they represent. For example, neural code converters (Ho et al., 2023; Yamada et al., 2015) learn subject-specific transformations that map brain activity from one individual to another, enabling the prediction of neural responses in the target brain. Hyperalignment (Chen et al., 2015; Guntupalli et al., 2016; Haxby et al., 2011, 2020) constructs a shared high-dimensional representational space in which individual brain activity patterns are projected, thus maximizing cross-individual similarity in representational geometry. These approaches offer effective tools for comparing neural representations across individuals.

This chapter introduces and discusses functional alignment methods, including hyperalignment, neural code converters, their variants, and latent space-based approaches. It aims

to provide a basic understanding of the principles of these methods and how they contribute to exploring neural representations across individuals and brain decoding models.

3.2 Hyperalignment

Hyperalignment is one of the classic models for functional alignment, introduced by Haxby et al. (2011, 2020). It estimates a shared representational space across individuals, allowing for comparison of brain activity patterns. A foundational assumption underlying hyperalignment is that the brain's response to a stimulus can be represented as a vector corresponding to the specific location in a high-dimensional space, with each dimension reflecting a local measurement or element of the brain response pattern (Haxby et al., 2014). When multiple stimuli are involved, these patterns form matrices that capture the brain's response to each stimulus as a geometric structure in the high-dimensional space. Each row of the matrices corresponds to the brain's response to a specific stimulus, and each column reflects activity in a specific brain region (such as an fMRI voxel). The geometric structure of these response patterns has been shown to be highly similar across individuals (Connolly et al., 2016, 2012b, 2012a; Diedrichsen & Kriegeskorte, 2017; Guntupalli et al., 2016; Kriegeskorte & Kievit, 2013). Prior to hyperalignment, these functional pattern vectors are instantiated in the spatially organized cortical topography of a given individual. Hyperalignment uses the inter-individual similarity in vector geometry to align brain activity patterns across subjects into a shared representational space, minimizing differences between corresponding vectors while preserving the geometric structure that reflects underlying neural representations (Figure 3.1).

Hyperalignment algorithm begins by selecting one subject as the initial reference template and aligning another subject's brain activity to this template using Procrustes transformation (Schönemann, 1966). After this alignment, the reference template is updated by averaging the aligned brain activity with the current template. This process continues iteratively: each subsequent subject's brain activity is aligned to the most recently updated template, and the template is further refined by incorporating the aligned brain activity of the new subjects. In the second step, each subject's original brain activity is aligned to the refined template (obtained after the first step). The aligned activity in this template space is averaged to create the finalized shared template. Finally, in the third step, each subject's brain activity is aligned to the final template, and an orthogonal transformation matrix is derived for each subject.

While the study of Haxby et al. (2011) is originally used to align brain data in the human ventral temporal (VT) cortex, hyperalignment has been extended to the whole cortex. This extension brings substantial computational challenges due to the drastically increased data dimensionality compared to region-specific alignment. There are two primary strategies to

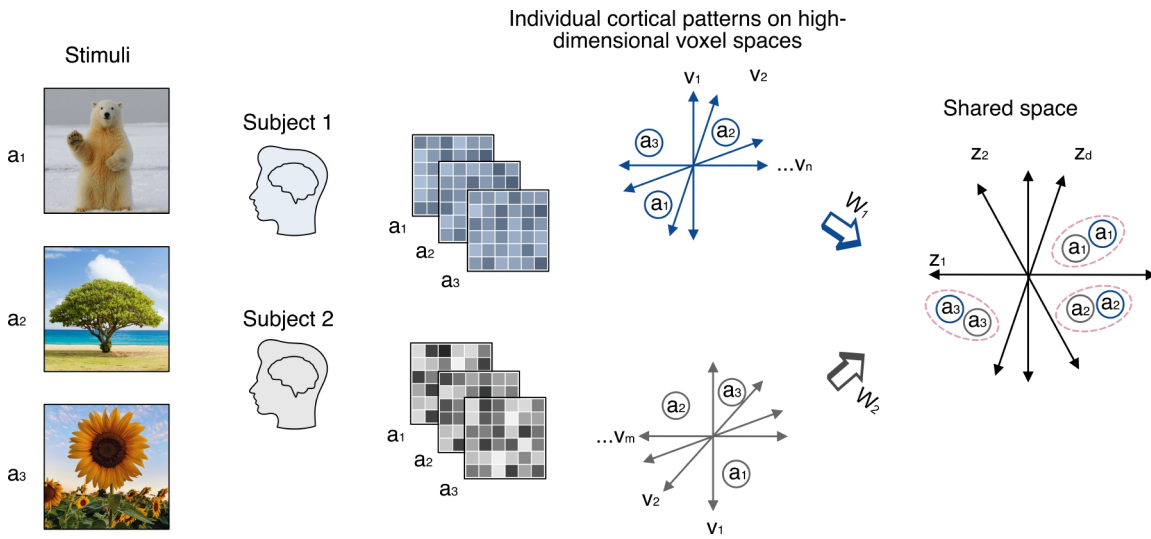


Fig. 3.1 Hyperalignment. The cortical responses of two subjects are in the high-dimensional individual brain spaces. Hyperalignment transforms cortical response vectors into a shared model space where patterns for the same brain state (corresponding to a given stimulus, condition, or task) are aligned across subjects.

perform functional alignment on the whole cortex: the searchlight (Guntupalli et al., 2016; Kriegeskorte et al., 2006) and the piecewise aggregation strategy (Bazeille et al., 2021). Both frameworks rely on first learning local functional transformation relationships within different small regions of the brain, and then integrating these local transformations into a large-scale alignment scheme that covers the whole cortex (Figure 3.2). The searchlight alignment strategy draws on an idea similar to that of a “sliding window,” where a large number of small, overlapping spherical regions with a fixed radius are defined in the cortex. The advantage of this approach is that the researcher does not have to rely on any prior functional or anatomical boundaries (e.g., such as those derived from parcellation atlases) and is able to explore in a data-driven manner. Within each spherical region, the hyperalignment algorithm is performed, and it learns a localized functional transformation. The final alignment is obtained by pooling the transformation results from all overlapping regions. A limitation of this approach is that the aggregated transformation may lose the properties (e.g., orthogonality, isometry) preserved in the local transformations. The aggregated transformation is also confined by searchlight size. Larger searchlight size enables accommodating more variation in individual topographies but reduces spatial specificity. The optimal searchlight size likely varies across brain regions and tasks, yet this remains underexplored.

The piecewise alignment strategy (Bazeille et al., 2019) is another way to align brain data across the whole cortex, which can also be regarded as a variant of hyperalignment. It divides the cortex into a set of non-overlapping regions or parcels. These parcels can be

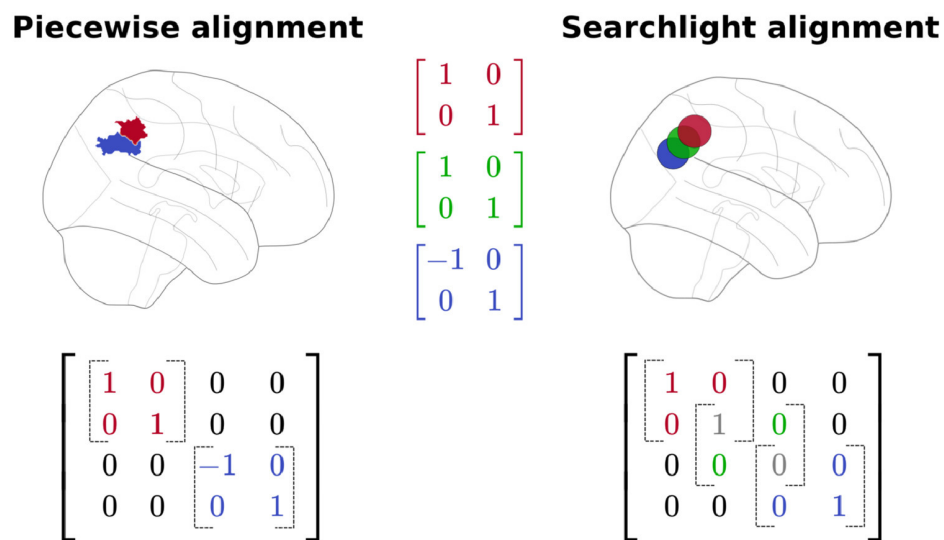


Fig. 3.2 Piecewise and searchlight alignment. Transformations are computed separately for the red, blue, and green regions. The green region appears only in the searchlight method, where it overlaps both red and blue areas. In the piecewise approach, the green region is not included because it corresponds to a searchlight overlapping both the red and blue areas. Piecewise transformations from non-overlapping parcels can be combined into a larger orthogonal matrix. In contrast, searchlight transformations are averaged in overlapping areas, resulting in a final transformation that is no longer strictly orthogonal. Adopted from Bazeille et al. (2019), used under CC BY 4.0 license.

delineated based on a priori anatomical or functional information, or they can be learned directly from the dataset using data-driven techniques like k-means clustering. Then, local functional transformation is calculated for each separate parcel. Finally, all individual parcel-specific transformations are concatenated to form a single transformation matrix covering the whole cortex. In contrast to the searchlight method, the aggregated transformation generated via this piecewise approach can preserve the desired mathematical regularities that were imposed locally. However, this approach relies on these partition boundaries (whether predefined or learned). This dependence can potentially introduce artificial discontinuities, sometimes referred to as “staircase effects,” or other functionally irrelevant changes into the final transformation, particularly at the borders between adjacent parcels.

In summary, hyperalignment methods offer powerful tools for neuroimaging research by addressing inter-individual variability in functional brain organization. They enable more precise investigations of shared neural representations at the group level. However, applying hyperalignment at the whole-brain (whole-cortex) level requires processing large-scale neural response matrices, resulting in significant computational demands. To manage these challenges across larger brain regions, approaches such as searchlight alignment and piecewise alignment have been developed. The searchlight method offers flexibility, but this may not guarantee the regularity of the final transformation. In contrast, the piecewise approach preserves the mathematical structure but may introduce artifacts related to the imposed parcel boundaries. The choice between these strategies should be guided by the specific research question, the nature of the data, and the trade-off between preserving transform properties and avoiding artificial discontinuities.

3.3 Neural code converter

A neural code converter is another functional alignment method, introduced by Yamada et al. (2015). The core idea of this method is to learn the statistical relationships between brain activity patterns evoked by pairs of subjects when presented with identical stimuli (Figure 3.3). Unlike hyperalignment, which estimates a shared representational space, the neural code converter is a pairwise alignment method that directly transforms brain activity patterns from one subject to another. This direct mapping enables seamless integration with downstream models trained on the target subject, providing greater flexibility.

In this pairwise alignment approach, both subjects are presented with an identical set of visual stimuli, which typically consists of random-patterned images in their study. The fMRI responses evoked in each subject by the same stimuli are then paired and used to train a series of linear regression models. Each model predicts the fMRI signal of a specific voxel

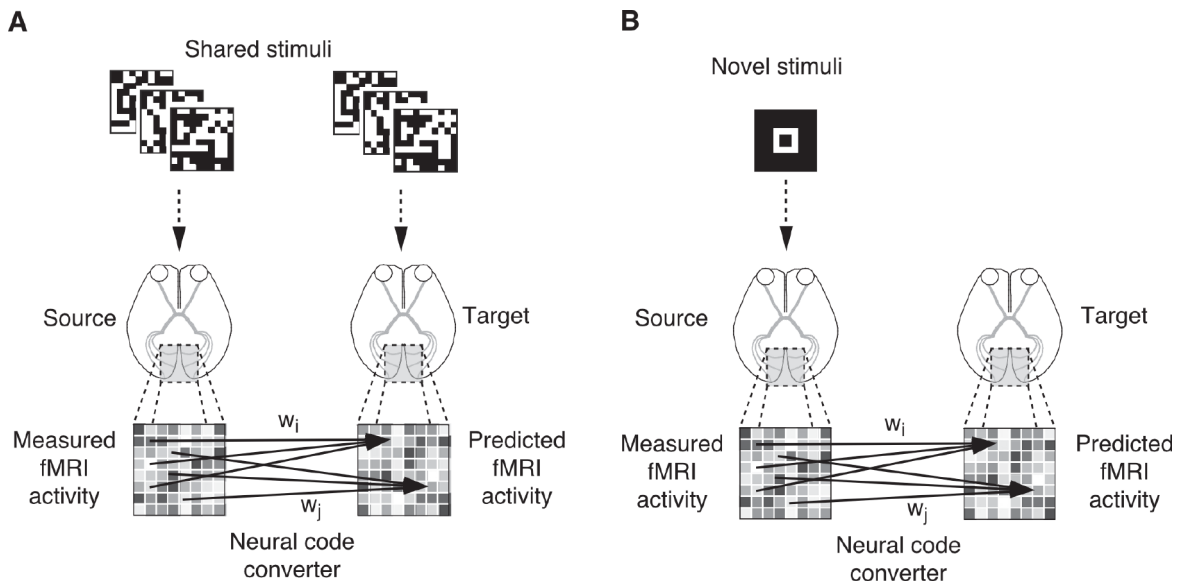


Fig. 3.3 Neural code converter. (A) A pair of subjects is presented with an identical (or shared) set of stimuli. A neural code converter model is then trained to predict the brain activity patterns of the target subject based on those of the source subject. (B) When a novel stimulus is presented to the source subject, the trained neural code converter predicts the corresponding brain activity pattern in the target subject, representing the same content. Adapted from Yamada et al. (2015), copyright 2015 with permission from Elsevier.

in the target subject, using the corresponding brain activity pattern from the source subject. Together, these linear regression models are the neural code converter. In the testing phase, the converter is applied to brain activity patterns evoked by new visual stimuli that were not used during training. This allows the brain activity of the source subject to be transformed into a predicted pattern for the target subject, representing the same content.

Although the original neural code converter relied on a specific linear regression model, its framework can accommodate other regression architectures as well. In the original study, Yamada et al. (2015) focus on the early visual area (V1), and the sample size of fMRI data used to train the model is often smaller than the number of voxels within that ROI, which tends to lead to overfitting problems. To mitigate this issue, they use sparse regression methods based on automatic relevance determination (ARD) a priori, which automatically selects a small number of voxels relevant to the prediction. This approach constrains the complexity of the model and improves the generalization ability of the model. Ho et al. (2023) extended the neural code converter method to the broader region (whole visual cortex), using a ridge regression model as the architecture. In contrast to ARD, which promotes sparsity by selecting only a subset of relevant voxels, ridge regression employs an L2 regularization that shrinks all coefficients but retains all features, making it more suitable when one assumes

most voxels contribute weakly to the prediction. Similar ridge regression-based approaches are also used in the cross-individual visual decoding analysis Ferrante et al., 2024 and language decoding analysis (Tang & Huth, 2025). Beyond linear models, researchers have also explored nonlinear architectures for neural code conversion, particularly in scenarios where the underlying relationships between brain activity and cognitive states are highly complex or noisy. For instance, Lu and Golomb (2023) introduced a U-Net-based nonlinear model for EEG signal conversion. Given the inherently low signal-to-noise ratio of EEG data, the U-Net’s hierarchical feature extraction and skip connections allow it to capture intricate nonlinear mappings across individuals. Their results demonstrate the effectiveness of nonlinear architectures in improving conversion performance for EEG signals. These approaches reflect the broader idea of the neural code converter framework. The notion of a neural code converter encompasses any model capable of capturing the statistical mapping of brain activity patterns across individuals, regardless of the specific architecture used.

3.4 Other alignment methods

While many functional alignment methods operate directly in the original brain activity space, such as classical hyperalignment and neural code converters, a growing number of derived approaches perform alignment through a latent space. These methods project individual brain data into shared latent spaces to reduce variability and enable cross-individual comparison. Although they differ in implementation, their core principle is consistent: aligning neural representations in a common representational space across individuals.

The SVD-Hyperalignment algorithm proposed by Chen et al. (2014), applies a joint Singular Value Decomposition (SVD) to response matrices across subjects in order to reduce the dimensionality of fMRI data prior to alignment. In this framework, dimensionality reduction and functional alignment are treated as two separate, sequential steps: the joint SVD extracts a shared low-dimensional feature space, and hyperalignment is subsequently performed within this reduced space. This approach improves computational efficiency and can enhance alignment robustness by focusing on the most informative components of brain activity. Further, the Shared Response Model (SRM; Chen et al., 2015) introduces a popular approach by combining dimensionality reduction and functional alignment into a joint optimization framework. SRM assumes that, during shared stimuli presentation, each subject’s brain activity is generated from a common set of latent variables via subject-specific linear projections. By jointly optimizing the shared latent response and subject-specific projection matrices, SRM effectively reduces the dimensionality of voxel-wise data while preserving shared stimulus-driven dynamics. Richard et al. (2019) similarly introduced the

fast version of SRM algorithm based on an intermediate atlas-based representation. These methods are particularly well-suited for tasks involving group-level representation learning, such as inter-individual decoding or functional clustering.

Gradient Hyeralignment (GHA; Xu et al., 2018) addresses both functional alignment and high-dimensionality in multi-subject fMRI data by combining Independent Component Analysis with stochastic gradient ascent. Unlike traditional approaches relying on closed-form transformations (e.g., Procrustes analysis), GHA learns subject-specific linear mappings that maximize statistical independence, iteratively updating a shared template based on the average of projected data. This enables scalable, dynamic alignment suited for large datasets. Regularized Canonical Correlation Analysis (rCCA; Bilenko and Gallant, 2016), offers an alternative by maximizing shared variance across subjects in a common low-dimensional space. It extends classical CCA with L2 regularization and supports both linear and nonlinear mappings, achieving robust alignment without requiring anatomical templates, and has been validated on tasks such as time-segment matching and movie classification. Manifold alignment methods (Gao et al., 2021; Nenning et al., 2015) assume individual brain activity lies on subject-specific low-dimensional manifolds and seek to align these for better cross-individual comparison. Gao et al. (2021) showed that nonlinear techniques like Diffusion Maps can reveal shared brain dynamics, while Nenning et al. (2015) proposed a two-step method using manifold alignment followed by coupled joint diagonalization to refine functional correspondence. These approaches are particularly powerful for capturing complex, nonlinear patterns in brain function that traditional linear methods may overlook.

Recently, deep learning-based methods have been proposed to model complex, nonlinear mappings of neural representations across subjects. Chen et al. (2016) introduced a Convolutional Autoencoder (CAE) for whole-brain functional alignment, treating fMRI volumes as 3D spatial data and using convolutional layers to learn hierarchical features that preserve local spatial structure. By minimizing reconstruction error, CAE captures shared functional patterns in a compact latent space while reducing subject-specific noise, without requiring anatomical normalization. Similarly, Deep Hyeralignment (DHA; Yousefnezhad and Zhang, 2017) extends the hyperalignment framework using subject-specific neural networks to learn nonlinear transformations into a common space. Combining stochastic gradient descent with rank-reduced SVD, DHA achieves scalable, subject-independent alignment and outperforms classical hyperalignment methods, especially on high-dimensional datasets.

Moving beyond alignment alone, recent approaches have explored multi-subject brain decoding by integrating functional alignment into representation learning, enabling more robust generalization across individuals. Chen et al. (2023a) introduced MinD-Vis, which implicitly aligns subject data by learning a shared sparse-coded latent space through self-

supervised masked modeling. They proposed Sparse-Coded Masked Brain Modeling (SC-MBM), a voxel-masking strategy inspired by masked autoencoders (MAE; He et al., 2021) and sparse coding, to extract context-aware fMRI features that guide a latent diffusion model for image reconstruction. Similarly, fMRI-PTE (Qian et al., 2023) pretrains a large-scale transformer encoder on surface-based fMRI using a reconstruction objective, then transfers it to downstream tasks such as semantic classification or brain-to-text/image mapping. While MinD-Vis focuses on image generation, fMRI-PTE builds a general-purpose encoder for flexible decoding. Related strategies are also used in MinD-Video (Chen et al., 2023b) to reconstruct videos from brain activity. By learning shared latent representations from large-scale data, these approaches reduce inter-subject variability and enable more robust, scalable multi-subject brain decoding.

As alignment strategies shift from unsupervised pretraining to brain decoding task-driven supervision, Zhou et al. (2024) introduced CLIP-MUSED, which integrates representation alignment and decoding in a unified model. The model introduces learnable subject-specific tokens into a Transformer encoder to disentangle individual variability while learning shared global representations. To encourage cross-individual consistency, it employs CLIP-derived representational similarity analysis (Kriegeskorte, 2008) as a regularization signal, aligning neural responses based on the topology of visual embeddings from the CLIP model (Radford et al., 2021). In parallel, the model is trained to predict semantic categories through a supervised classification loss applied to the shared representations. This joint optimization enables CLIP-MUSED to effectively align multi-subject data while preserving task-relevant information for downstream decoding. MindEye2 (Scotti et al., 2024) adopts a similar strategy by first using subject-specific ridge regression to project individual fMRI responses into a shared latent space (Figure 3.4). These aligned representations are then mapped to CLIP image embeddings through a unified deep network, which conditions a generative diffusion model for image reconstruction. Similar to CLIP-MUSED, the training process combines alignment with downstream supervision, optimizing both representational consistency and reconstruction quality across subjects. This design enables MindEye2 to achieve high-fidelity reconstruction using only limited fMRI data per subject, while supporting generalization to new individuals and stimuli. Similar strategies have been used in other multi-subject reconstruction methods (Bao et al., 2024; Gong et al., 2024; Quan et al., 2024).

Classical functional alignment methods, such as hyperalignment (Haxby et al., 2020) and neural code converters (Ho et al., 2023; Yamada et al., 2015), typically require shared stimuli across subjects to establish voxel- or pattern-level correspondence in representational space. This dependence on time-locked or stimulus-matched data poses challenges for broader applications, particularly in large-scale datasets across collection sites where such

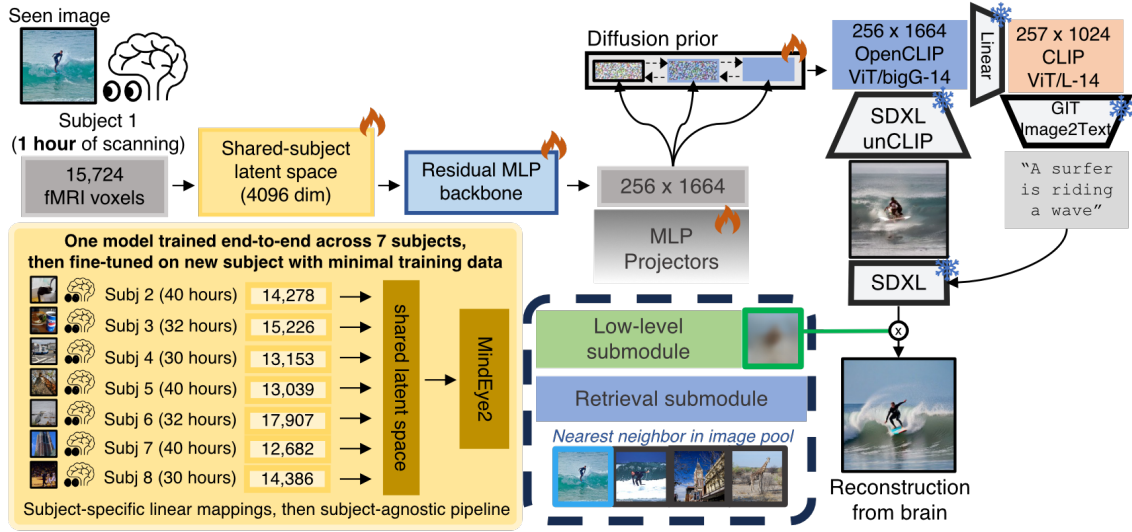


Fig. 3.4 The overview of the MindEye2. The model first projects individual fMRI responses into a shared latent space using subject-specific ridge regression. These aligned representations are then transformed into CLIP image embeddings, which subsequently guide image reconstruction via a generative diffusion model. Adopted from Scotti et al. (2024), used under CC BY 4.0 license.

requirement is often unavailable. Recent deep learning-based approaches have partially alleviated this constraint by performing alignment implicitly within decoding frameworks (Bao et al., 2024; Gong et al., 2024; Quan et al., 2024; Scotti et al., 2024; Zhou et al., 2024), where cross-individual correspondence emerges through task supervision such as classification or reconstruction. However, in these models, alignment is not treated as an explicit objective, and the structure of the resulting shared representational space remains underexplored.

3.5 Application of functional alignment

Understanding whether and how neural representations of visual information are shared across individuals is a fundamental goal in cognitive neuroscience. Despite significant anatomical and functional variability between brains, humans often perceive and interpret similar stimuli in highly consistent ways. Functional alignment techniques offer a powerful framework to investigate this shared mechanism by transforming individual brain activity patterns into a common representational space. This shared space enables researchers to examine the generalizability of neural representations across subjects, providing insight into population-level mechanisms of human perception. The central idea is to first train an



Fig. 3.5 The inter-individual reconstructed artificial images from the V1 area. Adopted from Yamada et al. (2015), copyright 2015 with permission from Elsevier.

encoding or decoding model using brain data from one subject in a shared space (which can be the subject's native brain space), and then apply it to another subject whose brain data have been aligned to the same space. If the model still has good performance on the new subject, this suggests that the neural representations it captures are, to some extent, shared across individuals.

Several studies have provided empirical support for this framework by demonstrating that visual features can be shared across individuals after functional alignment. The human brain encodes fundamental image properties such as contrast and spatial layout at the early stages of visual processing. Research on inter-individual visual image reconstruction has shown that these low-level features are decodable not only within individuals but also across individuals when their brain activity is appropriately aligned. For example, Miyawaki et al. (2008) first used local decoders to reconstruct contrast patterns from fMRI signals in V1, providing a tool to visualize the perceptual experience. Extending this work, Yamada et al. (2015) trained neural code converters using brain responses to random patches from two different subjects. They used neural code converters to align brain activity across subjects and found that image reconstructions based on converted brain activity closely resembled those based on the measured activity of the target subject (Figure 3.5). These findings suggest that contrast-related neural representations are, to some extent, shared across individuals.

The visual system is organized hierarchically, beginning with the processing of basic visual features in early areas such as the retina and primary visual cortex (V1), and progressing to higher-order regions that handle complex tasks like object recognition and spatial perception. This layered structure allows the brain to construct a coherent understanding of the visual world from simple sensory inputs, and has inspired the design of artificial vision systems. At higher levels of this hierarchy, the brain encodes increasingly abstract visual information, including object categories and identities. Haxby et al. (2011) applied hyperalignment to fMRI data collected while participants viewed a movie, enabling the

construction of a shared representational space across subjects. They aligned brain data of different subjects into this shared space, and trained object classifiers on aligned data from one participant and tested them on others. The classifiers achieved similar accuracy in within-subject and between-subject scenarios. This result demonstrates that representations of object identity, including human faces, animal faces, houses, and other objects, can be preserved across individuals after alignment. These findings support the view that high-level categorical representations are organized in a consistent manner across individual brains, particularly within regions such as the fusiform face area (FFA), lateral occipital cortex (LOC), and parahippocampal place area (PPA), which are known for their selectivity to specific object categories as shown in previous studies (Epstein & Kanwisher, 1998; Kanwisher et al., 1997; Kourtzi & Kanwisher, 2000).

A recent study by Ho et al. (2023) provides further evidence that hierarchical visual representations can also be shared across individuals using functional alignment. Following Yamada et al. (2015), they trained a neural code converter between pairs of subjects using fMRI activity patterns from the whole visual cortex, recorded while participants viewed the same natural image stimuli. They performed ablation analysis by excluding individual source visual areas from the converter model to assess their impact on prediction accuracy. The largest performance drops typically occurred when the excluded area matched the target region, showing a hierarchical correspondence. This result suggests that the model automatically learns to align brain patterns in a way that respects the hierarchical structure of the visual system without anatomical guidance. They also studied decoded DNN features from converted brain activity and identified the brain region that best predicted each DNN unit, as Nonaka et al. (2021) did. The distribution of top visual areas across DNN layers showed a hierarchical correspondence, where lower-level brain regions best predicted features from early DNN layers, while higher-level regions were more strongly associated with high-level layers. Image reconstructions using DNN features decoded from the converted brain activity also capture the main features of the natural image stimuli, including shapes and colors (Figure 3.6). These results support that the hierarchical visual representations can also be shared across individuals after neural code conversion.

In addition to hierarchical visual representations, the extent to which retinotopic organization is preserved after functional alignment has attracted considerable interest. Retinotopy refers to the orderly mapping of visual information from the retina onto the visual cortex, where neighboring points in the visual field correspond to neighboring regions in the brain. This organization, as detailed by Engel et al. (1994), is most clearly observed in the primary visual cortex (V1). Yamada et al. (2015) showed that neural code converters preserved both eccentricity and polar angle information when converting brain activities from one individual

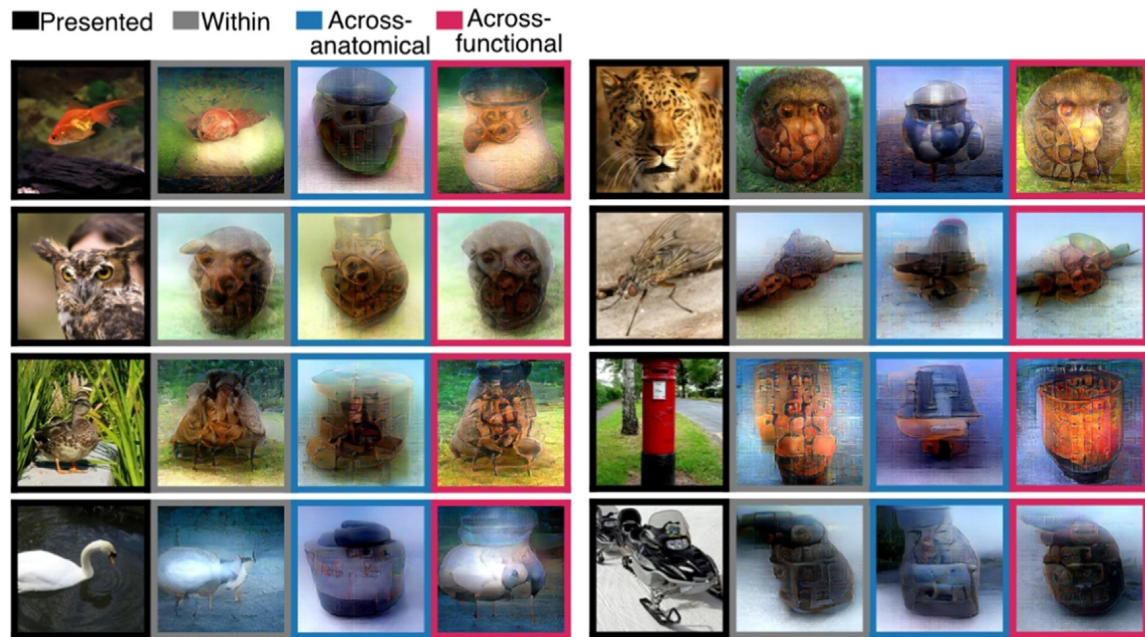


Fig. 3.6 The inter-individual reconstructed natural images from the whole visual cortex. Adopted from Ho et al. (2023), used under CC BY 4.0 license.

to another. They found that the transformation matrix gave higher weights to source voxels with similar eccentricity and polar angle as the target voxel. This indicates that when a stimulus moves across the source subject's visual field, the converted activity in the target brain reflects the expected retinotopic pattern at the corresponding cortical location. Subsequent studies by Bilenko and Gallant (2016) and Guntupalli et al. (2016) further demonstrated that retinotopic structure can be maintained across individuals after functional alignment beyond the primary visual cortex (V1). Feilong et al. (2023) also used the Individualized Neural Tuning (INT) model to examine whether visual field encoding, specifically eccentricity and polar angle, is consistent across individuals. They trained the INT model using brain activity collected during naturalistic movie viewing and used it to predict retinotopic maps based on leave-one-subject-out cross-validation. Predicted eccentricity and polar angle maps were compared to actual maps obtained from the experiment. The predicted maps closely matched the measured ones, with higher correlations than those between separate runs of the localizer. This suggests that the model preserved detailed spatial organization in early visual areas. These results suggest that not only the representational content of visual information but also its retinotopic organization are preserved after functional alignment.

While object recognition has been a central focus in visual neuroscience, real-world perception often involves understanding complex natural scenes filled with actions, relationships, and contextual cues. Natural scenes contain a wide range of semantic information that goes

beyond simply identifying objects. Understanding such scenes involves interpreting actions, relationships, and the broader context in which events occur. For instance, perceiving a classroom scene includes identifying students and desks, but also understanding the ongoing activity, the teacher-student interaction, and the educational setting implied by the environment. These types of semantic information are represented not in a single brain region, but across widespread neural systems, reflecting how the brain integrates different elements to form a coherent understanding of complex situations (Huth et al., 2016). Van Uden et al. (2018) investigated how to build semantic encoding models that generalize across individuals. Participants watched a natural movie while each time point was labeled with word embeddings describing semantic content. They trained three types of models: individual-specific models, between-subject models using anatomical alignment, and between-subject models using hyperalignment. To assess performance, they predicted brain activity for held-out movie segments and calculated the correlation between predicted and actual voxel responses. Anatomical alignment led to lower prediction accuracy and more spatially diffuse tuning, with broad regions showing similar semantic selectivity, making it hard to tell which voxels were selective for specific semantic information. In contrast, hyperalignment produced sharper, more localized tuning maps that closely resembled the within-subject models. Moreover, models using hyperalignment achieved higher voxel-wise prediction scores across most of the cortex, particularly in temporal and parietal regions, indicating better preservation of individual semantic encoding patterns across brains. Their study demonstrated that functional alignment enables semantic models to generalize across subjects while preserving semantic representations.

Semantic information is represented not only in visual processing but also in language comprehension. Semantic representations reflect the abstract concepts and relationships that underlie both what we see and what we hear. These representations are not limited to a single sensory modality and are distributed across higher-level brain regions. Recent studies have shown that semantic information can be decoded from both visual and auditory inputs, including spoken stories, silent videos, and even imagined language. Tang et al. (2023) developed a method to reconstruct continuous, semantic language from fMRI data. Their approach combined an encoding model that mapped semantic features to brain activity with a decoding strategy that selected language sequences consistent with observed neural signals. The reconstructed texts closely matched the semantic information of the original stimulus, and the method generalized to imagined speech and nonverbal stimuli. However, decoding models were highly participant-specific and failed to generalize across individuals directly, highlighting the need for individualized training. Based on this language decoding work, Tang and Huth (2025) further investigated whether semantic decoders trained on

one individual could be transferred to another using functional alignment, and whether this alignment could be achieved through nonverbal stimuli, such as silent movies. Using a two-step process, they first trained linear neural code converters to project brain activity from a target subject into the space of a reference subject, based on paired brain responses to either shared spoken stories or shared silent films. A semantic decoder trained on the reference brain was then applied to the aligned activity patterns of the target subject. Their results demonstrated that cross-individual decoder transfer was feasible. Decoders trained on one subject could reconstruct semantic word sequences from another's brain activity, with semantic similarity scores significantly above baseline. In addition, converters trained on silent movies performed as well as those trained on language input, suggesting that nonverbal stimuli can provide a sufficient basis for semantic alignment. Moreover, the approach remained robust across cortical regions and even under simulated cortical lesions. Cross-individual decoders also outperformed within-subject decoders when training data for the target participant was limited. These findings show that semantic representations are not only stable across modalities but also widely shared across individuals.

3.6 Summary

This chapter reviewed a range of functional alignment methods aimed at addressing individual variability in brain function by transforming brain activity into a shared representational space. These approaches allow researchers to compare neural representations across individuals and support generalizable brain models.

Classical models such as hyperalignment and neural code converters offer foundational methods for aligning brain activity, either by preserving geometric relationships in representational space or directly mapping activity between subjects. Both have been extended from early visual areas to the whole brain, demonstrating reliable alignment of object categories, image features, and spatial layouts. Other alignment methods adopt a common strategy of projecting individual brain activity into a shared latent space, using techniques such as dimensionality reduction or deep representation learning. These approaches reduce inter-subject variability and support flexible alignment, often integrated with downstream tasks such as decoding or reconstruction.

Empirical applications across vision and language have illustrated the use of functional alignment in capturing shared neural representations. Not only low-level visual features (e.g., contrast and retinotopy) but also high-level semantic content can be preserved and decoded across individuals after alignment. Moreover, semantic language alignment can be achieved even using nonverbal stimuli, such as silent movies. These results suggest that

functional alignment helps reveal shared neural representation and provides practical tools for enhancing the generalizability of brain decoding models.

While these methods have shown promising results, they may also have limitations. Classical alignment methods, such as hyperalignment and neural code converters, provide interpretable frameworks for aligning brain activity but typically rely on shared stimuli, limiting their use in naturalistic or clinical settings where shared stimuli is often unavailable. In contrast, recent deep learning approaches embed alignment within decoding models, enabling subject data to be aligned without shared stimuli. However, the alignment is often not an explicit objective in these models, and the resulting shared spaces tend to be less interpretable. Future research should integrate the strengths of both paradigms: combining the interpretability of classical methods with the flexibility of deep learning may yield more robust and generalizable alignment models that support scalable brain decoding.

Chapter 4

Within-individual brain decoding

4.1 Introduction

Decoding perceptual and cognitive states from brain activity is an important goal in cognitive neuroscience and brain-machine interface research. A standard approach is within-individual brain decoding, which aims to infer mental representations by building models based solely on data from the same individual. Unlike across-individual decoding methods that require normalization or alignment to address inter-individual variability in brain structure and function, within-individual decoding directly uses data from the same individual, leveraging functional architecture that is stable within the individual but idiosyncratic across individuals. This approach usually enables more accurate and personalized inference, providing insights into how perceptual experiences are represented in the brain at the individual level.

Recent advances in deep neural networks (DNNs) have further enhanced the within-individual brain decoding approach. DNNs trained on large-scale visual datasets extract hierarchical feature representations that closely resemble the organization of the human visual cortex, making them effective intermediaries between stimulus features and brain activation patterns. Integrating these powerful representations into brain decoding enables not only the decoding of object information but also the reconstruction of visual images from brain activity.

This chapter reviews DNN representations and within-individual brain decoding. Section 4.2 introduces the DNN architectures used as feature extractors, highlighting their hierarchical structure and relevance to visual processing. Section 4.3 reviews brain decoding and visual image reconstruction methods. Section 4.4 outlines our within-individual decoding setup, detailing the selection of DNNs, decoding models, and reconstruction pipelines. These components establish a framework for decoding within-individual visual experiences and

provide a methodological foundation for extending decoding analyses across individuals in the following chapters.

4.2 DNN architectures and content representations

Deep neural networks have transformed artificial intelligence by enabling effective extraction of meaningful representations from complex data. Their architectural design plays a central role in shaping the scale, abstraction level, and richness of learned features. In this subsection, we provide an overview of representative DNN architectures and discuss the key principles underlying their design and the resulting content representations.

DNNs are inspired by the structure and function of biological neural systems. Early models such as the Perceptron, proposed by Rosenblatt (1958), laid the theoretical foundation for neural network research. Progress slowed during the 1970s and 1980s, despite important developments like the Neocognitron (Fukushima, 1980) and LeNet (LeCun et al., 1989, 1998), which introduced key ideas later used in convolutional neural networks. The resurgence of interest in the 2010s was driven by the availability of large-scale labeled datasets, increased computational power, and advances in training algorithms. A major turning point came with the success of AlexNet (Krizhevsky et al., 2012) in the 2012 ImageNet challenge, which demonstrated the practical potential of deep convolutional networks for image recognition and accelerated progress in deep learning research.

Following the success of AlexNet, convolutional neural networks (CNNs) have emerged as the dominant architecture for image representation and analysis. A typical CNN consists of an input layer for receiving image data, multiple hidden layers for hierarchical feature extraction, and an output layer for classification or prediction. The hidden layers are composed of several essential components, including convolutional layers, pooling layers, fully connected layers, and activation functions. Convolutional layers form the backbone of CNNs, responsible for detecting local patterns in the input image. These layers apply a set of learnable filters (also known as kernels) that slide across the input image, performing convolution operations to detect local patterns such as edges, textures, and shapes in the lower layers, and more complex, abstract features in the deeper layers. Pooling layers, such as max pooling and average pooling, follow convolutional layers to downsample feature maps, reducing spatial dimensions and computational costs while imparting translation invariance. Finally, fully connected layers are usually placed at the end of the network, where each unit connects to all activations from the previous layer. They learn high-level abstract features and produce the final feature vector for the input image, which can be used to classify it. These foundational principles established a solid basis for subsequent advancements, driving

researchers to investigate deeper, wider, and more efficient CNN architectures for enhanced feature learning and representation.

One major direction of architectural advancements focused on increasing network depth to improve representation capacity and model performance. VGGNet (Simonyan & Zisserman, 2014) demonstrated this approach, particularly through its variants VGG-16 and VGG-19 (with 16 and 19 layers, respectively). By systematically stacking multiple 3×3 convolutional filters within a highly uniform and straightforward architecture, VGG demonstrated that increasing network depth could significantly enhance feature extraction capabilities and classification accuracy. However, the network's heavy reliance on large fully connected layers resulted in a substantial number of parameters, leading to increased computational and memory demands. ResNet (He et al., 2016) addressed the difficulty of training very deep networks by introducing residual blocks and skip connections. These designs allowed networks to learn identity mappings more easily, enabling effective training of extremely deep models without degradation in performance. ResNet variants, such as ResNet-50 (50 layers), ResNet-101 (101 layers), and ResNet-152 (152 layers), successfully pushed the boundaries of network depth.

While VGGNet and ResNet primarily focused on exploring greater depth, GoogLeNet (Szegedy et al., 2014), also known as Inception, increased both the depth and width of convolutional networks, with a particular emphasis on multi-scale feature extraction through a widened architecture. GoogLeNet exemplified this approach by introducing the Inception module, which performs parallel convolutional operations with different filter sizes (1×1 , 3×3 , and 5×5) within the same layer. This design enabled the network to capture features at multiple scales while maintaining computational efficiency. To further control model complexity, 1×1 convolutions were heavily utilized for dimension reduction, allowing a significant expansion in width without incurring prohibitive computational costs. With its 22 parameterized layers, GoogLeNet achieved state-of-the-art performance in the ImageNet 2014 competition while using substantially fewer parameters than previous deep architectures.

Another important direction focused on improving the efficiency of convolutional networks. DenseNet (Huang et al., 2018) embodied this idea by introducing dense connectivity, where each layer receives the feature maps of all preceding layers as input. This design promotes feature reuse, strengthens gradient propagation, and alleviates the vanishing gradient problem, allowing DenseNet to achieve high parameter efficiency and competitive accuracy with fewer parameters compared to traditional deep networks. EfficientNet (Tan & Le, 2020) proposed a systematic model scaling approach that simultaneously balances network depth, width, and input resolution through a compound scaling method to uniformly scale all three dimensions with a set of fixed coefficients. It is built upon a baseline network designed by

neural architecture search and achieves significantly higher computational and parameter efficiency compared to previous architectures. These networks reflect a broader shift toward holistic network design strategies, moving beyond simply making models deeper or wider.

To achieve more scalable and flexible representations, Transformer-based architectures have been introduced into vision tasks, replacing the traditional convolutional paradigm. Vision Transformer (ViT; Dosovitskiy et al., 2021) introduced a novel approach by applying Transformer models, initially developed for natural language processing, directly to image analysis. In ViT, an input image is divided into fixed-size patches, which are treated as tokens and processed through a standard Transformer encoder. This design leverages self-attention mechanisms to capture global dependencies across the entire image, replacing the local receptive fields of convolutional layers. With sufficient data and computational resources, ViT demonstrated that Transformer-based models could surpass traditional CNNs in image classification performance. Following this, Contrastive Language-Image Pretraining (CLIP; Radford et al., 2021) extended the Transformer framework into a multi-modal setting. CLIP jointly trains an image encoder and a text encoder to map images and natural language descriptions into a shared embedding space using a contrastive learning objective. This design enables zero-shot transfer to a wide range of vision tasks without task-specific fine-tuning, representing a significant step toward more generalized and scalable visual understanding systems.

Similar to vision tasks, early neural network models for audio processing largely drew inspiration from convolutional architectures originally developed for images. Convolutional neural networks were successfully adapted to handle time-frequency representations such as spectrograms, treating them as two-dimensional inputs analogous to images. Several key models include SoundNet (Aytar et al., 2016), which transferred visual representations to the audio domain, and VGGish (Hershey et al., 2017) and VGGish-ish (Iashin & Rahtu, 2021), both of which are VGG-style networks adapted for large-scale audio classification tasks. To further capture the temporal dynamics inherent to audio signals, Recurrent Neural Networks (RNNs) (Rumelhart et al., 1986) and Long Short-Term Memory (LSTM) networks (Hochreiter & Schmidhuber, 1997) were incorporated alongside CNNs in hybrid architectures. More recently, Transformer-based models such as the Audio Spectrogram Transformer (Gong et al., 2021) have been introduced, applying self-attention mechanisms to spectrogram patches in a manner similar to Vision Transformers, leading to improved performance across a range of audio recognition tasks.

Motivated by their strong representational capabilities, DNNs have attracted increasing interest as models for biological visual processing. In particular, CNNs have been proposed to capture key organizational principles of the primate ventral visual stream. Classic architec-

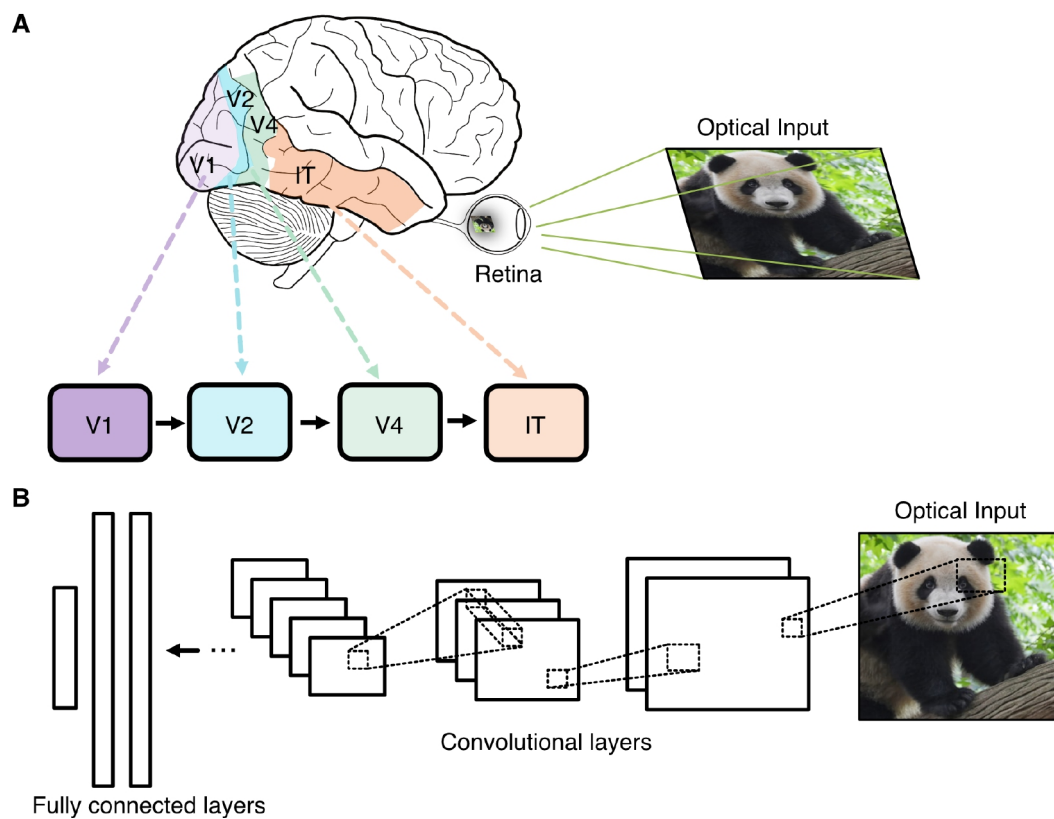


Fig. 4.1 Hierarchical processing in the human visual system and CNN. (A) The hierarchical processing of the ventral visual stream progresses from early visual areas like V1 to V2 and V4 and, finally, to high-level regions such as the inferior temporal (IT) cortex. (B) The hierarchical processing of the CNN progresses from low-level convolutional layers to high-level fully connected layers.

tures such as AlexNet (Krizhevsky et al., 2012), VGGNet (Simonyan & Zisserman, 2014), and ResNet (He et al., 2016) implement a hierarchical, layer-by-layer feature extraction process that mirrors the progression from early visual areas (e.g., V1, V2, V4) to high-level regions such as the inferior temporal (IT) cortex (Figure 4.1). In this hierarchy, early CNN layers detect simple features like edges and colors, similar to V1 neurons, while deeper layers encode increasingly complex patterns and semantic representations, resembling those observed in higher visual areas.

Motivated by their strong representational capabilities, DNNs have attracted increasing interest as models for biological visual processing. In particular, CNNs have been proposed to capture key organizational principles of the primate visual cortex. Classic architectures such as AlexNet (Krizhevsky et al., 2012), VGGNet (Simonyan & Zisserman, 2014), and ResNet (He et al., 2016) use a hierarchical, layer-by-layer feature extraction process inspired by the structure of the ventral visual stream, progressing from early visual areas like V1 to V2 and V4, and finally to high-level regions such as the inferior temporal (IT) cortex (Figure 4.1). In CNNs, early layers detect simple features (e.g., edges, colors), analogous to V1 neurons, while deeper layers progressively capture more complex patterns and semantic structures, resembling representations found in higher visual areas.

Human neuroimaging studies have revealed a striking parallel between the hierarchical processing in CNNs and the human visual system. Activation patterns in early CNN layers show strong correspondence with activity in early visual areas, such as V1 and V2, while activations in deeper layers align with higher-level ventral stream regions, including the lateral occipital cortex (LOC) and inferior temporal (IT) cortex. Using representational similarity analysis (RSA) (Kriegeskorte, 2008), Khaligh-Razavi and Kriegeskorte (2014) and Cichy et al. (2016) demonstrated that the representational geometry learned by deep supervised CNNs closely matches the structure of human fMRI activity patterns, with predictive power approaching the estimated noise ceiling in certain visual areas. Encoding model approaches (Güçlü & Van Gerven, 2015) have similarly revealed hierarchical correspondences between CNN activations and human cortical responses. Furthermore, decoding studies have provided additional evidence for this hierarchical relationship. Horikawa and Kamitani (2017) showed that features extracted from different CNN layers contributed differentially to decoding performance across brain areas, with lower CNN layers better aligning with early visual areas and higher CNN layers corresponding to higher-order regions such as the LOC.

Beyond human neuroimaging findings, similar correspondences have been observed at the single-neuron level in non-human primates. Neurophysiological studies have shown that units in CNNs optimized for object recognition account for a significant portion of the explainable variance in mid-level (V4) and high-level (IT) neuronal responses (Cadieu et al.,

2014; Yamins et al., 2014), with performance-optimized CNNs predicting approximately 50-60% of the variance. Although early computational models captured relatively little neural variance, these findings suggest that deep CNNs represent a significant advance in building biologically relevant models of visual processing.

In our study, we use the hierarchical feature representations learned by DNNs to decode perceptual experiences from brain activity patterns. We select DNN architectures that have demonstrated a strong correspondence to brain representations in the visual cortex, such as VGGNet (Simonyan & Zisserman, 2014), VGGish-ish (Iashin & Rahtu, 2021), and AlexNet (Krizhevsky et al., 2012). In addition, we include features extracted from CLIP (Radford et al., 2021), a model based on a fundamentally different architecture, for supplementary analysis. By using content representations from these neural networks, we aim to capture the neural representations that underlie perceptual experiences across individuals.

4.3 Brain decoding and visual image reconstruction

Brain decoding refers to the process of inferring a person's perceptions or thoughts from measurements of their brain activity, offering a valuable approach for investigating the neural representations underlying cognition and perception. Within-individual decoding involves training and testing models on data collected from the same individual, using the consistency of that person's brain responses to infer perceptual or cognitive states. This approach leverages subject-specific functional organization, thereby avoiding the anatomical and functional variability that typically arises across different individuals. As a result, within-individual decoders generally achieve higher accuracy compared to across-individual decoders, which must contend with variations in cortical folding, voxel correspondence, and neural tuning. This subsection reviews key methods for within-individual decoding and stimulus reconstruction from brain activity patterns.

4.3.1 Brain decoding

The success of brain decoding critically depends on the analytical techniques used to interpret brain signals. Traditional univariate analyses of brain imaging data focus on identifying localized changes in activation across individual voxels, which have proven valuable for mapping functional regions. However, these methods often overlook the rich, distributed patterns of activity that encode complex cognitive and perceptual states (Norman et al., 2006). To address this limitation, multi-voxel pattern analysis (MVPA) was introduced, offering a framework for detecting spatially distributed information in brain signals (Haxby

et al., 2001). Haxby et al. (2001) demonstrated that object categories such as faces and houses evoke overlapping but distinguishable activation patterns in the ventral temporal cortex. Even after removing voxels with maximal selectivity for a category, the remaining distributed pattern still enabled accurate classification of the viewed object. MVPA represents a critical foundation for modern brain decoding approaches, enabling researchers to infer perceptual, cognitive, or emotional states from brain activity with higher sensitivity (Haynes & Rees, 2006). Building on MVPA principles, more sophisticated decoding models have been developed, ranging from simple classifiers to advanced stimulus reconstruction techniques (Horikawa & Kamitani, 2017; Kamitani & Tong, 2005; Shen et al., 2019b).

An early demonstration of brain decoding using MVPA was presented by Kamitani and Tong (2005), who showed that fine-grained visual information, such as the orientation of a grating stimulus, could be reliably decoded from fMRI activity patterns in early visual areas. They trained a linear support vector machine as a classifier on brain activity patterns and found that ensemble patterns across voxels in areas V1-V4 contained sufficient information to accurately predict the specific orientation being viewed by the subject. Their study introduced the concept of “ensemble feature selectivity,” highlighting that weak biases in individual voxels could be pooled together to extract meaningful feature information. Furthermore, they demonstrated that attention could modulate these patterns: when subjects attended to one of two overlapping gratings, the ensemble activity shifted to reflect the attended orientation, allowing for successful “mind-reading” of subjective perceptual states. This work provided critical empirical evidence that MVPA can uncover fine-scale feature representations and subjective experiences from non-invasive brain imaging data, establishing a foundation for modern decoding and stimulus reconstruction studies.

Building on this foundation, decoding techniques rapidly progressed from simple feature discrimination to the identification of complex naturalistic stimuli. Kay et al. (2008) introduced a decoding framework that predicted voxel responses to visual stimuli using quantitative receptive-field models. By training voxel-wise models on responses to natural images and matching predicted patterns to measured brain activity, they successfully identified which novel natural image subjects had seen from a large candidate set. This approach demonstrated that fine-grained tuning to spatial position, orientation, and spatial frequency in early visual areas carries sufficient information to decode complex visual experiences, moving beyond simple category-level decoding and towards the identification of rich, naturalistic percepts.

Beyond decoding externally presented stimuli, researchers have increasingly explored the possibility of inferring internal cognitive states from brain activity. Harrison and Tong (2009) demonstrated that the contents of visual working memory could be reliably decoded from early visual areas, including V1 through V4. Using pattern classification techniques on fMRI

data collected during delay periods, they showed that even when overall activation levels dropped to near baseline, the specific orientation held in memory could be predicted with high accuracy. This finding challenged the traditional view that early visual areas only process incoming sensory information, suggesting instead that these regions can also maintain fine-grained feature representations over extended periods to support higher cognitive functions.

Haynes et al. (2007) provided direct empirical evidence that hidden intentions could be decoded from distributed activation patterns in medial and lateral prefrontal cortex. Using a delayed intention task combined with multivariate pattern classification, they demonstrated that participants' freely chosen intentions to perform addition or subtraction could be predicted with above-chance accuracy during the preparatory delay period. This study revealed that prospective goals are spatially encoded in the prefrontal cortex and laid foundational work for decoding high-level cognitive states beyond simple perceptual experiences. Soon et al. (2008) extended this line of research by showing that free decisions can be predicted from brain activity several seconds before reaching conscious awareness. Using a free-choice task combined with multi-voxel pattern analysis of fMRI data, they demonstrated that patterns in the frontopolar cortex and precuneus encoded upcoming left or right button choices up to ten seconds in advance. Their findings suggest that unconscious processes in high-level control regions shape volitional decisions long before they enter subjective experience.

Beyond cognitive states such as memory contents, intentions, and free choices, Horikawa et al. (2013) achieved decoding of internally generated visual imagery during sleep. By training classifiers on brain activity evoked by visual stimuli during wakefulness and applying these decoders to fMRI patterns preceding awakening, they accurately predicted the visual content reported by participants from their dreams. Their findings provided compelling evidence that dream imagery shares neural representations with visual perception and demonstrated that even spontaneous mental content during sleep can be objectively decoded using multi-voxel brain patterns.

Motivated by advances in DNN representations, Horikawa and Kamitani (2017) developed a generic decoding approach capable of identifying arbitrary seen and imagined object categories from brain activity (Figure 4.2). Leveraging hierarchical visual features extracted from deep convolutional neural networks, they trained decoders to predict visual representations for presented images from brain activity patterns. This enabled accurate identification of both seen and imagined object categories, including categories not encountered during training, thus achieving a major step towards generalized brain decoding. Their results also revealed a close correspondence between the hierarchical structure of visual feature representations in artificial neural networks and the organization of human visual cortex,

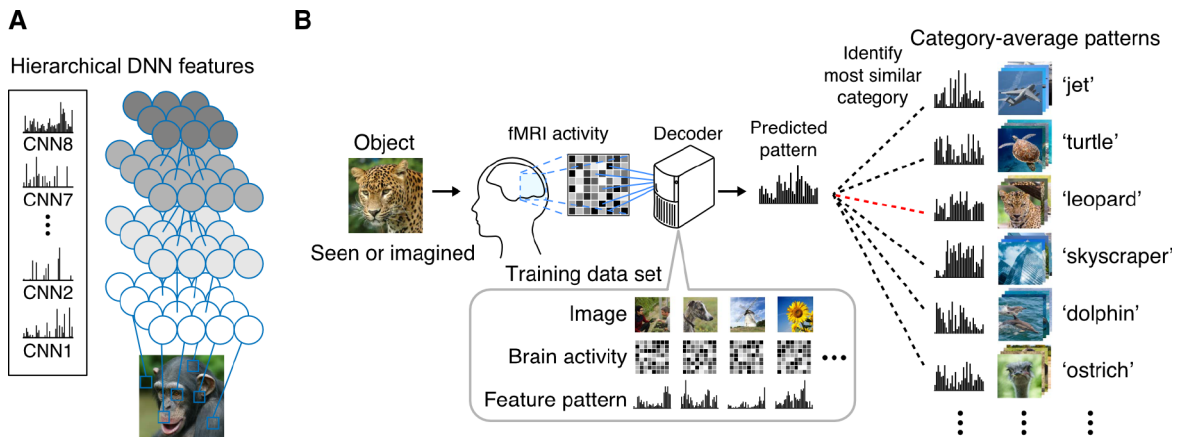


Fig. 4.2 The schematic of generic object decoding using DNN representations. Adapted from Horikawa and Kamitani (2017), used under CC BY 4.0 license.

highlighting the potential of DNN feature-based methods for advancing brain decoding technologies.

4.3.2 Visual image reconstruction

Reconstruction of visual images from brain activity constitutes a major milestone in brain decoding research. It seeks to reproduce the visual experiences of individuals using only their brain activity patterns. Early attempts at image reconstruction used relatively simple techniques and stimuli. For example, Miyawaki et al. (2008) showed subjects random binary patterns and trained a linear model to reconstruct those 10×10 pixel patterns from fMRI signals in the visual cortex (Figure 4.3). In their approach, the visual images were decomposed into multiple local elements at different spatial scales (1×1 , 1×2 , 2×1 , and 2×2 patches). For each local patch, a dedicated decoder was trained using sparse multinomial logistic regression, automatically selecting the most informative voxels from early visual areas (V1 and V2) without relying on explicit retinotopic mapping. By optimally combining the outputs of all local decoders, they successfully reconstructed the full image from brain activity. Their method allowed accurate reconstruction of both random patterns and novel structured shapes, such as geometric figures and alphabet letters, even though these structured images were never used during training. This study provided strong evidence that complex visual information is distributed across multi-voxel patterns in the brain and that multiscale representations are crucial for achieving visual image reconstruction.

Extending from simple stimuli to complex natural images, Naselaris et al. (2009) developed a Bayesian reconstruction framework that integrated multiple sources of information about brain responses and image structure. Their approach first involved building two sepa-

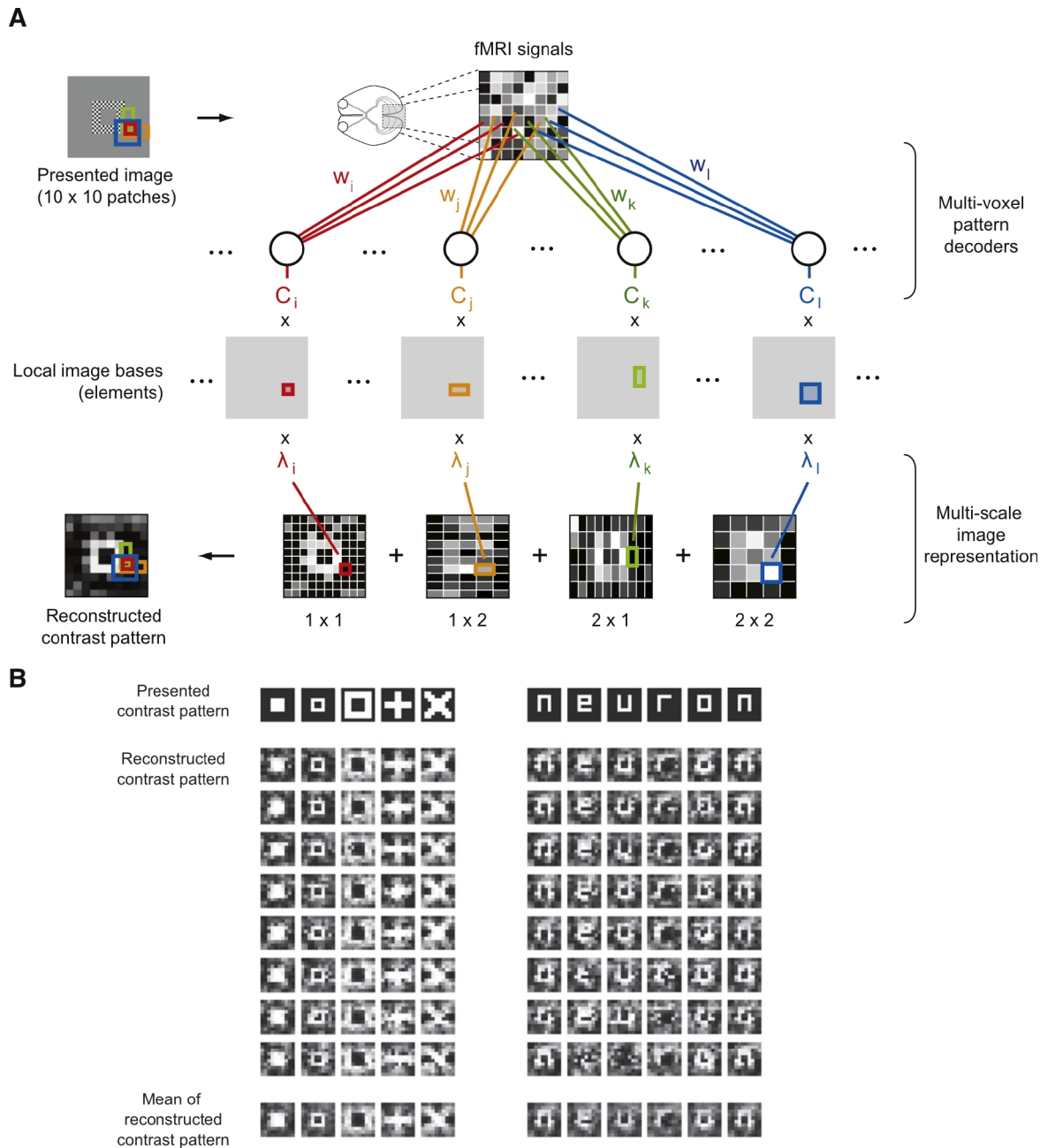


Fig. 4.3 Visual image reconstruction by multiscale local image decoders. (A) The reconstruction procedure. The fMRI activity is used with local decoders to predict contrast values for multiscale local image bases. The predicted contrasts are multiplied by their corresponding bases and linearly combined using combination coefficients to reconstruct the image. (B) The reconstructed artificial images. Adapted from Miyawaki et al. (2008), copyright 2008 with permission from Elsevier.

rate encoding models: a structural model, which predicted voxel responses based on low-level Gabor-like features extracted from the image, and a semantic model, which predicted voxel responses based on the presence of object categories and high-level semantic attributes. In the reconstruction phase, they applied Bayes' rule to invert these encoding models, combining the likelihood of the observed fMRI signals with a strong natural image prior that constrained the solution space to visually plausible images. Using this hybrid method, they successfully reconstructed natural scenes that preserved both low-level spatial structure and high-level semantic content, demonstrating that accurate image reconstruction requires the integration of hierarchical brain representations and external priors derived from natural image statistics. Further, Nishimoto et al. (2011) extended static visual image to dynamic visual experiences. They recorded fMRI responses while subjects watched natural movie clips and developed a motion-energy encoding model, which separately captured fast visual motion features and slow hemodynamic dynamics. Using a Bayesian decoding framework and a large prior of sampled movies, they successfully reconstructed coarse but recognizable dynamic visual scenes from brain activity, preserving motion trajectories and basic spatial layouts. This study demonstrated the feasibility of decoding continuous visual experiences from brain activity under naturalistic conditions.

Recent advances in deep learning have significantly accelerated the progress of visual image reconstruction from brain activity. Deep neural networks provide powerful feature representation capabilities, capturing hierarchical and abstract patterns from visual data, while deep generative models, such as generative adversarial networks (GANs) and variational autoencoders (VAEs), offer strong tools for synthesizing realistic images. These advances have enabled the high-fidelity reconstruction of images from brain activity using deep generative models.

Seeliger et al. (2018) reconstructed natural images from fMRI activity by training a linear mapping from brain signals into the latent space of a pre-trained deep convolutional GAN (DCGAN). This approach enabled the generation of plausible visual reconstructions corresponding to the perceived stimuli. Güçlütürk et al. (2017) developed the deep adversarial neural decoding model, combining variational inference with adversarial training to map brain activity into a GAN latent space optimized for faces, achieving high-fidelity face reconstructions. VanRullen and Reddy (2019) reconstructed face images from fMRI data by first training a VAE-GAN on a large face dataset to define a 1024-dimensional latent space, and then learning a linear mapping from brain activity to this space. The decoded latent vectors were fed into the generator to synthesize photorealistic faces, enabling accurate reconstruction of gender and identity. Shen et al. (2019a) advanced this direction by proposing an end-to-end image reconstruction framework, directly mapping fMRI activity to

reconstructed images. Furthermore, Ren et al. (2021) proposed a Dual-VAE/GAN framework that introduced a visually guided cognitive latent space, improving the semantic consistency and perceptual quality of reconstructed images, especially for unseen categories. Gaziv et al. (2022) developed a self-supervised approach for reconstructing natural images and decoding semantic categories without relying on extensive paired datasets, by training dual neural networks to enforce cycle consistency between images and brain activity and applying high-level perceptual similarity constraints.

Beyond reconstructing externally presented stimuli, recent studies have extended brain decoding to reconstruct subjective visual experiences, such as mental imagery, attention-modulated perception, and visual illusions. Based on hierarchical features extracted from DNNs, Shen et al. (2019b) proposed a powerful reconstruction method that integrates fMRI decoding with feature inversion to reconstruct both perceived and imagined images. Their approach involves decoding brain activity into hierarchical DNN features and then optimizing pixel values so that a generated image matches these decoded features. They incorporated a deep generator network as a natural image prior to enhance the realism of the reconstructions. This method produced recognizable images not only for viewed natural and artificial stimuli but also for imagery task (Figure 4.4A and 4.4B), marking a significant advancement in reconstructing subjective visual content from brain activity. Using the same reconstruction framework, Horikawa and Kamitani (2022) showed that top-down attention modulates neural representations. When participants focused on one of two superimposed images, only the attended stimulus appeared in the reconstruction (Figure 4.4C), revealing the influence of selective attention on decoded perception. Cheng et al. (2023) further reconstructed illusory experiences from brain activity. They mapped brain activity to the DNN features and then used a GAN-based image generator to reconstruct visual illusions such as illusory lines and neon color spreading (Figure 4.4D). Their results demonstrated that brain activity associated with illusions could be translated into perceptual feature representations and visualized as reconstructed images, revealing that early and higher visual areas differentially contribute to the representation of illusory percepts.

Recently, diffusion models have emerged as a powerful alternative for image synthesis, offering superior generation quality and robustness. These models gradually denoise a random distribution to produce coherent visual outputs and have been increasingly applied to brain decoding tasks, enabling more accurate and high-resolution image reconstruction. Takagi and Nishimoto (2023) pioneered this direction by employing a latent diffusion model conditioned on hierarchical visual features decoded from fMRI. Their model produced high-resolution reconstructions of perceived natural images, leveraging Stable Diffusion to generate realistic outputs while enabling interpretability of internal processes through layer-

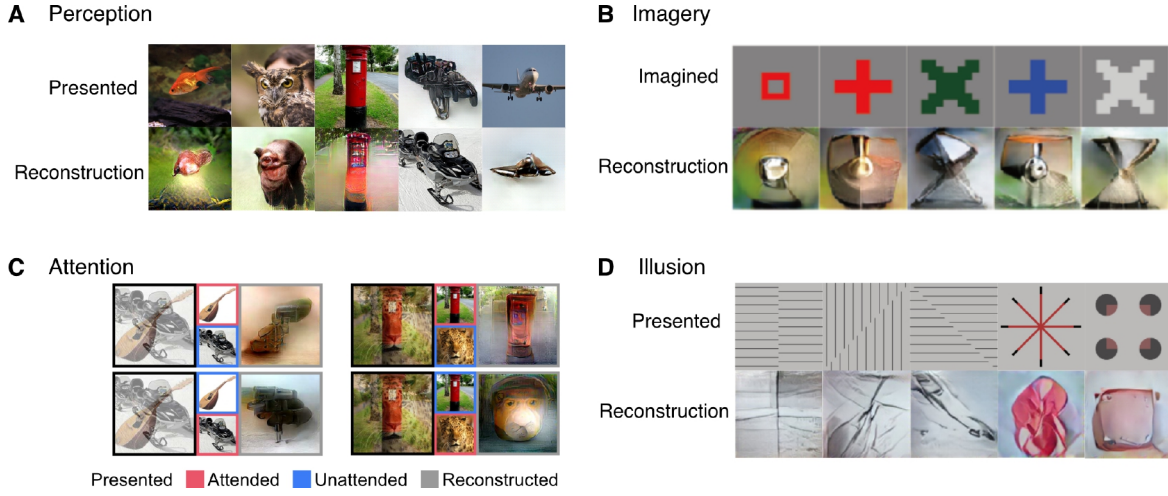


Fig. 4.4 Image reconstruction of perceptual and subjective experiences. (A) Reconstructed natural images. Adapted from Shen et al. (2019b), used under CC BY 4.0 license. (B) Reconstructed imagery images. Adapted from Shen et al. (2019b), used under CC BY 4.0 license. (C) Reconstructed attended images. Adapted from Horikawa and Kamitani (2022), used under CC BY 4.0 license. (D) Reconstructed illusory images. Adapted from Cheng et al. (2023), used under CC BY 4.0 license.

wise analyses. Scotti et al. (2023) proposed MindEye, a dual-path framework combining contrastive learning and a diffusion prior to bridge brain and image embeddings in CLIP space. By aligning fMRI-derived embeddings with CLIP image space, and feeding them into Versatile Diffusion, MindEye achieved good reconstruction and retrieval performance. Chen et al. (2023a) introduced MinD-Vis, which integrates sparse masked modeling of fMRI with a double-conditioned latent diffusion model. This architecture allowed for high-quality image generation from limited paired data, surpassing previous methods in semantic fidelity and perceptual quality. Similar diffusion-based methods (Bao et al., 2024; Gong et al., 2024; Quan et al., 2024; Scotti et al., 2024) have contributed to the semantic fidelity of visual image reconstructions. However, when using diffusion models for brain decoding, it is important to consider the limited diversity of stimuli in the dataset, as overlap of semantic and visual contents between training and test sets may pose a risk of spurious reconstructions (Shirakawa et al., 2024).

4.4 Setup of within-individual brain decoding

Building on the theoretical and empirical foundations laid out in the previous subsections, we now define the methodological setup for within-individual brain decoding used in the subsequent chapters of this thesis. We adopt a within-individual decoding framework in which

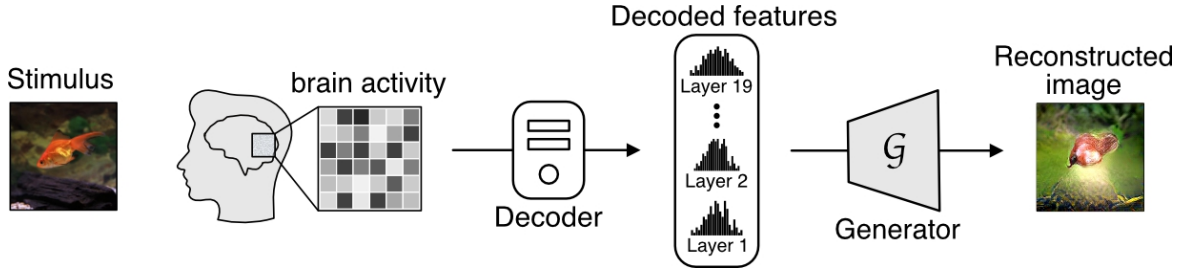


Fig. 4.5 The schematic of within-individual decoding and image reconstruction. When a stimulus is presented to a subject, the resulting brain activity patterns are decoded into DNN features using a decoder trained exclusively on data from the same subject. These decoded DNN features are then fed into a generator to reconstruct the image as perceived by the subject.

DNN features are predicted from brain activity patterns using decoders trained exclusively on data from the same subject, and these features are subsequently used for image reconstruction (Figure 4.5). This within-individual decoding and visual image reconstruction approach will serve as the comparison method in our cross-individual analysis in the subsequent chapters.

Selection of DNN architecture

We selected three convolutional neural network models: VGG19 (Simonyan & Zisserman, 2014), AlexNet (Krizhevsky et al., 2012), and VGGish-ish (Iashin & Rahtu, 2021) as sources of hierarchical feature representations, given their strong correspondence with human visual cortical hierarchies. Both VGG19 and AlexNet have demonstrated strong representational similarity with human visual cortical hierarchies, particularly in early and intermediate visual areas. Their well-characterized architecture and widespread use in previous decoding studies (Horikawa & Kamitani, 2022, 2017; Shen et al., 2019b) make them reliable feature extractors for perceptual content. VGGish-ish (Iashin & Rahtu, 2021), a modified VGG-style model designed for large-scale audio classification, was included to extract features from audio clips and to explore potential extensions to auditory decoding tasks. In addition, we incorporated features from CLIP, which provides a fundamentally different architecture and learning objective, offering a complementary perspective on perceptual representations and supporting flexible, high-level semantic decoding. These architectures allow us to extract a range of features spanning from low-level visual features to high-level semantic information, providing a rich representational basis for decoding.

Selection of brain decoding method

For brain decoding, we adopt the generic decoding framework developed by Horikawa and Kamitani (2017), which maps fMRI activity to DNN feature representations. Unlike categorical classifiers, this approach preserves rich representational information across multiple levels of visual processing, enabling the decoding of both seen and imagined objects within individuals. This approach serves as a robust foundation for linking stimulus representation with neural representation. Following Horikawa and Kamitani (2017), we use a linear decoder combined with voxel selection to predict DNN features from brain activity. Linear regression provides a computationally efficient and interpretable method for modeling the relationship between fMRI signals and high-dimensional feature spaces. Despite its simplicity, previous studies have shown that linear models can capture much of the relevant information contained in distributed voxel activity (Cheng et al., 2023; Horikawa & Kamitani, 2017). To further improve prediction accuracy and avoid overfitting, we apply voxel selection using sparse linear regression, which identifies a subset of informative voxels with non-zero weights for each feature unit. This strategy not only enhances decoding performance but also reveals meaningful spatial correspondences between cortical areas and the hierarchical structure of DNN representations (Horikawa & Kamitani, 2017).

Selection of reconstruction method

To reconstruct images from the decoded DNN features, we primarily adopted the method proposed by Shen et al. (2019b). This approach performs image reconstruction through feature inversion by iteratively optimizing the input image so that its DNN features, extracted by a fixed pretrained network, match the decoded features obtained from brain activity. We applied the same natural image prior and extended the loss function by adding a Deep Image Structure and Texture Similarity (DISTS) loss component (Muraki, 2024), which leverages the spatial characteristics of feature maps to enhance reconstruction detail. Unlike methods that rely solely on low-level feature inversion or pixel-wise similarity, this approach aligns brain-decoded features with DNN representations across multiple hierarchical layers, enabling the capture of both fine-grained visual details and abstract semantic structure. It generalizes well across diverse image types, even when trained only on natural images. The method has also proven effective for reconstructing not only externally viewed stimuli but also subjective visual content, such as mental imagery and attentionally modulated perception. In our within-individual reconstruction, for reconstructing illusion experiences, we additionally adopted the GAN-based generator introduced by Cheng et al. (2023).

4.5 Summary

In this chapter, we introduced within-individual decoding and reconstruction approaches. We began by reviewing the hierarchical feature representations extracted by deep neural networks. Then, we reviewed the evolution of decoding and reconstruction methods, which serve as tools for investigating and externalizing the neural representations of stimuli. Finally, we described the setup of within-individual decoding and image reconstruction used in this thesis, including the selection of DNN representations (VGG19, AlexNet, VGGish-ish, and CLIP), the DNN feature decoding approach, and visual image reconstruction methods. These elements lay the foundation for the decoding and reconstruction analysis presented in the subsequent chapters.

While the within-individual decoding framework offers reliable decoding and reconstruction performance, it typically requires large amounts of fMRI data from each individual to train the models. Given the practical difficulty of collecting such brain data, especially across a large variety of stimuli, this limitation highlights the need for more scalable and generalizable across-individual decoding strategies, which will be further explored in the subsequent chapters.

Chapter 5

Neural code conversion using content representations

5.1 Introduction

Individual differences in brain organization are observed at different scales, from macroscopic anatomy to fine-grained functional topography (Cox & Savoy, 2003; Fischl et al., 2008; Guntupalli et al., 2016; Haxby et al., 2011; Van Essen, 2005, 2004). These individual differences may complicate and confound group-level analyses, and make it challenging to generalize individual-specific models to other individuals. While anatomical alignment can help mitigate anatomical differences (Fischl et al., 2008; Van Essen, 2005, 2004), it does not fully account for the variability in functional topography across individuals, likely due to idiosyncratic neural representations at finer-grained scales (Cox & Savoy, 2003; Guntupalli et al., 2016; Haxby et al., 2011). Functional alignment can mitigate individual differences in functional topography. However, it generally requires presenting identical (shared) stimuli to different individuals to obtain paired brain activity data for model training, which is often unavailable and thus limits its broader application (Bazeille et al., 2021; Chen et al., 2015; Guntupalli et al., 2016; Haxby et al., 2011, 2020; Ho et al., 2023; Yamada et al., 2015).

In this chapter, we introduce a flexible pairwise functional alignment model using content loss-based optimization, which relies on the latent feature representation of stimulus content and does not require subjects to be exposed to shared stimuli. For each pair of subjects, we designate one as the target subject and the other as the source subject. First, we pre-train decoders to predict the latent features of stimuli perceived by the target subject based on their measured brain activity (Horikawa & Kamitani, 2017). Then, a neural code converter model is optimized to minimize the content loss between the latent features of the stimulus and

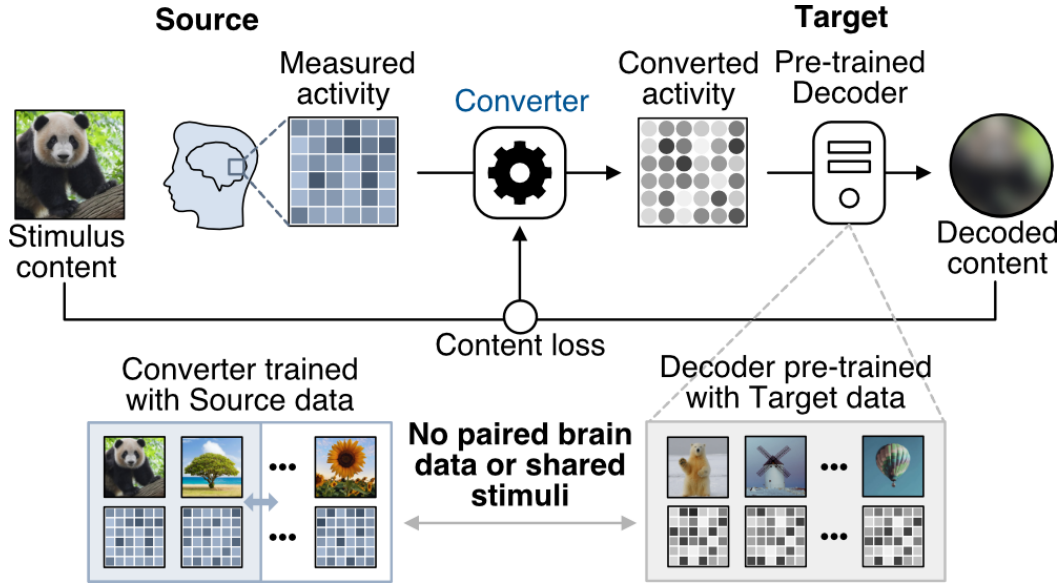


Fig. 5.1 The schematic of the content-loss-based neural code converter. The target subject's training data is used to pre-train the target decoder, whereas the source subject's training data is used for the converter training. The converter is optimized so that the converted brain activity is decoded into content representations that closely resemble that of the stimulus given to the source subject. During the training stage, there was no need for paired brain data or shared stimuli between the source and the target.

those decoded from the converted brain activity patterns of the target subject (Figure 5.1). This training approach for the converter does not rely on paired brain activity (shared stimuli). To evaluate the method's ability to convert brain activity patterns between subjects, we compared the proposed converters optimized with visual content with converters optimized with paired brain activity in terms of pattern correlation and profile correlation. We also conducted a sub-wise conversion analysis by training separate converters for each pair of ROIs between source and target subjects, in order to assess whether our method can capture hierarchical neural representations. We then examine whether the method remains effective without shared stimuli, by evaluating conversion performance with both overlapping and non-overlapping stimuli between converter and decoder trainings.

5.2 Methods

5.2.1 Content loss-based neural code converter

The content loss-based neural code converter for each pair of subjects uses a nonlinear Multi-Layer Perceptron (MLP) to predict the brain activity patterns of one subject (target)

from the brain activity patterns of another subject (source). It has two hidden layers with instance normalization and ReLU functions, and the number of units in each hidden layer is half that of the input layer. The converter Φ takes a source subject's brain activity pattern $\mathbf{x}_i \in \mathbb{R}^m$ consisting of m voxels' values, and predicts the target subject's brain activity pattern as $\Phi(\mathbf{x}_i) \in \mathbb{R}^n$, consisting of n voxels' values. Then, a pre-trained target decoder takes the converted brain activity pattern $\Phi(\mathbf{x}_i)$ and predicts the decoded feature pattern of the image stimulus as $\mathbf{d}_{il} = \mathbf{W}_l \Phi(\mathbf{x}_i) + \mathbf{b}_l$, where $\mathbf{d}_{il} \in \mathbb{R}^{d_l}$ is the decoded feature pattern consisting of d_l units' values in the l -th DNN layer for the i -th image stimulus. $\mathbf{W}_l \in \mathbb{R}^{d_l \times n}$ and $\mathbf{b}_l \in \mathbb{R}^{d_l}$ are the decoding matrix and bias vector, respectively, of the pre-trained decoder. The converter is optimized to make decoded features from the converted brain activity similar to the true features of the source subject's stimulus. It is trained to minimize the objective function

$$\mathcal{L}(\Phi) = \sum_i^N \sum_l^L \eta_l \|\mathbf{v}_{il} - (\mathbf{W}_l \Phi(\mathbf{x}_i) + \mathbf{b}_l)\|^2, \quad (5.1)$$

where $\mathbf{v}_{il} \in \mathbb{R}^{d_l}$ represents true units' values in the l -th DNN layer for the i -th image stimulus, N is the number of training samples, and L is the number of DNN layers. η_l is the parameter that weighs the contribution of the l -th layer, which is set to be $1/\|\mathbf{v}_{il}\|^2$.

The converter was trained through an iterative process, where each iteration applied a stochastic decoding strategy across all layers of the VGG19 model (Simonyan & Zisserman, 2014), which comprises 16 convolutional layers and 3 fully connected layers. 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. During each iteration, within each convolutional layer, a feature map was chosen at random, and all its units were subjected to the decoding process. Due to the much fewer units in the fully connected layers, all units from these layers were decoded in each iteration. 1,024 iterations were used to ensure the comprehensive involvement of all DNN units from every layer of the image stimulus during the training stage.

5.2.2 Brain loss-based neural code converter

The brain loss-based neural code converter consists of a set of regularized linear regression models (Ho et al., 2023). It 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 as $\hat{\mathbf{y}}_i = \mathbf{M}\mathbf{x}_i + \mathbf{c}$, consisting of n voxels' values. $\mathbf{M} \in \mathbb{R}^{n \times m}$ is the conversion matrix, and $\mathbf{c} \in \mathbb{R}^n$ is the bias vector. The

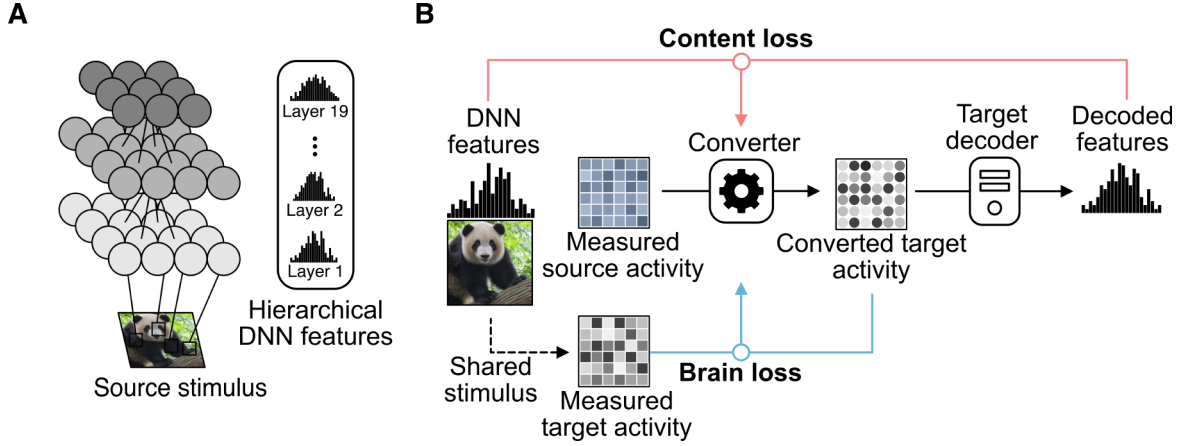


Fig. 5.2 The comparison of converter training using content loss-based optimization and brain loss-based optimization. (A) Hierarchical DNN features as visual content. (B) The content loss-based optimization minimizes the loss between the DNN features decoded from the converted target activity and those extracted from the corresponding stimulus. The brain loss-based optimization minimizes the loss between the converted target activity and the measured target activity.

converter is trained to minimize the objective function

$$\mathcal{L}(\mathbf{M}, \mathbf{c}) = \sum_i^N \|\mathbf{y}_i - (\mathbf{M}\mathbf{x}_i + \mathbf{c})\|^2 + \lambda \|\mathbf{M}\|_F^2, \quad (5.2)$$

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 that is set as described in Ho et al. (2023), and $\|\cdot\|_F$ represents the Frobenius norm.

Different from the content loss-based optimization, which minimizes the loss between the DNN features decoded from the converted target activity and those extracted from the corresponding stimulus, the brain loss-based optimization minimizes the loss between the converted target activity and the measured target activity (Figure 5.2).

5.2.3 Other brain loss-based functional alignment methods

While our analysis primarily focuses on content loss-based neural code converters and their comparison with brain loss-based converters, we also included other functional alignment methods based on brain loss for additional comparison, such as pairwise Procrustes transformation and template-mediated Procrustes transformation (hyperlalignment; Haxby et al., 2011).

Procrustes transformation

The Procrustes transformation is originally introduced by Haxby et al. (2011) for hyperalignment analysis. This transformation estimates an orthogonal matrix $\mathbf{M} \in \mathbb{R}^{n \times m}$ that minimizes the discrepancy between the source subject's brain activity pattern $\mathbf{x}_i \in \mathbb{R}^m$ and target subject's brain activity pattern $\mathbf{y}_i \in \mathbb{R}^n$, under the constraint $\mathbf{M}^T \mathbf{M} = \mathbf{I}$, where $\mathbf{I}$ denotes the identity matrix. The orthogonal matrix $\mathbf{M}$ is optimized by minimizing the objective function

$$\mathcal{L}(\mathbf{M}) = \sum_i^N \|\mathbf{y}_i - (\mathbf{M}\mathbf{x}_i)\|^2, \quad (5.3)$$

where N is the number of training samples.

Template-mediated Procrustes transformation

The hyperalignment (Haxby et al., 2011) is a template-based alignment method for estimating a shared representational space across subjects. We also performed pairwise alignment via hyperalignment, conducting analyses similar to those in the neural code converter. This template-mediated pairwise alignment first estimates a shared template across subjects using hyperalignment. Then, it constructs a pairwise Procrustes transformation by first mapping a source subject's brain activity onto the template, followed by an inverse mapping from the template to a target subject's brain space.

Hyperalignment algorithm begins by selecting one subject as the initial reference template and aligning another subject's brain activity to this template using Procrustes transformation. After this alignment, the reference template is updated by averaging the aligned brain activity with the current template. This process continues iteratively: each subsequent subject's brain activity is aligned to the most recently updated template, and the template is further refined by incorporating the aligned brain activity of the new subjects. In the second step, each subject's original brain activity is aligned to the refined template (obtained after the first step). The aligned activity in this template space is averaged to create the finalized shared template. Finally, in the third step, each subject's brain activity is aligned to the final template, and an orthogonal transformation matrix is derived for each subject.

5.2.4 Conversion accuracy

We evaluated the conversion accuracy using two metrics: pattern correlation and profile correlation. Pattern correlation is the Pearson correlation coefficient between the converted and measured voxel patterns for a test stimulus (Figure 5.3A). Profile correlation, which calculates the Pearson correlation coefficient between the sequences of converted and measured

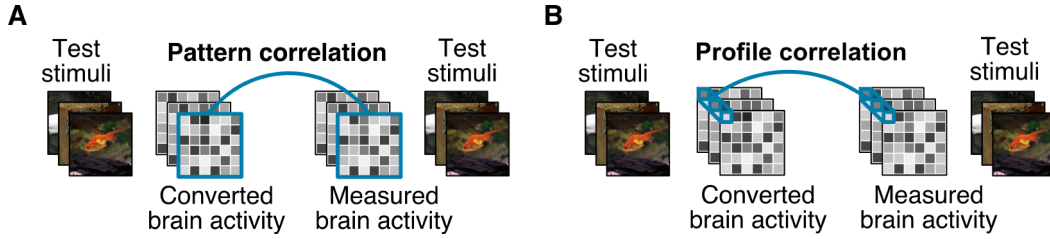


Fig. 5.3 Evaluation of conversion accuracy. (A) Pattern correlation. Pearson correlation coefficient between the converted and measured voxel patterns for a test stimulus was calculated for pattern correlation. (B) Profile correlation. Pearson correlation coefficient between the sequences of converted and measured responses of a single voxel to the 50 natural test stimuli was calculated for profile correlation.

responses of a single voxel to the 50 natural test stimuli (Figure 5.3B). Repeated measures of the brain responses to an identical stimulus in fMRI data are subject to measurement noise, which impacts the evaluation of conversion accuracy. To address this issue, we performed the noise ceiling estimation (Hsu et al., 2004; Lescroart & Gallant, 2019). The noise ceiling is calculated by averaging the correlation coefficients between repeated responses to identical stimuli, which reflects the maximum performance that the converter model can achieve given measurement noise in fMRI data. We excluded samples or voxels if their noise ceilings fell below a specified threshold (the 99th percentile of the distribution from random pairs). The obtained correlation coefficients were then normalized by dividing the raw values by their respective noise ceilings. For each individual pair and each ROI, the normalized pattern correlation coefficients for all test stimuli were averaged to calculate the mean conversion accuracy (pattern), while the normalized profile correlation coefficients for all voxels were averaged to calculate the mean conversion accuracy (profile).

5.2.5 Visual subarea-wise conversion

Our main conversion analysis used visual cortex (VC) activity patterns from a source subject to estimate voxel responses in a target subject. This method, referred to as whole VC conversion, does not distinguish between functional subareas within the cortex. In contrast, we also introduce another conversion strategy that aligns the brain activity patterns between subareas of the source and target subjects, which is subarea-wise conversion (Figure 5.4).

In the whole VC approach, the converter was trained to predict activity patterns in the target VC using the whole VC of the source subject, with conversion accuracy evaluated separately for each ROI. In the subarea-wise approach, separate converters were trained for each ROI to predict activity patterns in target subareas using only the corresponding source subareas, with accuracy calculated for each ROI individually. We compared these

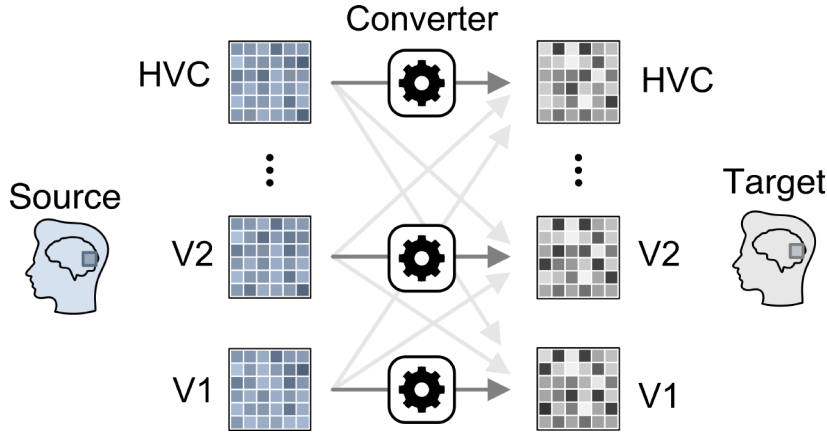


Fig. 5.4 The schematic of subarea-wise conversion. The converters are trained for each combination of subareas between source and target subjects.

two approaches to evaluate the impact of integrating prior anatomical information into the conversion framework.

To further investigate whether the conversion reflects the hierarchy of visual subareas, we extended our analysis to all combinations of subareas between source and target subjects. Unlike the subarea-wise conversion within corresponding ROIs, this analysis involved predicting activity patterns in a target subarea using different source subareas (e.g., using source V1, V2, V3, V4, or HVC individually to predict target V1). Similarly, we calculated conversion accuracy for each combination of subareas individually.

5.2.6 Conversion without shared stimuli

We investigated whether we could perform neural code conversion without shared stimuli. To examine this, we randomly divided the training samples from the Deeprecon dataset into two distinct halves based on the categories of stimuli. The source subject was provided with 3000 training samples (600 images from 75 randomly selected categories out of 150 categories, with five repetitions of each image), and the target subject were given a different set of 3000 training samples (the remaining 600 images with five repetitions each), a condition we refer to as “Non-overlapping” (Figure 5.5). This strategy was designed to prevent any pairing in their brain activity patterns. We then trained neural code converters for each individual pair under this condition and evaluated their conversion accuracy. For comparison, we also performed the same analysis under an “Overlapping” condition, where the target subject used training samples the same as the source subject, resulting in overlapping stimuli between the converter and decoder trainings.

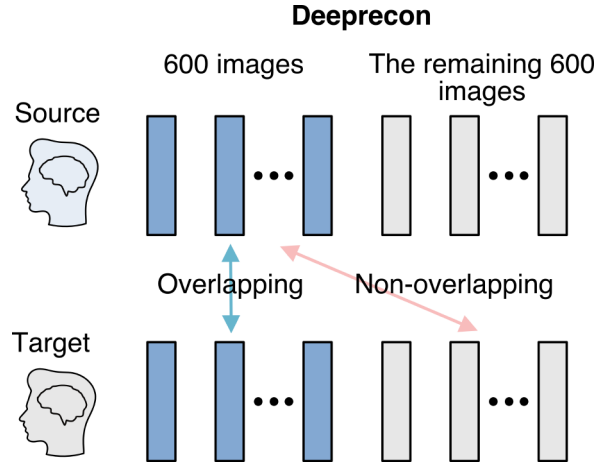


Fig. 5.5 Non-overlapping stimuli for source and target subjects. The Deeprecon training samples were randomly split into two distinct halves. Source and target subjects were allocated different halves of training samples, ensuring no overlap in image stimuli between the source and target subjects.

5.2.7 Group statistics

We performed group-level statistical inference by showing the group means and confidence intervals for conversion accuracy. Each data point represents the mean conversion accuracy corresponding to a unique pair of subjects. Due to potential correlations between pairs involving the same subject (dyadic dependency), we applied bootstrapping to these dyadic data points to calculate the group mean and 95% confidence interval. We performed bootstrap sampling separately on the source and target IDs, obtaining data points of the original sample size while ensuring that pairs where the source and target IDs were the same were excluded. The obtained data points were then used to calculate the mean. This process was repeated 1000 times to generate bootstrap replicates, from which we calculated the 95% confidence interval.

In the comparison of the content loss-based converter with the brain loss-based converter, we performed post hoc statistical analyses to evaluate the differences. For conversion accuracy within each ROI, we used Levene's test to assess the homogeneity of variance of the two methods, where Levene's statistic and p-values were computed using mean conversion accuracies from 20 individual pairs as data points.

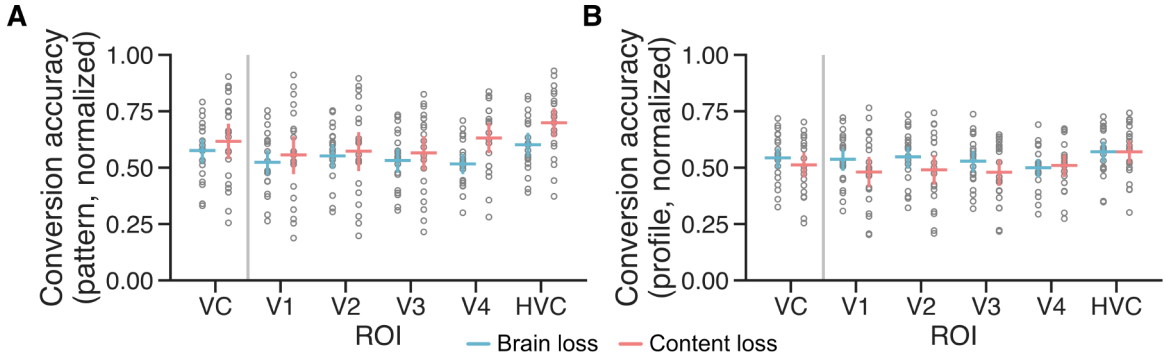


Fig. 5.6 Conversion accuracy. (A) Conversion accuracies measured by pattern correlation. Distributions of the pattern correlation coefficients of 20 pairs are shown for the VC and visual subareas. The horizontal bars represent the mean accuracies across the 20 pairs, with vertical bars representing the 95% C.I., and each dot indicates an individual pair's mean correlation coefficient over stimuli. (B) Conversion accuracies measured by profile correlation. The two converters are compared as in (B) but with profile correlation coefficients.

5.3 Results

5.3.1 Visual neural code conversion

We first investigated whether brain activity patterns could be reliably aligned across subjects using a content loss-based converter within the Deeprecon dataset. The whole VC conversion was performed, and conversion accuracy was evaluated by two metrics: (a) pattern correlation, which is the Pearson correlation coefficient between the converted and measured voxel patterns for a test stimulus (Figure 5.2C), and (b) profile correlation, which is the Pearson correlation coefficient between the sequences of converted and measured responses of a single voxel to the 50 natural test stimuli (Figure 5.2D). The correlation coefficients obtained were normalized against their noise ceilings to account for the noise present in fMRI brain responses across repeated measurements using the same stimulus (Hsu et al., 2004; Lescroart and Gallant, 2019; see Subsection 5.2.4). To summarize the results, we averaged the pattern correlation coefficients across all stimuli and profile correlation coefficients across all voxels within each individual pair and each ROI. For comparison, we also evaluated the brain loss-based neural code converter (Ho et al., 2023; Yamada et al., 2015; Figure 5.2B; see Subsection 5.2.2).

Figure 5.6A and 5.6B present the distributions of pattern correlation coefficients and profile correlation coefficients across all conversion pairs for different ROIs in the target brain space. Each data point in the Figures corresponds to an individual pair. The mean pattern correlation for the whole VC using content loss-based neural code converter was 0.62 ± 0.08 (mean with 95% confidence interval) for 20 individual pairs (Figure 5.6A for

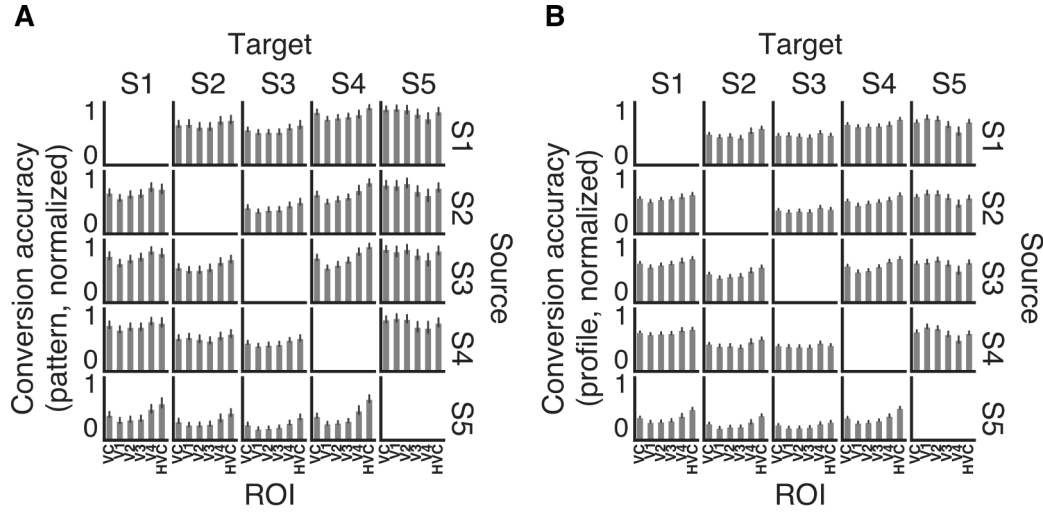


Fig. 5.7 Conversion accuracies of individual pairs. (A) Pattern correlation. The pattern correlation coefficients for 50 visual stimuli were used to calculate the mean conversion accuracy (pattern) and the 95% confidence (error bar) in each pair of subjects. (B) Profile correlation. The profile correlation coefficients for individual voxels were used to calculate the mean conversion accuracy (profile) and the 95% confidence interval (error bar) in each pair of subjects.

group results; Figure 5.7A for individual pairs), whereas the mean profile correlation for the VC was 0.51 ± 0.05 for 20 individual pairs (Figure 5.6B for group results; Figure 5.7B right for individual pairs). The accuracy of the brain loss-based converter aligned with the results reported by Yamada et al. (2015) and Ho et al. (2023). While the content loss-based converter was not explicitly trained to align brain activity patterns, its performance measured by correlations between converted and measured brain activity patterns was comparable to, or only slightly lower than, the brain loss-based method across all examined visual subareas. Although some differences in variance are observed between the two methods, they were not statistically significant (post-hoc Levene's test, pattern correlation: $p > 0.05$ for all ROIs except V2, where $p = 0.039$; profile correlation: $p > 0.05$ for all ROIs). Both methods demonstrated relatively consistent alignment, with individual pairs showing correlated performance between the two methods (Figure 5.8). We also tested other standard methods for functional alignment (Haxby et al., 2011; see Subsection 5.2.3), showing similar conversion accuracies (Figure 5.9).

5.3.2 Hierarchical correspondence of neural representations

We extended our analysis to different selections of visual ROIs. We performed subarea-wise conversion analyses by training separate converters for paired ROIs between source and target

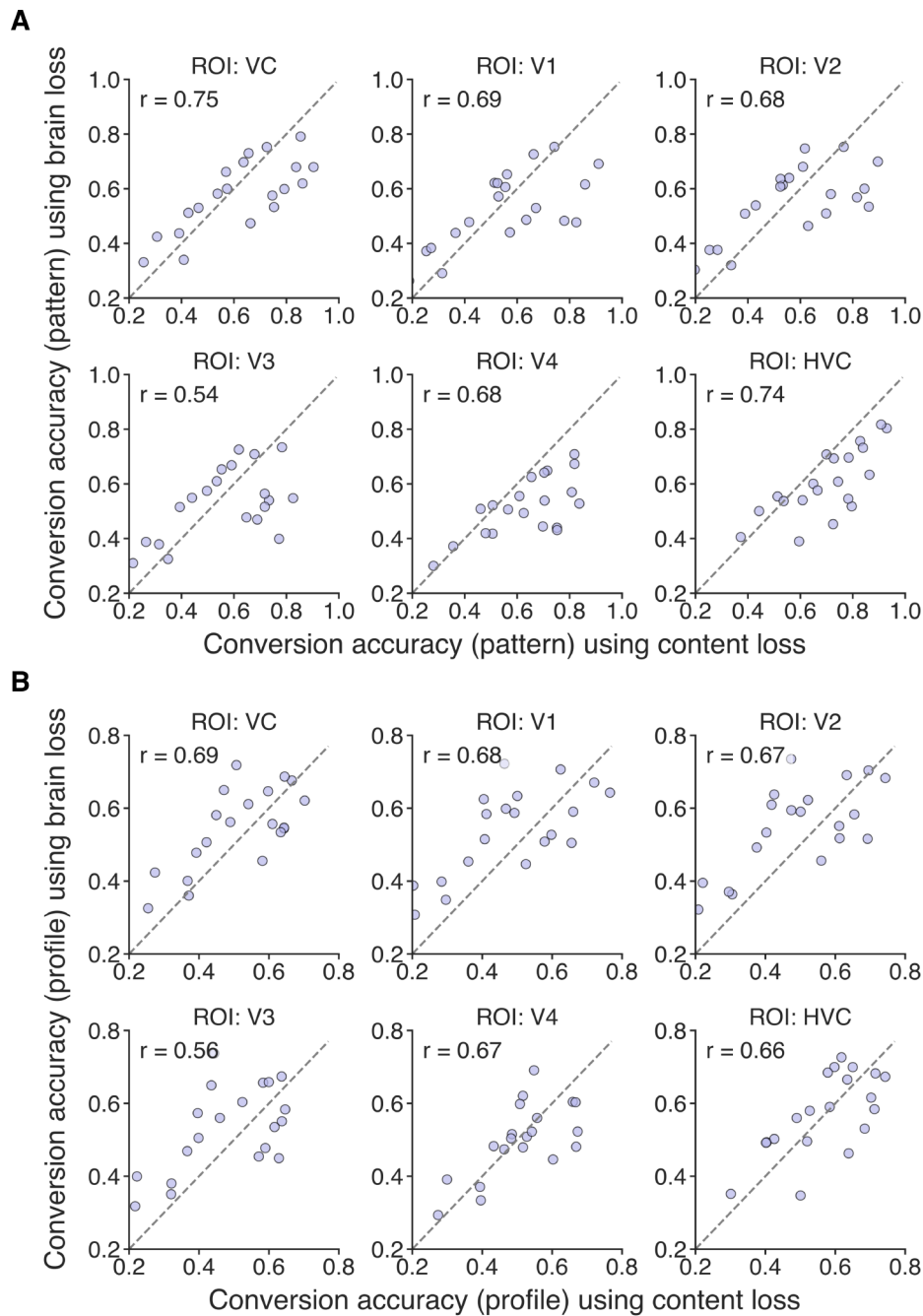


Fig. 5.8 Correlation in conversion accuracy between brain loss and content loss. (A) Scatter plots comparing conversion accuracies quantified using pattern correlations across different ROIs. The vertical axis shows accuracies obtained using brain loss, while the horizontal axis represents accuracies obtained using content loss. Each data point represents an individual subject pair, with a total of 20 pairs. A diagonal reference line indicates where accuracies from both methods would be equal, allowing direct comparison of their performance. Pearson correlation coefficients (r) are displayed for each ROI to quantify the relationship between the two methods. (B) Scatter plots comparing conversion accuracies quantified using profile correlations across different ROIs. The axes and reference lines follow the same conventions as in (A), with correlation coefficients (r) shown for each ROI.

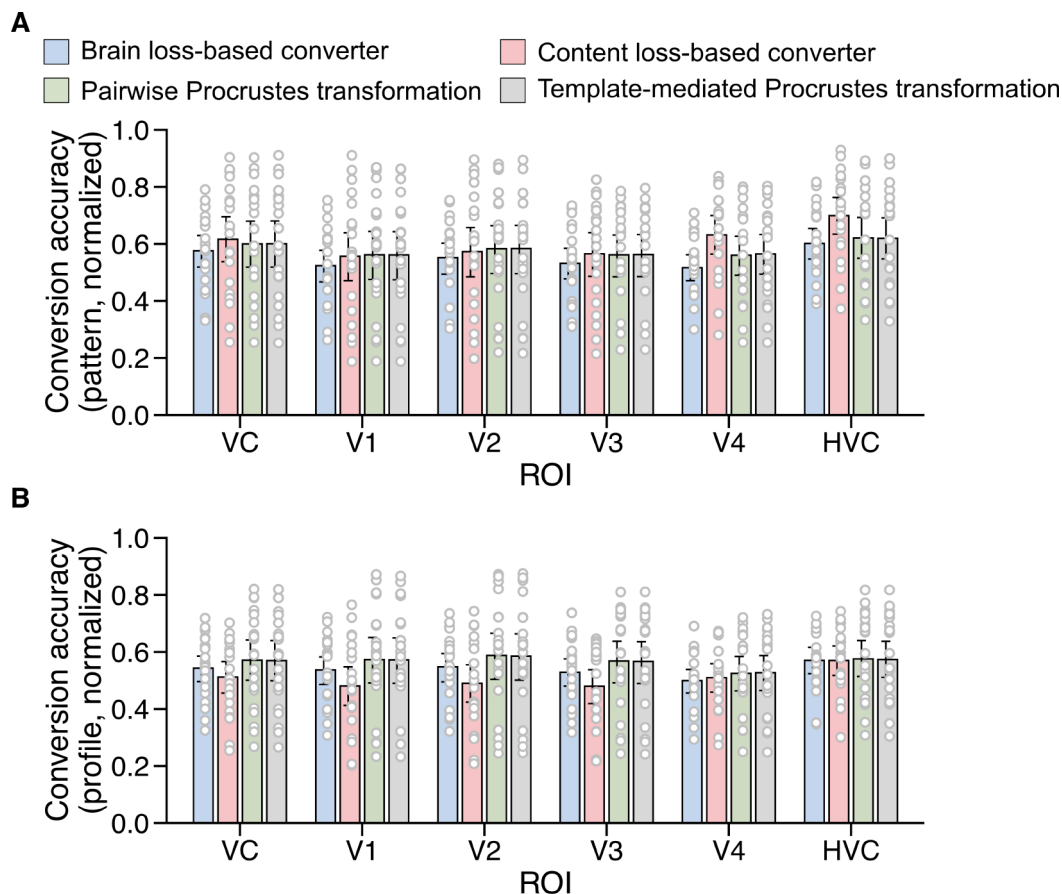


Fig. 5.9 Comparison of conversion accuracies across functional alignment methods (A) Conversion accuracies quantified using pattern correlations for four different alignment methods in the VC and visual subareas. Correlation coefficients are shown for 20 subject pairs, where each dot represents a pair's mean correlation coefficient across stimuli. Bars indicate mean accuracies across all pairs (error bars, 95% confidence intervals). The comparison includes the brain loss-based and content loss-based converters, along with two established methods: the pairwise Procrustes transformation and the template-mediated Procrustes transformation. (B) Conversion accuracies quantified using profile correlations. The analysis follows the same methodology as in (A), comparing all four alignment methods.

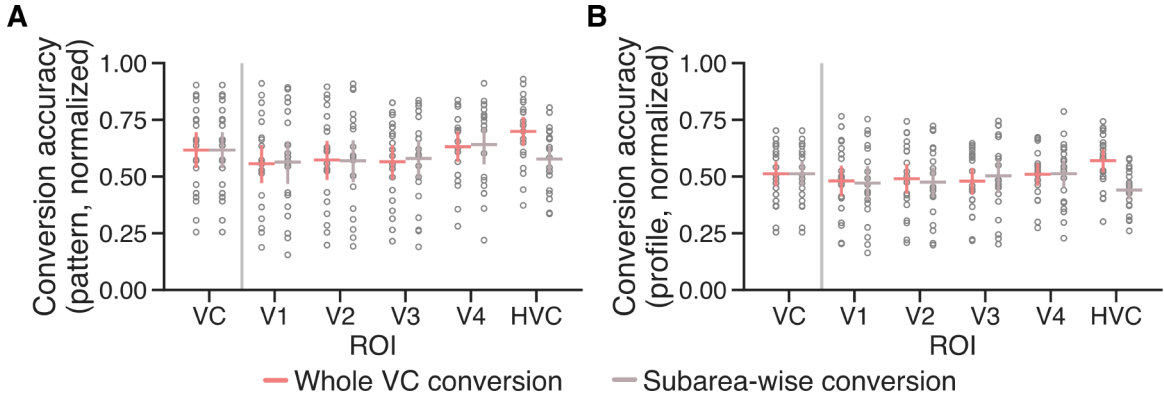


Fig. 5.10 Comparison of conversion accuracies between whole VC and subarea-wise approaches. (A) Conversion accuracies quantified using pattern correlations across different ROIs. The horizontal bars represent the mean accuracies across the 20 pairs, with vertical bars representing the 95% C.I., and each dot indicates an individual pair's mean correlation coefficient over stimuli. (B) Conversion accuracies quantified using profile correlations across different ROIs. The analysis follows the same methodology as in (A), comparing whole VC and subarea-wise approaches.

subjects (Ho et al., 2023). When using corresponding ROIs, both methods achieve similar levels of performance, while in HVC, the subarea-wise method shows poorer performance (Figure 5.10). There were no significant differences in variance between the two methods across ROIs (Levene's test, pattern, $p > 0.05$ for all ROIs; profile, $p > 0.05$ for all ROIs). The results indicate that incorporating anatomical constraints provides no substantial benefit and suggest that the flexible mapping afforded by the whole VC method may be advantageous. When predicting target ROI activity from different source ROIs, in most subareas, the highest conversion accuracy is achieved using the corresponding source subareas (Figure 5.11), suggesting that neural code conversion accuracy can potentially serve as a quantitative measure of homology between brain areas. These results indicate that the content loss-based approach can effectively convert brain activity patterns across subjects, preserving hierarchical neural representations.

5.3.3 Conversion without shared stimuli

We have shown the conversion results when the training of the converter and decoder involves overlapping (shared) stimuli. However, it remains unclear whether we can perform neural code conversion without shared stimuli. To examine this, we randomly divided the training samples from the Deeprecon dataset into two distinct halves based on the categories of stimuli.

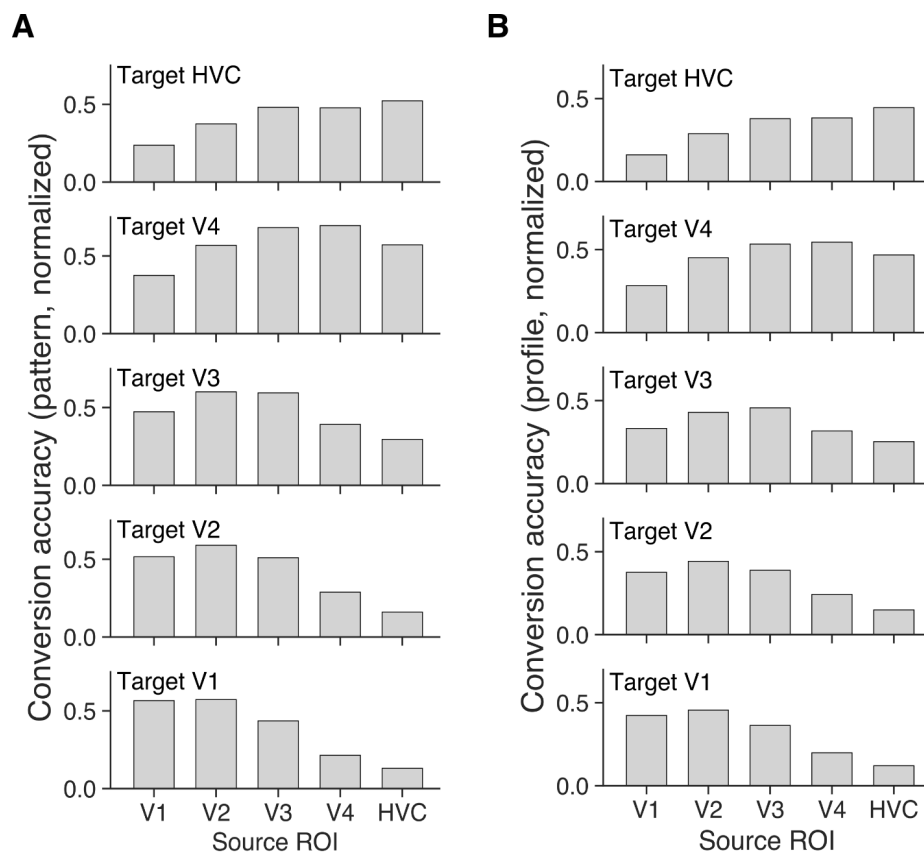


Fig. 5.11 Conversion performance between different visual subareas. (A) Conversion accuracies quantified using pattern correlations for all combinations of subareas between source and target subjects. Unlike the subarea-wise conversion within corresponding ROIs shown in Figure 5.10, this analysis involved predicting activity patterns in a target subarea using different source subareas (e.g., using source V1, V2, V3, V4, or HVC individually to predict target V1). The results from a representative subject are shown. (B) Conversion accuracies quantified using profile correlations. The analysis follows the same methodology as in (A).

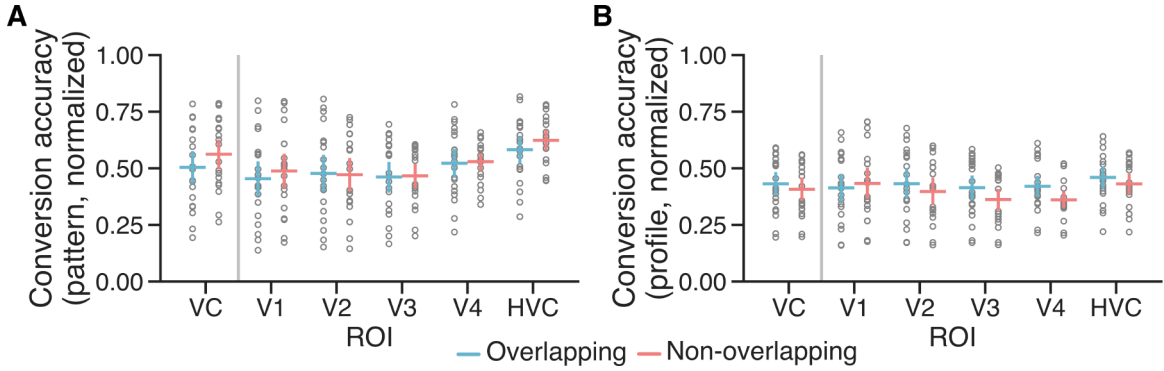


Fig. 5.12 The effect of overlapping (shared) stimuli between converter and decoder trainings. (A) Conversion accuracy measured by pattern correlation. Distributions of the pattern coefficients of 20 pairs are shown for the VC and visual subareas. The horizontal bars represent the mean accuracies across the 20 pairs, with vertical bars representing the 95% C.I., and each dot indicates an individual pair's mean correlation coefficient over stimuli. (B) Conversion accuracy measured by profile correlation. The two conditions are compared as in b but with profile correlation coefficients.

Figure 5.12 presents the conversion accuracy for the whole VC and visual subareas under two analytical conditions for all individual pairs. The mean pattern correlation for the whole VC under the Non-overlapping condition was 0.56 ± 0.06 (mean with 95% confidence interval) for 20 individual pairs, whereas the mean profile correlation was 0.41 ± 0.05 for 20 individual pairs. For pattern correlation, the content loss-based converter demonstrates comparable performance under both conditions across ROIs. There were no significant differences in variance between the two conditions in most ROIs (Levene's test, pattern correlation: $p > 0.05$ for all ROIs except V4, where $p = 0.035$). Similar results was observed in profile correlation. There were no significant differences in variance between the two conditions across ROIs (Levene's test, $p > 0.05$ for all ROIs). These results demonstrated that the content loss-based converters can convert brain activity patterns across subjects without the need for overlapping (shared) stimuli, even between decoder and converter trainings.

5.4 Discussion

This chapter aimed to develop a functional alignment method that eliminates the need for identical sets of stimuli (shared stimuli) to be presented to individuals during training. To achieve this, we proposed a flexible approach for training neural code converters using content loss-based optimization, which optimizes the converter parameters based on the stimulus content as represented by latent features of a DNN model, rather than relying on brain activity patterns paired by the same set of stimuli.

Our proposed converter accurately converted a source subject’s brain activity into a target subject’s brain space, as evaluated by pattern correlations and profile correlations. By converting brain activity patterns across all combinations of subareas between source and target subjects, the highest conversion accuracies were observed between subareas at similar hierarchical levels. These results indicate that the content loss-based approach can effectively convert brain activity patterns across subjects, preserving hierarchical neural representations. Training the converter with non-shared stimuli yielded comparable conversion accuracy, further confirming the effectiveness of our method.

A major difference between our functional alignment method and the previous ones is that we optimized the converter based on the stimulus content rather than on paired brain activity patterns. The optimization process is designed to ensure that the converted brain activity patterns can be decoded into latent feature representations that closely resemble the feature representations extracted from the stimulus. By performing the optimization within the stimulus content space, the content loss-based converter eliminated the need for paired brain activity patterns during the training stage. In contrast, conventional brain loss-based functional alignment methods (Bazeille et al., 2021; Ho et al., 2023; Yamada et al., 2015) optimize within a brain space, aiming to minimize the discrepancy between measured brain activity patterns of different individuals when presented with the same stimulus.

We used the brain loss-based converter, pairwise Procrustes transformation, and template-mediated Procrustes transformation for comparison in the context of shared stimuli. These methods broadly reflect functional alignment methods based on “brain loss” within innate brain space. It is important to note that “brain loss” can be defined in various ways. For instance, it can be measured using latent representations of brain data, enabling alignment within a latent brain space rather than the innate brain space (Chen et al., 2015; Chen et al., 2023a; Guntupalli et al., 2016; Haxby et al., 2011, Haxby et al., 2020; Qian et al., 2023; Jiahui et al., 2023; Huo et al., 2024). While shared stimuli are commonly used in brain loss-based alignment tasks, they may not be strictly necessary in other contexts. For example, connectivity hyperalignment (Guntupalli et al., 2016) aligns neural representations by leveraging functional connectivity of cortical fields, enabling alignment even when individuals experience different stimuli (Jiahui et al., 2023). Similarly, the Individualized Neural Tuning (INT) model (Feilong et al., 2023) separates stimulus-specific information from neural response tuning and captures stimulus-general neural tuning that can generalize across different stimuli. Furthermore, brain loss-based functional alignment is not limited to measured brain data. By using subject-specific pre-trained encoders to predict brain activity patterns for non-shared stimuli, brain loss-based optimization can be applied to predicted data in the same way as measured data with shared stimuli (Wasserman et al., 2024).

Chapter 6

Inter-individual decoding and image reconstruction

6.1 Introduction

Decoding perceptual experiences from brain activity across individuals is important for uncovering the shared neural basis of perception. Recent advances in deep neural networks (DNNs) and brain decoding have enabled the reconstruction of visual images from brain activity within individuals, providing a powerful tool for externalizing perceptual content and probing how sensory information is represented in the brain. However, extending such approaches across individuals presents challenges due to anatomical and functional variability between brains. Neural code conversion offers a promising approach to mitigating inter-individual variability by aligning brain activity patterns across subjects. This alignment enables models trained on one subject to be generalized to others, supporting inter-individual decoding and visual image reconstruction.

This chapter examines whether fine-grained visual information is preserved after neural code conversion, using inter-individual decoding and image reconstruction analyses. We begin by introducing the methodological framework, including DNN-based feature decoders, image reconstruction techniques, and evaluation metrics. Based on the neural code conversion method introduced in Chapter 5, we transformed brain activity patterns from a source subject into the brain space of a target subject. We then applied decoding and reconstruction models to the converted activity (Figure 6.1) and assessed whether it retained fine-grained visual features by comparing performance with within-individual decoding and a brain loss-based converter. To evaluate the flexibility of our method, we further tested its performance when the decoder and converter were trained on non-overlapping stimulus sets. Finally, we discuss

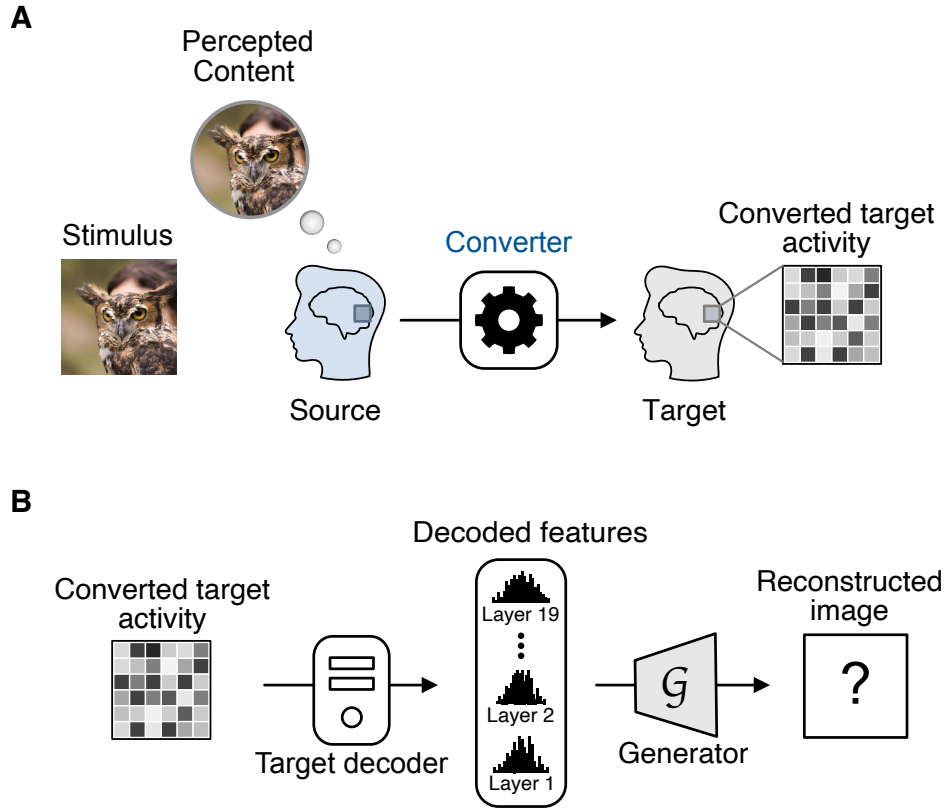


Fig. 6.1 Inter-individual decoding and image reconstruction. (A) Neural code conversion. Given a stimulus, the brain activity pattern of the source subject is converted into the target subject's brain space by a neural code converter. (B) Brain decoding and image reconstruction. The converted target activity is decoded into DNN features using a target subject decoder, and further used to reconstruct the image as the source subject perceived through a generator.

the results and highlight how content-optimized neural code conversion differs from previous approaches.

6.2 Methods

6.2.1 Deep neural networks

We used the VGG19 DNN model (Simonyan & Zisserman, 2014) implemented using the Caffe library (Jia et al., 2014) in the converter training and DNN feature decoding analysis. 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. Outputs from individual units before rectification were used. 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.

For the evaluation of reconstructed images, we used another DNN model, the AlexNet (Krizhevsky et al., 2012) implemented using the Caffe library to extract DNN features from the reconstructed images and the presented image. This model is also pre-trained with images in ImageNet to classify 1,000 object categories (available from https://github.com/BVLC/caffe/tree/master/models/bvlc_alexnet). The model consists of five convolutional layers and three fully connected layers. All the input images to the model were rescaled to 224×224 pixels. 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.

6.2.2 DNN feature decoding analysis

We used a ridge regression model as a DNN feature decoder. This model predicts the feature values of the stimulus, given an fMRI activity pattern evoked by the stimulus. We normalized both the feature values and voxel responses before model training and used a voxel selection procedure. This procedure involved calculating the Pearson correlation coefficients between sequences of voxel responses and feature values for all voxels. The 500 voxels exhibiting the highest correlations were selected for training. We set the ridge regularization parameter to 100 to enhance model robustness. The feature decoding analysis is detailed in the studies by Horikawa and Kamitani (2017) and Shen et al. (2019b). For testing the trained decoders, we used the average fMRI pattern over repetitions to improve the signal-to-noise ratio of the fMRI signal. We evaluated the decoding accuracy using pattern correlation and profile correlation. Pattern correlation is the Pearson correlation coefficient between the decoded feature patterns and the true feature patterns for a test stimulus, while profile correlation calculates the Pearson correlation coefficient between the sequences of the decoded and true feature values across all test stimuli for a single DNN unit. For each individual pair or each subject (Within), the pattern correlation coefficients for all test stimuli were averaged to calculate the mean decoding accuracy (pattern), while the profile correlation coefficients for all DNN units were averaged to calculate the mean decoding accuracy (profile). These mean accuracies were used as data points in group analysis.

6.2.3 Visual image reconstruction

The reconstruction method used in this study was extended from our original deep image reconstruction study (Figure 6.2; Shen et al., 2019b). The pixel values of an input image were optimized to make its image features match the decoded features from brain activity. Following Shen et al. (2019b), we used the feature values before the rectification operation from eight layers (conv1-5 and all fully connected layers). We applied the same natural image prior and extended the loss function by adding a Deep Image Structure and Texture Similarity (DISTS) loss component (Muraki, 2024), which uses spatial characteristics of feature maps to improve the detail of image reconstructions. Given the decoded features from multiple layers, an image was reconstructed by solving the following optimization problem:

$$\mathbf{z}^* = \arg \min_{\mathbf{z}} (\mathcal{L}_{\text{mse}}(\mathbf{z}) + \lambda_{\text{tex}} \mathcal{L}_{\text{tex}}(\mathbf{z}) + \lambda_{\text{str}} \mathcal{L}_{\text{str}}(\mathbf{z})), \quad (6.1)$$

where $\mathbf{z}$ is the latent vector, and $\mathcal{L}_{\text{mse}}(\mathbf{z})$ is the loss originally used in Shen et al. (2019b):

$$\mathcal{L}_{\text{mse}}(\mathbf{z}) = \sum_l^L \gamma_l \|\Psi_l(G(\mathbf{z})) - \mathbf{u}_l\|^2. \quad (6.2)$$

G is the deep generator network (DGN) to enhance the naturalness of the image (Nguyen et al., 2016), and the reconstructed image is obtained as $G(\mathbf{z}^*)$. Ψ_l is the function that maps the image to the DNN feature vector of the l -th layer. $\mathbf{u}_l \in \mathbb{R}^{P_l \times Q_l \times K_l}$ represents the decoded DNN feature vector of the image at the l -th layer, where P_l , Q_l , and K_l denote the width, height, and number of channels of the feature maps in the l -th layer, respectively. γ_l is the parameter that weights the contribution of the l -th layer and is set to be $1/\|\mathbf{u}_l\|^2$. L is the number of DNN layers. $\mathcal{L}_{\text{tex}}(\mathbf{z})$ is the texture similarity loss, and $\mathcal{L}_{\text{str}}(\mathbf{z})$ is the structure loss; together, they constitute the DISTS loss. λ_{tex} and λ_{str} serve as the coefficients of weights for the texture and structural similarity losses, respectively.

If we denote $\hat{\mathbf{u}}_l = \Psi_l(G(\mathbf{z})) \in \mathbb{R}^{P_l \times Q_l \times K_l}$, then the texture similarity $\mathcal{L}_{\text{tex}}(\mathbf{z})$ and $\mathcal{L}_{\text{str}}(\mathbf{z})$ are defined as follows:

$$\mathcal{L}_{\text{tex}}(\mathbf{z}) = - \sum_l^L \alpha_l \frac{1}{K_l} \sum_k \frac{\mu_k(\mathbf{u}_l) \mu_k(\hat{\mathbf{u}}_l) + \varepsilon}{\mu_k(\mathbf{u}_l)^2 + \mu_k(\hat{\mathbf{u}}_l)^2 + \varepsilon}, \quad (6.3)$$

$$\mathcal{L}_{\text{str}}(\mathbf{z}) = - \sum_l^L \beta_l \frac{1}{K_l} \sum_k \frac{\delta_k(\mathbf{u}_l, \hat{\mathbf{u}}_l) + \varepsilon}{\delta_k(\mathbf{u}_l) + \delta_k(\hat{\mathbf{u}}_l) + \varepsilon}, \quad (6.4)$$

where

$$\mu_k(\mathbf{u}_l) = \frac{1}{P_l Q_l} \sum_{p_l, q_l} \mathbf{u}_{p_l, q_l, k_l}, \quad (6.5)$$

$$\delta_k(\mathbf{u}_l) = \frac{1}{P_l Q_l} \sum_{p_l, q_l} (\mathbf{u}_{p_l, q_l, k_l} - \mu_k(\mathbf{u}_l))^2, \quad (6.6)$$

$$\delta_k(\mathbf{u}_l, \hat{\mathbf{u}}_l) = \frac{1}{P_l Q_l} \sum_{p_l, q_l} \mathbf{u}_{p_l, q_l, k_l} \hat{\mathbf{u}}_{p_l, q_l, k_l} - \mu_k(\mathbf{u}_l) \mu_k(\hat{\mathbf{u}}_l). \quad (6.7)$$

Here, $\mathbf{u}_{p_l, q_l, k_l} \in \mathbb{R}$ is a notation meaning the activity value of the (p, q, k) -th element for the l -layer. The α_l and β_l are hyperparameters that signify the weight assigned to each layer, which were tuned using data from a subject excluded from the result analyses. A small positive constant ε is included to avoid numerical instability when the denominator is close to zero. We solved the optimization problem using stochastic gradient descent with momentum with 200 iterations.

6.2.4 Identification analysis

We used pairwise identification analysis to quantify the accuracy of image reconstruction. For each reconstructed image, its features were reshaped into a one-dimensional vector and compared with the true feature vector of the presented image and the false alternative of another image. An identification was considered correct when the correlation coefficient of the reconstructed image's feature vector was greater with the true feature vector than with the false alternative. The false alternative image was sourced from the remaining images in the test dataset. For each reconstructed image, this process was repeated for all remaining test images as the false alternative. For Deeprecon natural images, 49 false alternatives were used per reconstructed image ($50 \text{ images} \times 49 \text{ comparisons} = 2450 \text{ comparisons}$). We defined the identification accuracy for a reconstructed image as the ratio of correct identifications. For each individual pair or each subject (Within), we calculated the mean identification accuracy by averaging the identification accuracies of all reconstructed images, which was used as a data point for group analysis.

6.2.5 Group statistics

We performed group-level statistical inference by showing the group means and confidence intervals in the figures. For DNN feature decoding and image reconstruction in within-individual analyses, each data point represents the mean accuracy for a subject. The mean accuracies from all subjects were used to calculate the group mean and its 95% confidence

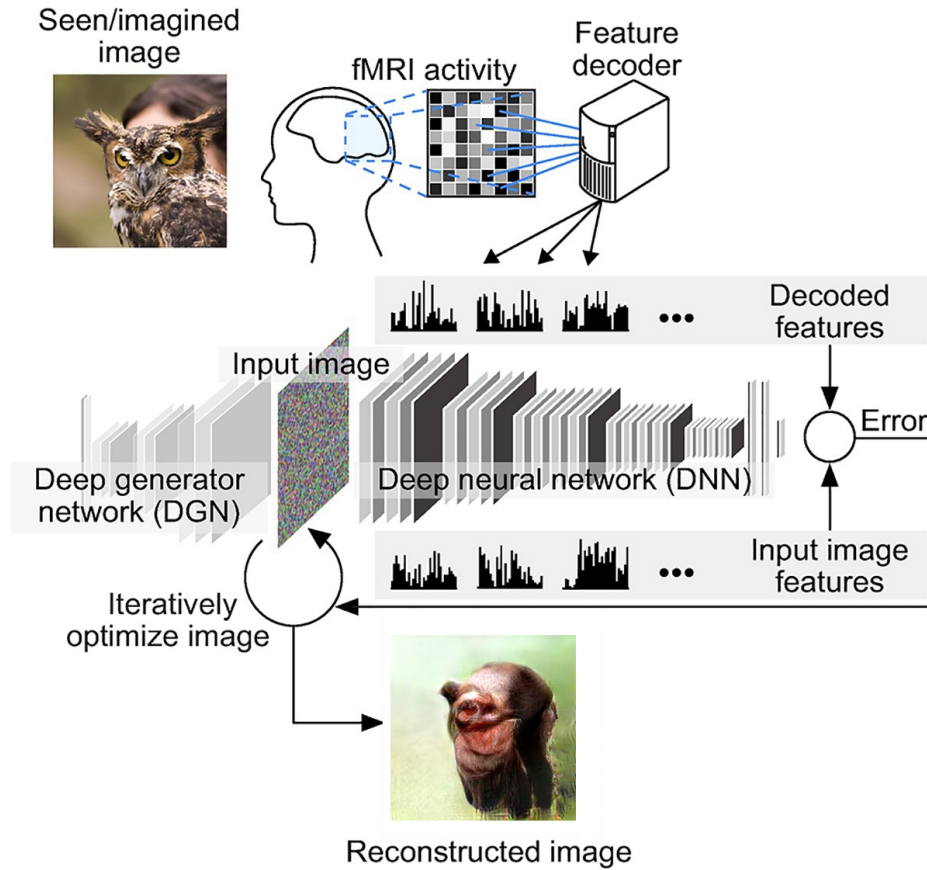


Fig. 6.2 The schematic of the reconstruction method. The brain activity patterns are decoded into DNN features using trained feature decoders, which are further used to reconstruct the image. A reconstructed image is iteratively optimized so that its DNN features align with the decoded ones. Adapted from Shen et al. (2019b), used under CC BY 4.0 license.

interval. For conversion accuracy, DNN feature decoding, and image reconstruction in inter-individual analyses, each data point represents the mean accuracy corresponding to a unique pair of subjects. Due to potential correlations between pairs involving the same subject (dyadic dependency), we applied bootstrapping to these dyadic data points to calculate the group mean and 95% confidence interval. We performed bootstrap sampling separately on the source and target IDs, obtaining data points of the original sample size while ensuring that pairs where the source and target IDs were the same were excluded. The obtained data points were then used to calculate the mean. This process was repeated 1000 times to generate bootstrap replicates, from which we calculated the 95% confidence interval.

6.3 Results

6.3.1 Inter-individual decoding

We have proposed the content loss-based neural code converter and demonstrated that the method can accurately convert brain activities across subjects in Chapter 5. We next investigated whether fine-grained feature representations of stimuli were preserved in the converted fMRI activity patterns by a DNN feature decoding analysis (Horikawa & Kamitani, 2017). Feature decoders were trained to predict hierarchical DNN feature values of the stimuli from the target subject’s fMRI activity patterns across the whole VC. These decoders were then tested on the converted brain activity to predict the hierarchical DNN features of the test stimuli (Figure 6.1; see Subsection 6.2.2). We evaluated the decoding accuracy using two metrics: (a) pattern correlation, which is the Pearson correlation coefficient between the pattern of decoded features and the true feature patterns for a test stimulus, and (b) profile correlation, which is the Pearson correlation coefficient between the sequences of the decoded and true feature values across all test stimuli for a single DNN unit. We further averaged the pattern correlations across all test stimuli and the profile correlations across all DNN units within each layer. To provide a baseline for comparison, we calculated the decoding accuracy for the standard within-individual decoding, which predicts DNN features using decoders trained exclusively on data from the same subject, referred to as the “Within” condition.

Figure 6.3 shows the DNN feature decoding accuracy for the whole VC across three analytical conditions: Within, Brain loss, and Content loss. The content loss-based converters exhibited comparable decoding accuracies with those obtained through the within-individual decoding, with consistently similar trends across various DNN layers. The content loss-based converters showed higher accuracy than those using brain loss regarding both pattern correlation and profile correlation. The results suggest that content loss-based converters effectively preserve the fine-grained representation of stimulus features and demonstrate an advantage in decoding hierarchical DNN features.

6.3.2 Inter-individual image reconstruction

Since we have decoded fine-grained DNN feature representations from converted brain activity, we next investigated whether we can reconstruct images from these decoded DNN features (Figure 6.1; see Subsection 6.2.3) and showed examples of the reconstruction of natural images and artificial images from the whole VC (Figure 6.4). The reconstructions of all pairs are presented in Figures 6.5 and 6.6. The reconstructed images obtained from

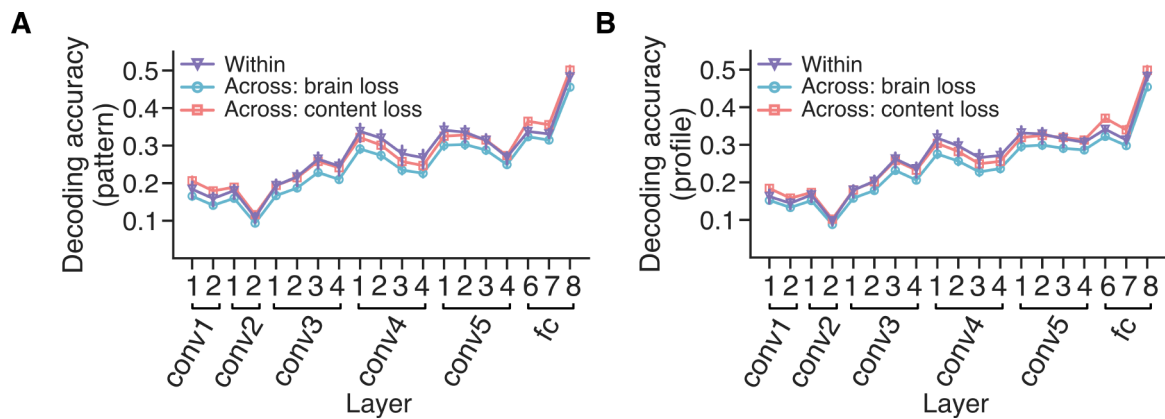


Fig. 6.3 Inter-individual decoding accuracy. (A) Feature decoding accuracy measured by pattern correlation. The DNN features from the convolutional (conv) layers and fully connected (fc) layers of the VGG19 model are analyzed. The mean pattern correlation for each layer of the VGG19 model is shown for the Within, Brain loss, and Content loss conditions (VC; error bars, 95% C.I. from five subjects for the Within condition, and from 20 pairs for the Brain loss and Content loss conditions). (B) Feature decoding accuracy measured by profile correlation. The three conditions are compared as in (A) but with profile correlation coefficients.

converters using the brain loss and content loss, captured the essential characteristics and details of the presented images, with visual objects being similarly recognizable to those in the within-individual reconstructions. Additionally, we reconstructed images from different ROIs to investigate whether the reconstruction could be performed from the subareas. Compared to reconstructions from the VC, those from subareas show a slight degradation in quality but still reflect the main characteristics of the presented images (Figures 6.7 and 6.8).

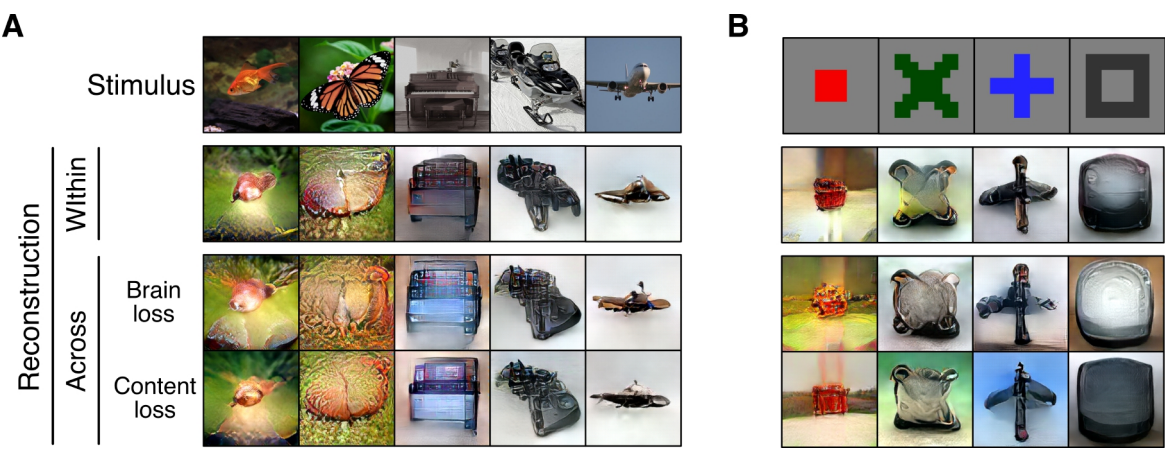


Fig. 6.4 Visual image reconstruction. (A) Reconstructions of natural images. The reconstructions under the three analytical conditions for each stimulus image were all from the same source subject (VC; source: Subject 1, target: Subject 2). (B) Reconstructions of artificial images.

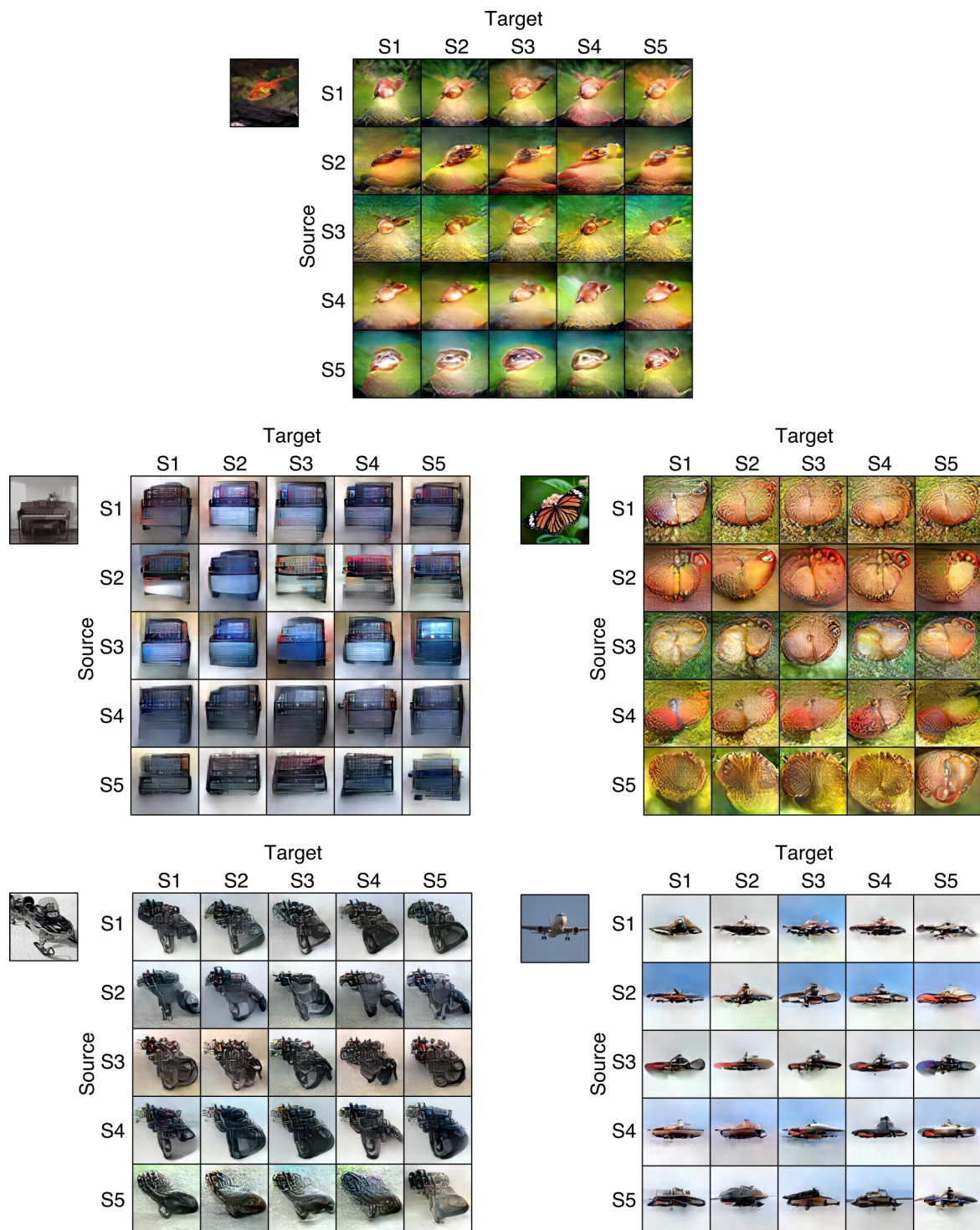


Fig. 6.5 Reconstructed natural images of all pairs. The reconstructions are shown for each pair of subjects for each stimulus image. The diagonal images in each block are reconstructed in the Within condition. All reconstructions were generated using the whole visual cortex (VC).

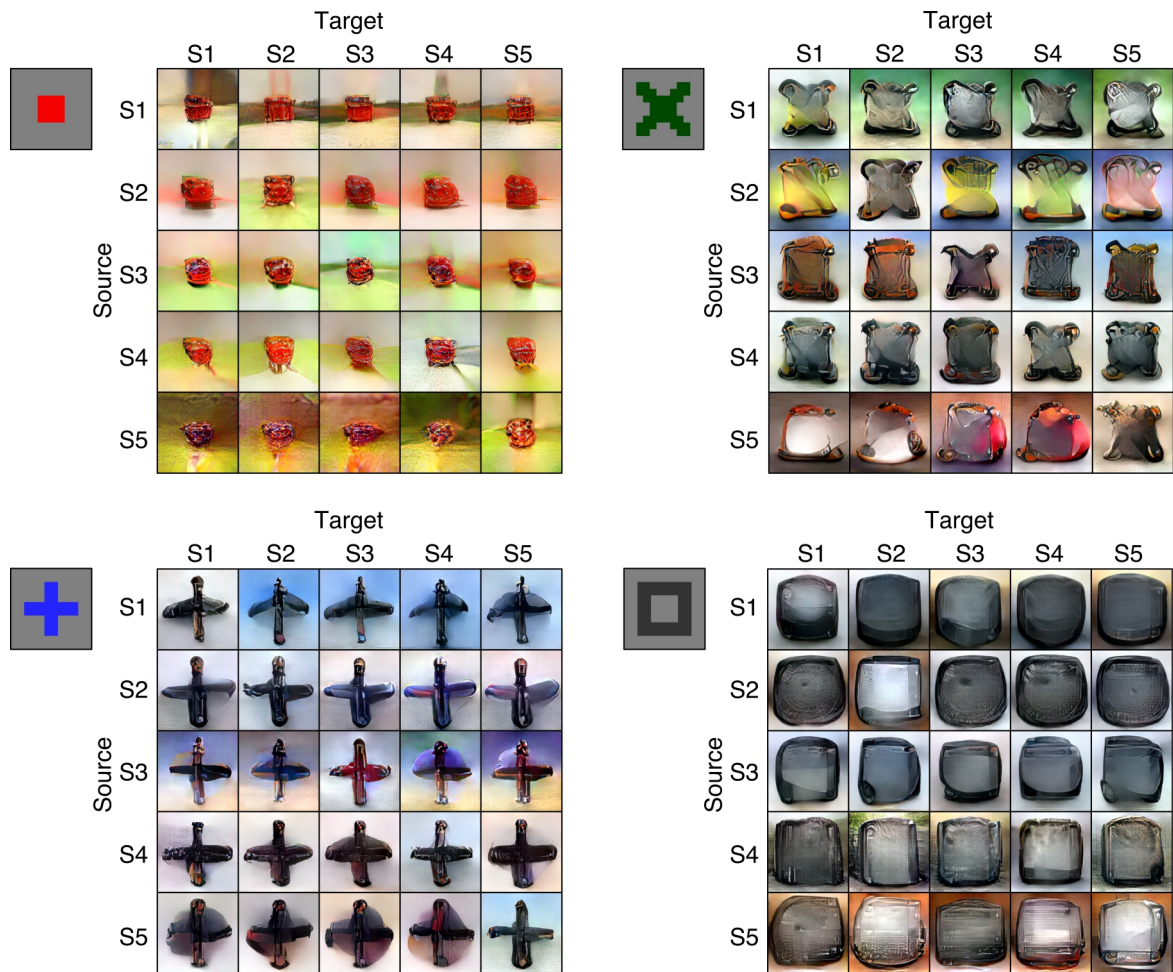


Fig. 6.6 Reconstructed artificial images of all pairs. The reconstructions are shown for each pair of subjects for each stimulus image. The diagonal images in each block are reconstructed in the Within condition. All reconstructions were generated using the whole visual cortex (VC).

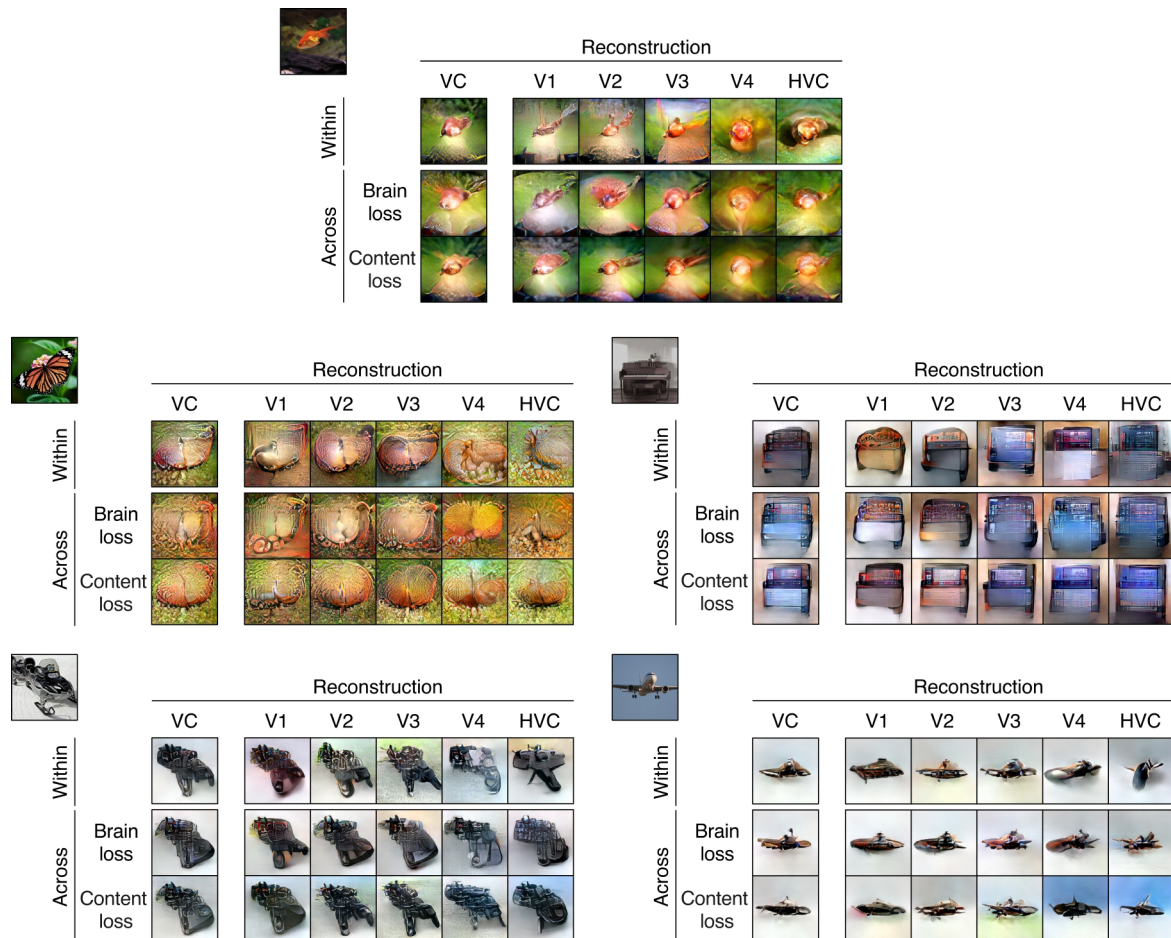


Fig. 6.7 Reconstructed natural images from different ROIs. Reconstructions were generated from different ROIs in the Within, Brain loss, and Content loss conditions. All reconstructions are from the same source subject (VC; source: Subject 1, target: Subject 2).

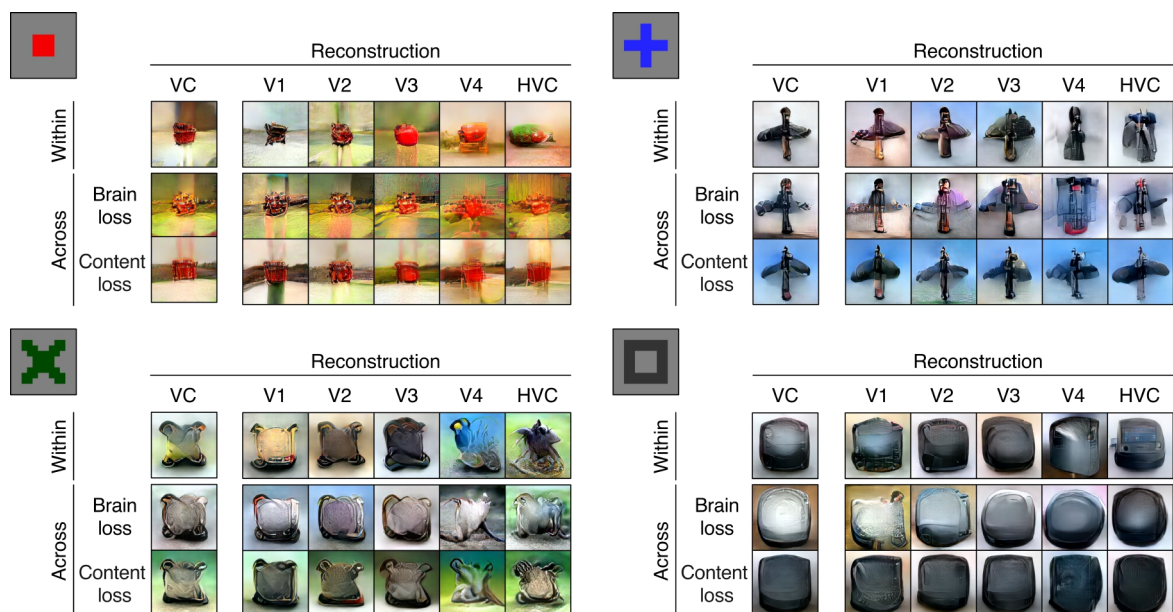


Fig. 6.8 Reconstructed artificial images from different ROIs. Reconstructions were generated from different ROIs in the Within, Brain loss, and Content loss conditions. All reconstructions are from the same source subject (VC; source: Subject 1, target: Subject 2).

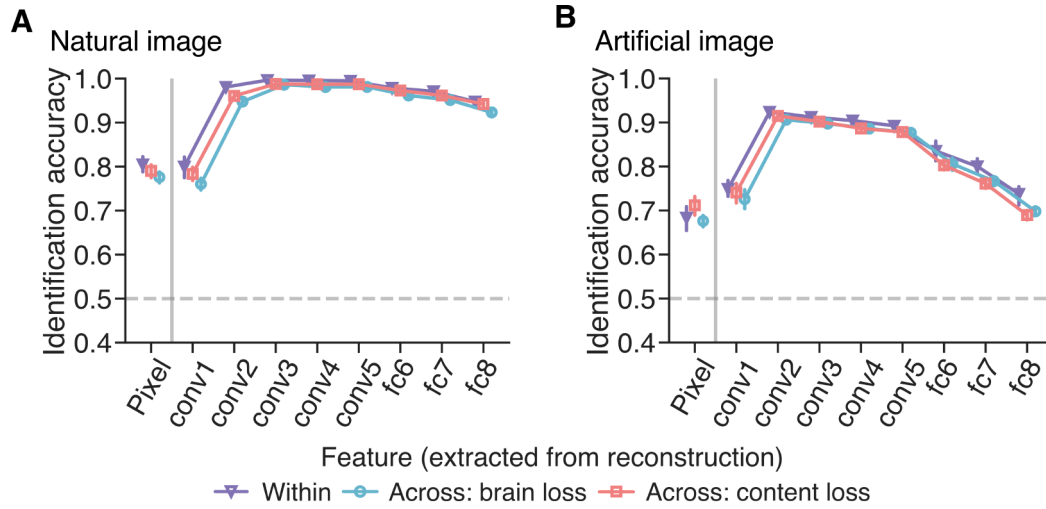


Fig. 6.9 Identification accuracy. (A) Identification accuracy of natural images. Pairwise identification was performed using the pixel values and the extracted DNN feature values (AlexNet) from the reconstructions. The mean identification accuracy was calculated over all reconstructed images for each subject or individual pair (error bars, 95% C.I. from five subjects or 20 pairs; dashed lines, 50% chance level). (B) Identification accuracy of artificial images.

We performed a quantitative evaluation of reconstruction performance from the whole VC under three analytical conditions: Within, brain loss, and content loss. We used a pairwise identification analysis, and this analysis used the pixel and DNN feature patterns of the reconstructed image to identify the true stimulus between two alternatives (see Subsection 6.2.4). We used the AlexNet model (Krizhevsky et al., 2012) to extract DNN feature patterns. The identification process was repeated with multiple false alternatives to obtain the accuracy for each reconstructed image. We then calculated the mean identification accuracy across all reconstructions for each individual pair, presenting these accuracies at the group level in Figure 6.9. Features extracted from reconstructions achieved higher identification accuracy at the mid-level layers, while correlation-based evaluation of decoded features showed higher performance at the high-level layers (Figure 6.3). We also performed a similar correlation analysis on the extracted features from reconstructions to provide a complementary insight into this discrepancy. The correlation analysis of extracted features from reconstructions also showed higher predictions at the high-level layers (Figure 6.10). This discrepancy between identification and correlation metrics can arise because successful identification requires features that distinctly differentiate between stimuli, whereas high correlations can occur even when predictions capture only common patterns across stimuli (Horikawa & Kamitani, 2017).

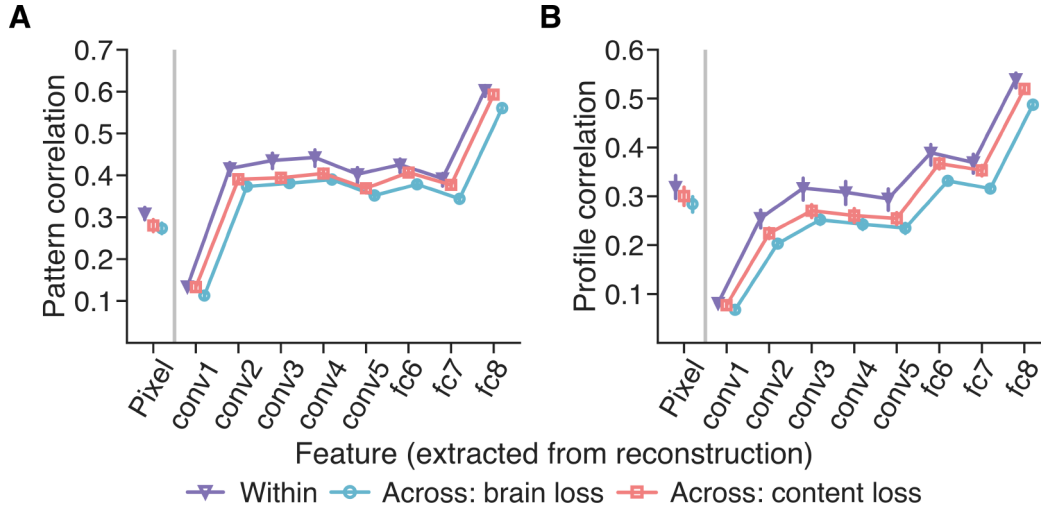


Fig. 6.10 Evaluation of reconstructions based on feature correlations. (A) Pattern correlation of natural images. Pattern correlation between the reconstruction and the presented image was calculated using the pixel values and the extracted DNN feature values (AlexNet) from the reconstructions. The mean pattern accuracy was calculated over all reconstructed images for each subject or individual pair (error bars, 95% C.I. from five subjects or 20 pairs; dashed lines, 50% chance level). (B) Profile correlation of natural images. The three conditions are compared as in (A) but with profile correlation.

In identification accuracy, we observed that both types of inter-individual converters achieved comparable but slightly lower identification accuracies than within-individual reconstruction (Figure 6.9). Among the inter-individual approaches, the content loss-based converter demonstrated better performance than the brain loss-based converter. This superiority was consistent in the evaluation of reconstructions based on feature correlations, where the content loss-based converter achieved higher performance across all layers (Figure 6.10).

6.3.3 Inter-individual decoding and image reconstruction using different alignment methods

In addition to the content loss- and brain loss-based neural code converters, we employed standard functional alignment methods as additional baselines for evaluating conversion accuracy (see Subsection 5.2.3). Here, we further assessed inter-individual decoding accuracy and image reconstruction performance using these methods. Figure 6.11 presents the mean decoding accuracy across DNN layers for the content loss-based converter, brain loss-based converter, pairwise Procrustes transformation, and template-mediated Procrustes transformation. Among these, the content loss-based converter achieved the highest decoding accuracy, followed by the brain loss-based converter. The pairwise and template-mediated

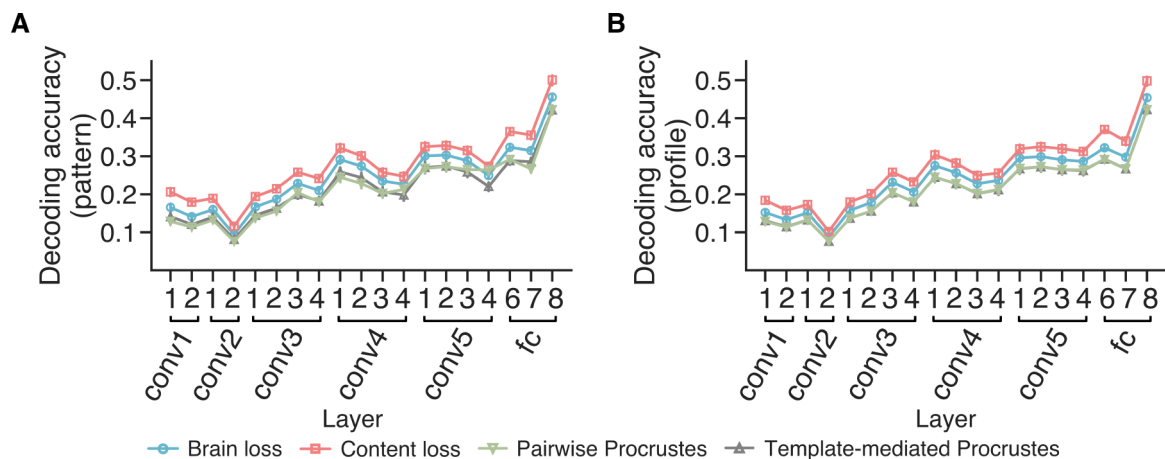


Fig. 6.11 Inter-individual decoding using different functional alignment methods. (a) Feature decoding accuracy measured by pattern correlation. The mean pattern correlation for each layer is shown for the brain loss-based converter, content loss-based converter, pairwise Procrustes transformation, and template-mediated Procrustes transformation (VC; error bars, 95% C.I. from 20 pairs for each condition). (B) Feature decoding accuracy measured by profile correlation. The four conditions are compared as in (A) but with profile correlation coefficients.

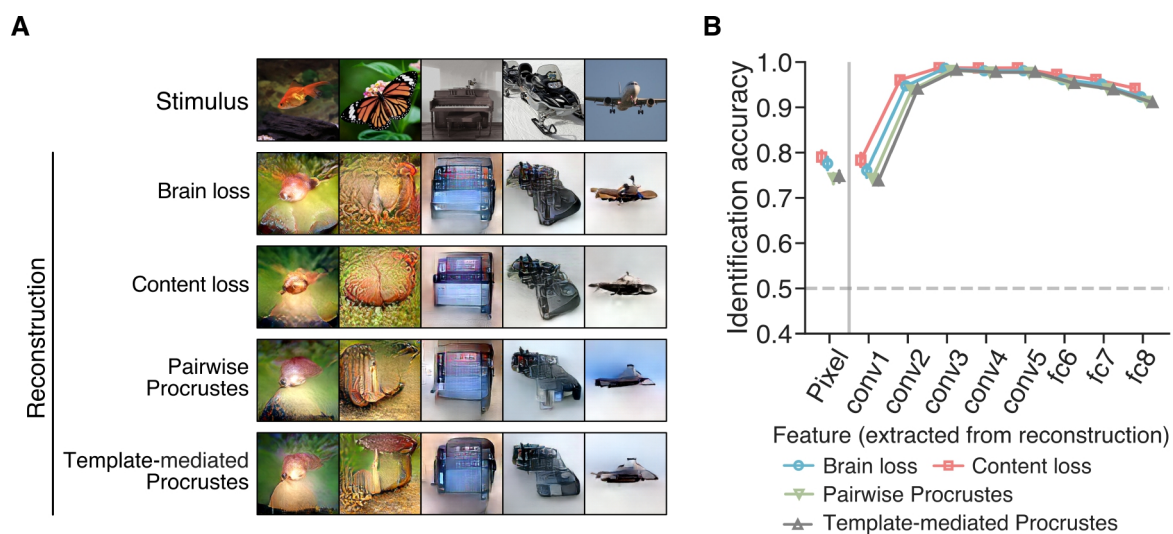


Fig. 6.12 Inter-individual image reconstruction using different functional alignment methods. (A) Reconstructions of natural images. The reconstructions under the four analytical conditions for each stimulus image were all from the same source subject (VC; source: Subject 1, target: Subject 2). (B) Identification accuracy of natural images. Pairwise identification was performed using the pixel values and the extracted DNN feature values (AlexNet) from the reconstructions. The mean identification accuracy was calculated over all reconstructed images for each subject or individual pair (error bars, 95% C.I. from 20 pairs; dashed lines, 50% chance level).

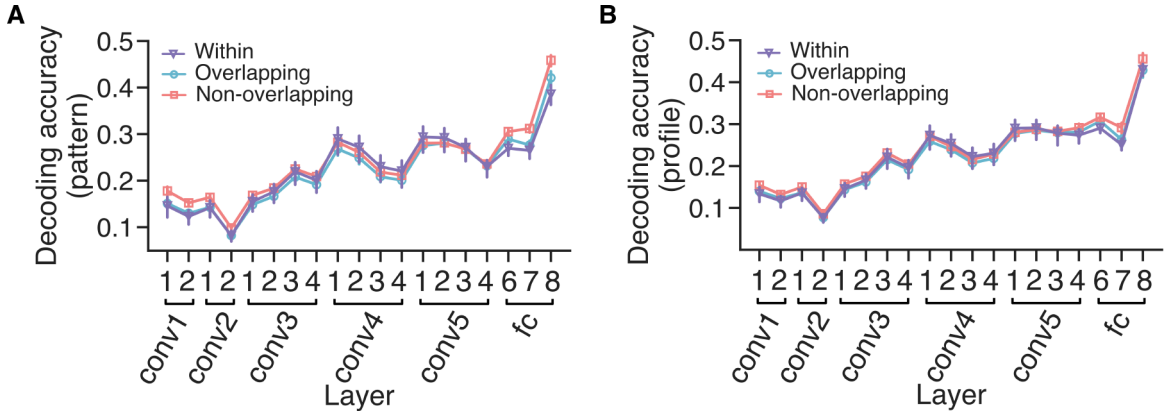


Fig. 6.13 Inter-individual decoding without overlapping stimuli. (A) Feature decoding accuracy measured by pattern correlation. The mean pattern correlation for each layer of the VGG19 model is shown for the Within, Overlapping, and Non-overlapping conditions (VC; error bars, 95% C.I. from five subjects for the Within condition, and from 20 pairs for the Overlapping, and Non-overlapping conditions). (B) Feature decoding accuracy measured by profile correlation. The three conditions are compared as in (A) but with profile correlation coefficients.

Procrustes transformations yielded comparable results and exhibited the lowest overall accuracy. Despite differences in absolute performance, all four methods displayed a consistent decoding pattern: lower accuracy in low-level DNN layers and higher accuracy in high-level layers. The reconstructed images across these methods preserved similar visual content, such as object color and shape. Quantitative evaluations further confirmed that the content loss-based converter produced the most accurate reconstructions, while other standard functional alignment methods achieved slightly lower accuracies (Figure 6.12).

6.3.4 The effect of overlapping (shared) stimuli between converter and decoder trainings

We examined whether the method remains effective without shared stimuli, by evaluating conversion accuracy with both overlapping and non-overlapping stimuli between converter and decoder trainings in Chapter 5 (see Subsection 5.2.6). Here, we further investigate whether the fine-grained visual features are preserved for the converters trained using non-overlapping stimuli.

As shown in Figure 6.13, converters using non-overlapping stimuli show a similar decoding tendency across all DNN layers to those using overlapping stimuli, with accuracy levels approaching those observed in the within-individual decoding. The images reconstructed under two conversion conditions were recognizable and closely matched the presented stimuli

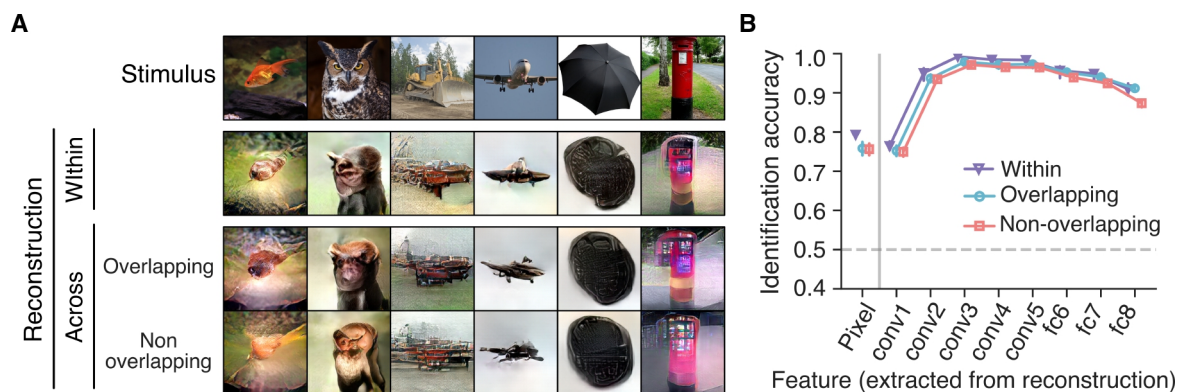


Fig. 6.14 Inter-individual reconstruction without overlapping stimuli. (A) Reconstructions. The reconstructions under three analytical conditions for each stimulus were all from the same source subject (VC; source: Subject 1 and Subject 2). (B) Identification accuracy. Pairwise identification was performed using the pixel values and the extracted DNN feature values from the reconstructions. The mean identification accuracy was calculated over all reconstructed images for each subject or individual pair (error bars, 95% C.I. from 20 pairs; dashed lines, 50% chance level).

as well as the within-individual reconstructions (Figure 6.14A). Quantitative evaluations of image reconstructions indicate that using either overlapping or non-overlapping stimuli between converter and decoder trainings results in similar identification accuracy (Figure 6.14). These results demonstrated that the content loss-based converters can convert brain activity patterns across subjects without the need for overlapping (shared) stimuli, even between decoder and converter trainings.

6.4 Discussion

This chapter explored whether fine-grained visual features are preserved using the content loss-based neural code conversion. By decoding converted brain activity into hierarchical DNN features and using these features for image reconstruction, we assessed to what extent the fine-grained visual information is preserved during inter-individual conversion. Our results demonstrate that converters optimized using visual content can effectively capture fine-grained visual details, achieving reconstruction performance comparable to converters trained with paired brain activity. Moreover, we found that high-quality inter-individual image reconstruction is attainable even when the converter and decoder are trained using non-overlapping stimuli, suggesting that shared stimuli are not necessary for our proposed approach.

In contrast, conventional brain loss-based functional alignment methods yielded slightly lower decoding and reconstruction performance, despite reconstructing images with similar overall contents. This discrepancy may arise from the inherent limitations of brain-based alignment. A brain's response to a stimulus comprises three components: a consistent stimulus-evoked response across individuals, an idiosyncratic stimulus-evoked response, and a noise component (Nastase et al., 2019). As a result, brain loss-based alignment methods might inadvertently incorporate idiosyncratic responses and noise, potentially limiting their effectiveness in capturing the consistent stimulus-evoked response. This could explain the inferior performance observed in Figures 6.3 and 6.9. In contrast, our proposed method may help mitigate the influence of idiosyncratic responses and noise components, leading to a more accurate representation of the consistent stimulus-evoked response across individuals.

Chapter 7

Hierarchical levels of content representations

7.1 Introduction

Perception is hierarchically represented in the human brain, with early visual areas encoding low-level features such as edges and textures, and higher-order areas processing more abstract, semantic representations. Deep neural networks (DNNs), particularly models like VGG19 (Simonyan & Zisserman, 2014), exhibit similar hierarchical feature structures and have therefore been widely adopted as computational models for visual processing in the brain. In the context of neural code conversion, this correspondence raises a critical question: which levels of the DNN feature hierarchy are most essential for visual content and neural code conversion? In previous chapters, we demonstrated that inter-individual neural code conversion optimized using content loss from all DNN layers can successfully preserve fine-grained visual information and support accurate decoding and image reconstruction. However, the contribution of each hierarchical level of content representation to the success of neural code conversion remains unclear.

This chapter investigates how visual features extracted from different levels of the DNN hierarchy contribute to inter-individual neural code conversion. We trained separate converters using content loss from individual DNN layers and compared their performance to that of a converter trained with content loss from all layers. Additionally, we examined whether the choice of converter architecture affects the conversion performance when the hierarchical DNN features were used as the visual contents. Through decoding, reconstruction, and identification analyses, we aimed to reveal the critical hierarchical levels that drive successful

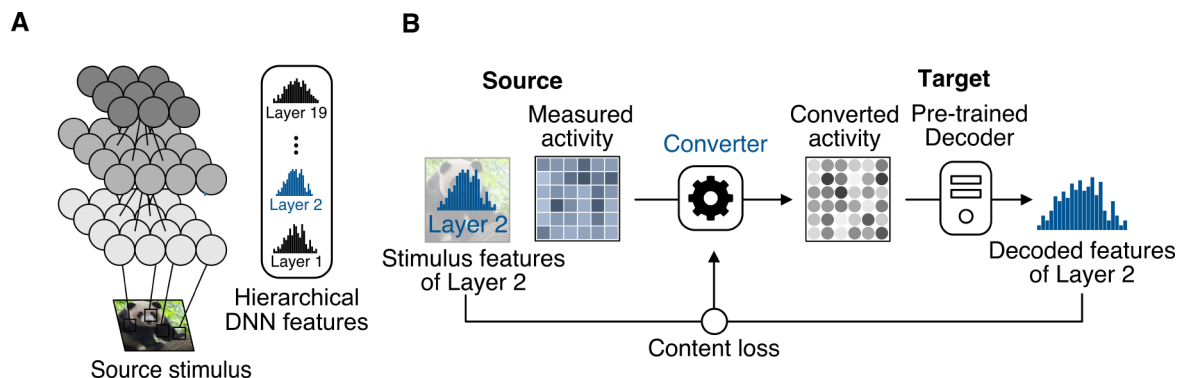


Fig. 7.1 Converter training using loss from different DNN layers. (A) Hierarchical DNN features are extracted from the stimulus across 19 layers. Each layer can be used for converter training, for example, Layer 2. (B) An example of converter training using the content loss based on the DNN feature from Layer 2. Other layers can be used to train the converter in a similar way.

neural code conversion and assess the robustness of our approach across different model configurations.

7.2 Methods

7.2.1 Converter training using the loss from different DNN layers

To examine how visual information from different hierarchical levels contributes to inter-individual neural code conversion, we trained separate converters using content loss from individual DNN layers (Figure 7.1). Unlike the previous approach in Chapter 5, which jointly optimized across all layers, each converter here was trained using the loss from a single DNN layer. Each layer-specific converter used the same architecture as in Chapter 5 for fair comparison and was trained to minimize the difference between decoded specific-layer features from converted brain activity and the true specific-layer features of the source subject's stimulus. We applied this training procedure to each layer of the VGG19 model, including both convolutional and fully connected layers. Finally, we compared the performance of these layer-specific converters with that of the converter trained using the loss from all DNN layers, to identify the most useful layers for inter-individual conversion.

7.2.2 Robustness across content-loss based converter architecture

We used the nonlinear MLP as the default architecture for the content loss-based converter. This model consists of two hidden layers with instance normalization and ReLU activation functions, and each hidden layer contains half the number of units as the input layer.

In addition to this default architecture, we examined several alternative architectures to assess the robustness of converter architecture with hierarchical DNN feature as the content representation. The first is a Linear Multi-Layer Perceptron (MLP), which shares the same structure as the nonlinear MLP (i.e., two hidden layers and the same number of units), but without any nonlinear normalization and activation functions. By removing nonlinear operations, this model serves as a purely linear baseline, allowing us to isolate the impact of nonlinearity on conversion performance. The second variant is a Residual MLP, which extends the original nonlinear MLP architecture by integrating three residual blocks (He et al., 2016). Each block includes shortcut connections that enable deeper learning while maintaining efficient gradient flow, effectively allowing the converter to capture more complex mappings between source and target brain activity patterns. The third architecture is a Vision Transformer (ViT)-based converter, adapted from the model proposed by Beyer et al. (2022). In this model, the source subject's brain activity pattern is first patchified into fixed-length patches, like image patches in standard ViT applications. These patches are then processed using a transformer architecture that models long-range dependencies and relationships across the entire input. The input and output layers were modified to accommodate voxel-wise brain activity vectors.

By comparing these architectures, we aimed to evaluate the extent to which the architectural choices impact the effectiveness of neural code conversion when optimized using the loss from all DNN layers.

7.2.3 Group statistics

For the converters using content loss from individual DNN layers, we performed group-level statistical inference by showing the group means and confidence intervals of the conversion accuracy, decoding accuracy, and identification accuracy. Each data point represents the mean accuracy corresponding to a unique pair of subjects. Similar to Subsection 5.2.7 and Subsection 6.2.5, we applied bootstrapping to these dyadic data points due to dyadic dependency to calculate the group mean and 95% confidence interval.

7.3 Results

7.3.1 Converter training using the loss from different layers

To examine the importance of the use of multiple hierarchical DNN layers in converter training, we compared the conversion performance using the loss from different sets of DNN layers. Figure 7.2 compares the conversion accuracy between these converters. For pattern correlation, the loss from all layers results in the highest accuracy. However, it was observed that the loss from all layers resulted in the highest accuracy. The conversion accuracies were lower when using loss from the low-level and mid-level layers (conv1_1 to conv5_4). The high-level layers, specifically fc6, fc7, and fc8, showed the worst conversion accuracy. Similar results were observed in the evaluation of profile correlation.

Figures 7.3 and 7.4 show the decoding accuracy of converters trained with the loss from each individual DNN layer, with the performance of the converter trained with the loss from all layers (the “All” condition). Decoding was evaluated in each DNN layer. The converter trained with all layers consistently achieved the highest decoding accuracy across all layers, serving as an upper bound. When a converter was optimized using the loss from a specific DNN layer, it performed best in decoding that corresponding layer, approaching the performance of the “All” condition. These results were consistent for both pattern correlation and profile correlation analyses, suggesting that optimizing the converter with loss from all DNN layers helps preserve hierarchical visual information.

Figure 7.5 shows the images reconstructed by converters trained using the loss from different DNN layers. Images reconstructed from the “All layers” condition produced visual contents with high perceptual quality. In the case of converters trained using the loss from single DNN layers, specifically layers from the lower and middle levels, the reconstructions also yielded recognizable visual contents. However, those trained with the high-level layers were found to be less effective in capturing the details of the objects, leading to reconstructed images with lower perceptual quality. Quantitative evaluations presented in Figure 7.6 reflected this qualitative assessment. The reconstructions from the “All layers” condition achieved higher identification accuracy compared to those from each single-layer condition, ranging from conv4_1 to fc8 layers. These results indicate that the use of the loss from all DNN layers is important for accurate neural code conversion.

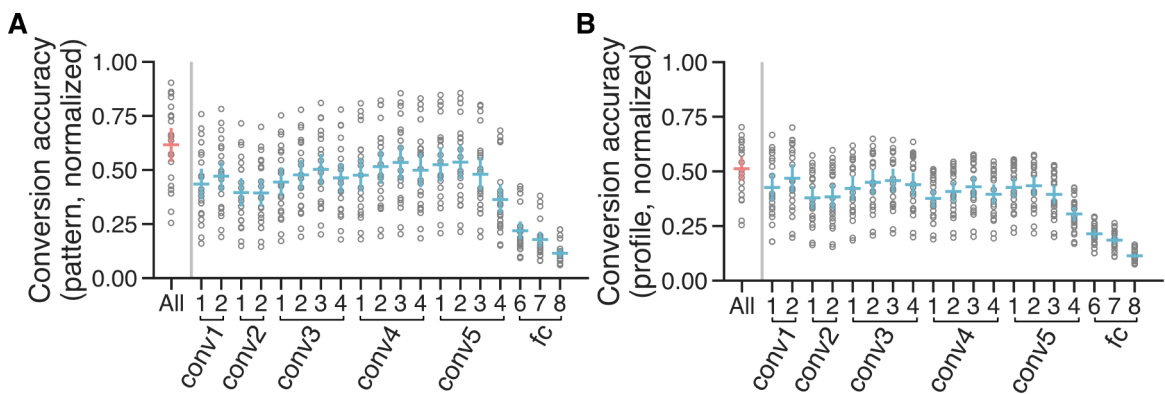


Fig. 7.2 Converters trained using the loss from different DNN layers. (A), Conversion accuracy measured by pattern correlation. To examine the significance of the use of multiple hierarchical DNN layers in converter training, we compared the conversion performance using the loss from different sets of DNN layers, including convolutional (conv) layers and fully connected (fc) layers of the VGG19 model. The pattern correlation coefficients calculated for 20 individual pairs (VC) are used to compare these converters trained using loss from different DNN layers. The horizontal bars represent the mean accuracies across the 20 pairs, with vertical bars representing the 95% confidence interval (C.I.), and each dot indicates an individual pair's mean correlation coefficient over stimuli. The results are shown with the converter trained using loss from all layers (All). (B) Conversion accuracy measured by profile correlation. The comparison is shown as in (A) but with profile correlation coefficients.

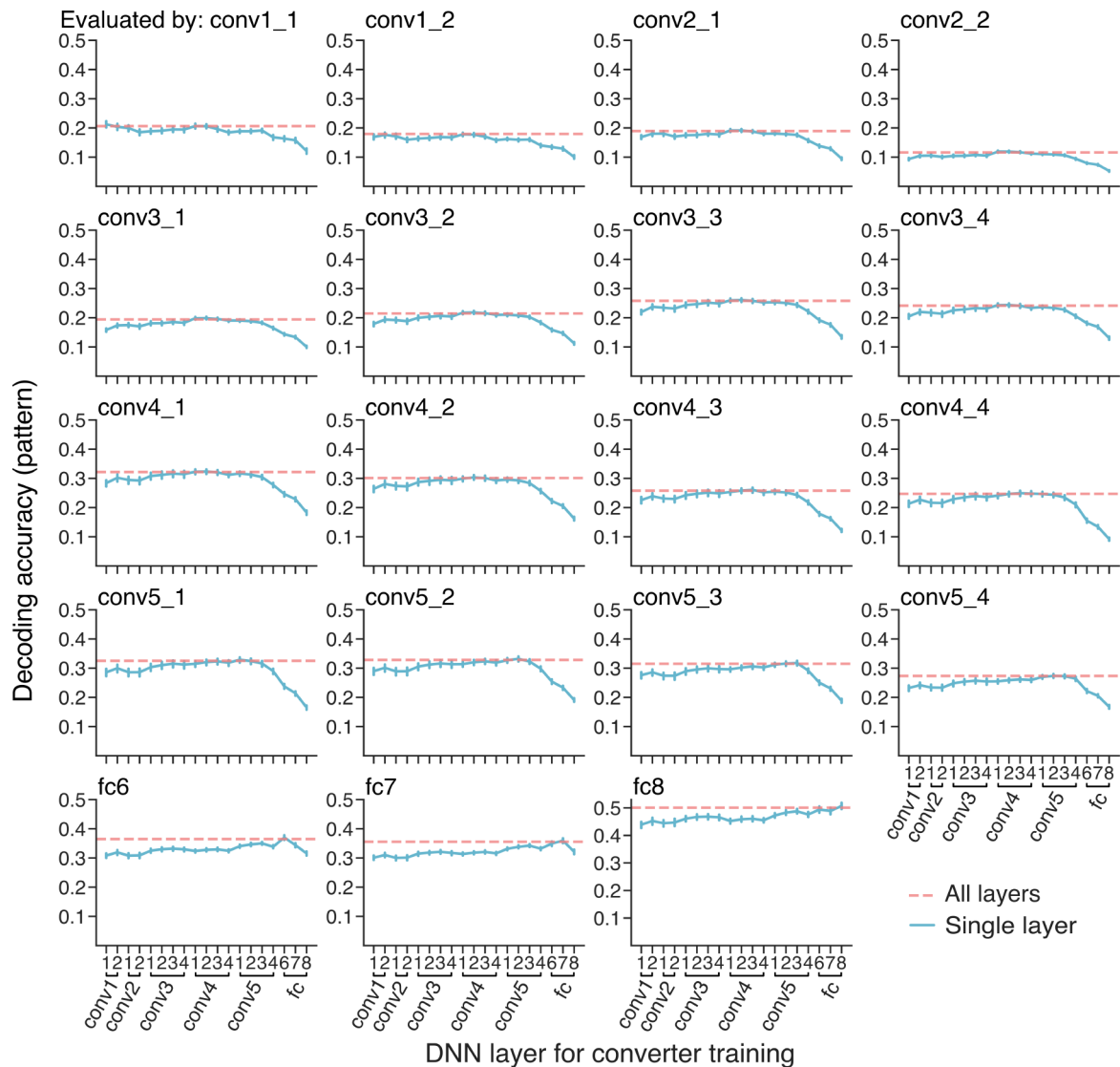


Fig. 7.3 DNN feature decoding performance (pattern) with converters trained using the loss from different DNN layers. Converters were trained using the loss from different layers of the VGG19 model. The “All layers” refers to converters trained using the loss from all DNN layers, while the “single layer” corresponds to converters trained using the loss from single DNN layers. The feature decoding performance of the converted brain activity was evaluated by the mean pattern correlation coefficient at each layer of the VGG19 model (VC; error bars, 95% C.I. from 20 pairs).

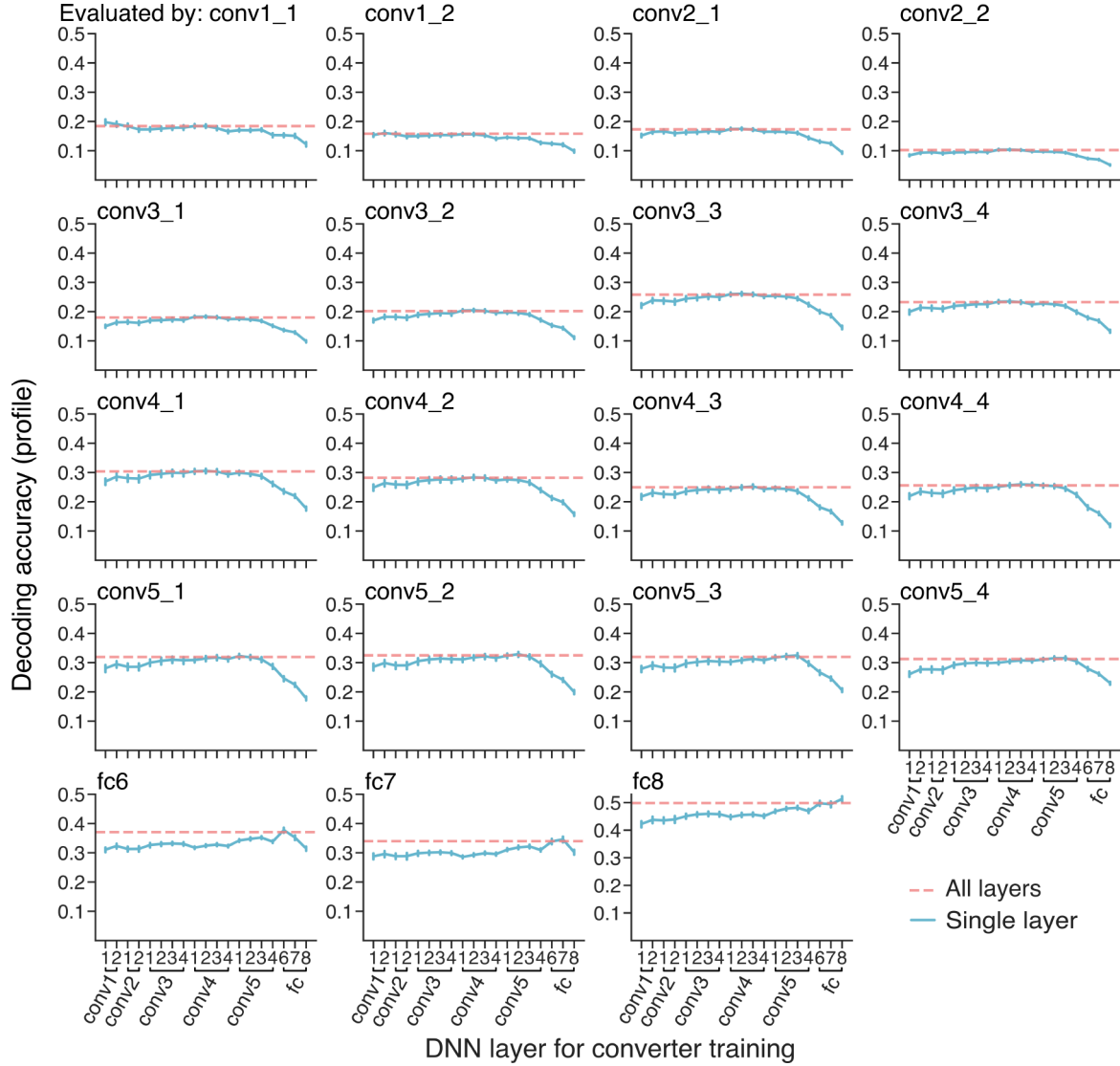


Fig. 7.4 DNN feature decoding performance (profile) with converters trained using the loss from different DNN layers. The decoded features, obtained from the converters and decoders described in Figure 7.3, were evaluated using profile correlation coefficients. The feature decoding performance was evaluated by the mean profile correlation coefficient at each layer of the VGG19 model (VC; error bars, 95% C.I. from 20 pairs).



Fig. 7.6 Identification accuracy of reconstructions with converters trained using the loss from different DNN layers. Pairwise identification was performed using the pixel values and the extracted DNN feature values (AlexNet) from the reconstructions. The “All layers” refers to converters trained using the loss from all DNN layers, while the “single layer” corresponds to converters trained using the loss from single DNN layers. The mean accuracy was calculated in each condition (VC; error bars, 95% C.I. from 20 pairs; dashed lines, 50% chance level).

7.3.2 Robustness of converter architecture

After confirming the importance of using multiple hierarchical DNN layers in converter training, we further investigated whether the converter architecture influences performance when optimized with the loss from all DNN layers. Figure 7.7 compares the conversion accuracy of four architectures: Nonlinear MLP, Linear MLP, Residual MLP, and Visual Transformer (ViT). In terms of pattern correlation, the linear MLP achieved the highest accuracy, whereas the residual MLP showed slightly lower accuracy. In contrast, all four architectures yielded comparable results in profile correlation, suggesting that they are similarly effective in capturing the overall response patterns.

All four architectures produced similar reconstructed images that captured key visual characteristics of the presented stimuli, such as object shape and color. Among them, the ViT-based converter showed slightly lower perceptual quality (Figure 7.8A). This observation was supported by the quantitative evaluation (Figure 7.8B), where the ViT-based architecture yielded the lowest identification accuracy, while the three MLP-based architectures demonstrated comparable levels of identification accuracy. Overall, these results suggest that all evaluated architectures are capable of effectively converting brain activity patterns across subjects and yielding recognizable inter-individual reconstructions, demonstrating the robustness of the training neural code converters approach using hierarchical levels of DNN features.

7.4 Discussion

This chapter examined how hierarchical DNN features influence the effectiveness of content loss-based neural code conversion. By systematically training converters using content loss from individual DNN layers, we demonstrated that the use of multiple hierarchical DNN layers significantly enhances conversion performance. Specifically, converters trained with loss from all layers of the VGG19 model consistently achieved better decoding accuracy and reconstruction quality compared to those trained on single layers.

Our content loss-based converter used hierarchical DNN features as latent representations for stimulus content. This method builds upon recent studies in brain decoding, which have identified parallels between the hierarchical representations in the human brain and the fine-grained features of DNNs (Horikawa & Kamitani, 2017). By reconstructing visual images from converted brain activity, we demonstrated that the converter using the loss from all layers indeed preserved fine-grained visual features (Figures 7.5 and 7.6). Furthermore, we found that the results were robust across various model architectures for the content loss-based converter, including Nonlinear MLP, Linear MLP, Residual MLP, and ViT (Figures 7.7

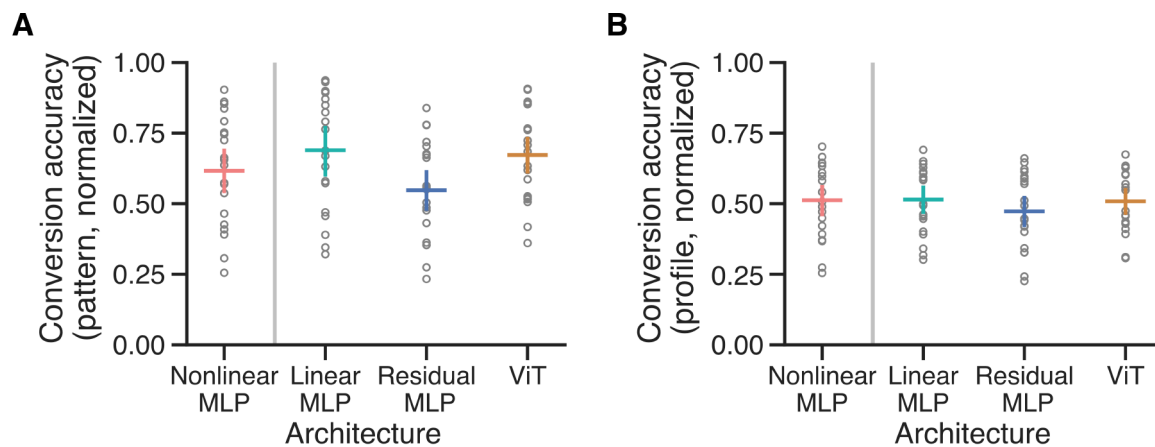


Fig. 7.7 Conversion accuracy using converters consisting of different architectures. (A) Conversion accuracies measured by pattern correlation. Nonlinear MLP, Linear MLP, Residual MLP, and Visual Transformer (ViT) converters are compared using the profile correlation coefficients calculated for 20 individual pairs (VC). The horizontal bars represent the mean accuracies across the 20 pairs, with vertical bars representing the 95% C.I., and each dot indicates an individual pair's mean correlation coefficient over stimuli. (B) Conversion accuracies measured by profile correlation. The four architectures are compared as in (A) but with profile correlation coefficients.

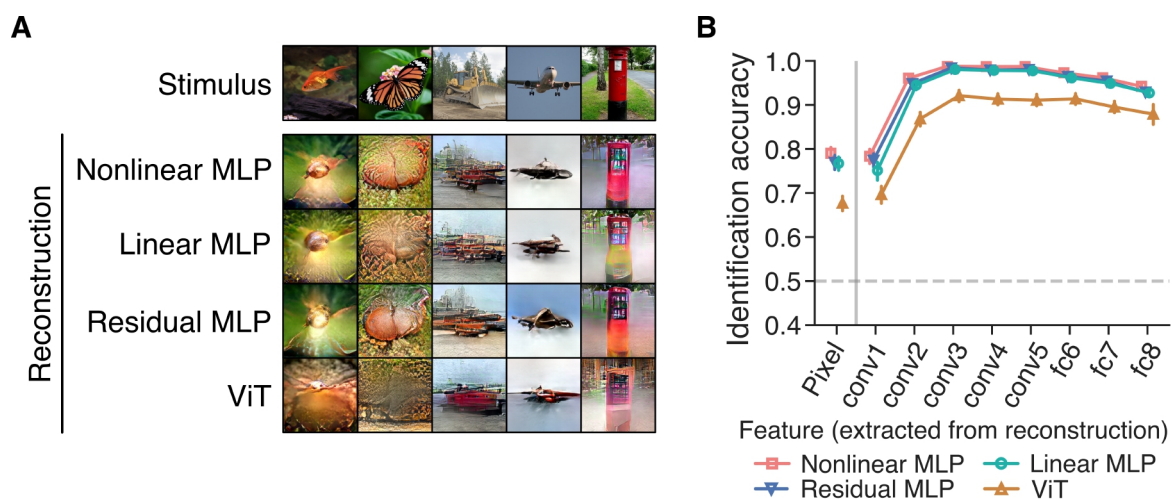


Fig. 7.8 Reconstruction using converters consisting of different architectures. (A) Reconstructions. Examples for source Subject 1 and target Subject 2 (VC) are shown. Some stimulus images have been replaced with visually similar alternatives for copyright reasons. Identification accuracy. The four architectures are compared using the identification accuracies based on pixel and extracted DNN feature values (error bars, 95% C.I. from 20 pairs; dashed lines, chance level = 50%).

and 7.8). However, the ViT-based converter, which divides and processes brain activity patterns in a transformer framework, demonstrated slightly lower performance, potentially due to its greater architectural complexity and data requirements. These findings support the effectiveness and robustness of using hierarchical DNN features as latent representations for stimulus content in neural code conversion.

Optimizing converters using content loss from multiple hierarchical DNN layers contributes to the improvement of neural conversion accuracy. However, when focusing on the quality of inter-individual image reconstructions, we found that using loss from single layers at either the lower or middle levels (specifically from conv1_1 to conv4_4) can match the performance of the loss from all layers (Figures 7.5 and 7.6). Relying on single-layer loss from the high-level DNN layers (fc6, fc7, and fc8) tends to degrade performance. This observation may be due to the substantial capacity of the low-level and mid-level DNN features to the image-level visual information, which is pivotal for reconstructing perceptually coherent images. Converters trained with high-level DNN layers may overlook some fine-grained visual details while preserving more abstract information. These results suggest that integrating visual features across multiple levels is important for accurately converting brain activity patterns, preserving not only image-level information but also capturing the different levels of abstraction in visual processing.

These findings underscore the importance of using hierarchical levels of DNN features in neural code converter training and demonstrate that such training is robust across diverse model architectures.

Chapter 8

Generalizability to different decoding tasks

8.1 Introduction

A central objective of neural code conversion is to enable flexible readout of perceptual content across subjects. Previous chapters have shown that content loss-based conversion using hierarchical DNN features can preserve fine-grained visual representations and support accurate image reconstruction. However, it remains unclear whether the converted brain activity contains generalizable representations that can be interpreted by alternative decoding models or under limited training conditions.

In previous analyses, the readout of converted brain activity was restricted to a specific decoder trained on VGG19 features, which aligns well with the hierarchical organization of the visual cortex. While effective in capturing certain aspects of visual content, this model is optimized for a specific feature space and may fail to reflect all aspects of neural representations. To achieve broad applicability, neural code conversion should yield brain activity patterns that can be interpreted by a range of decoders trained on different feature representations.

This chapter investigates the generalizability of neural code conversion from two perspectives. First, we examine whether brain activity converted with a VGG19-based converter can be decoded using a CLIP-ViT (Radford et al., 2021) decoder that was trained independently. Second, we assess how the amount of training data used to train the converter influences its performance, particularly in scenarios where data availability is constrained. By evaluating decoding accuracy and reconstruction performance under these conditions, we examine

whether content-based neural code conversion can capture generalized neural representations and support flexible inter-individual brain decoding.

8.2 Methods

8.2.1 Decoding the converted brain activity using different schemes

To evaluate whether the converted brain activity preserved generalizable visual representations beyond those optimized for a specific decoder (i.e., the VGG19-based decoder used during converter training), we used an additional decoding scheme using an alternative DNN model, CLIP-ViT (Radford et al., 2021).

The CLIP-ViT model, implemented using the PyTorch library (Paszke et al., 2019), was used to extract hierarchical visual features from natural images. This model was pre-trained on a large set of image-text pairs to enable general-purpose visual recognition without task-specific fine-tuning (available from <https://github.com/openai/CLIP>). Input images were rescaled to 224×224 pixels before feature extraction. The number of units in each layer for the image encoder is as follows: conv1, 37,632; each transformer block (resblocks 0-11), both attention and MLP outputs, 38,400; ln_post, 768; and model_output, 512.

A new feature decoder, independent of the converter, was trained to predict CLIP-ViT feature representations of visual stimuli from the target subject's fMRI responses. Similar to the training approach of the VGG19 decoder, this CLIP-ViT decoder was trained using a linear regression framework and 6000 training samples. After training, the CLIP-ViT decoder was applied to brain activity converted from the source subject's brain space to the target's via the VGG19-based converter in Chapter 5. This setup allowed for an independent evaluation of the generalizability of the converted brain activity, as no CLIP-ViT features were used during the converter training process. The predicted CLIP-ViT features were then used in the same image reconstruction pipeline as the VGG19 features, enabling a direct comparison between the two decoding schemes.

8.2.2 Varying the number of training data for conversion

To evaluate how the amount of training data affects the performance of inter-individual neural code converters, we systematically varied the number of training samples used for converter training. The converters were trained using DNN features extracted from the VGG19 model as the latent representation. Training sets of different sizes were prepared, including 300, 600, 900, 1,200, 2,400, 3,600, 4,800, and 6,000 stimulus-response pairs, corresponding to fMRI data acquisition times ranging from approximately 40 minutes to 13 hours.

For each data size, a separate converter was trained to predict the target subject’s brain activity from the source subject’s brain activity. To ensure fair evaluation, the decoders used for decoding converted brain activities and image reconstruction were fixed across all conditions. Specifically, we used two decoders in the test stage: one trained on VGG19 features and the other on CLIP-ViT features, both trained on the full 6,000 samples of the target subject. This ensured that any variation in reconstruction quality reflected the converter’s performance rather than the decoder’s.

This design allowed us to assess the visual fidelity of reconstructed images under different training sample sizes and the minimal data requirement for effective converter training in practical settings. By varying the training sample sizes, we were able to explore how the amount of data available influences the generalizability of the converter. Specifically, we could determine whether the converter, trained on limited data, still retained the ability to produce robust and generalizable brain activity representations, which could be decoded using pre-trained decoders. This assessment of generalizability is critical for practically applying the conversion method across individuals and tasks, where training data might be scarce.

8.2.3 Group statistics

We performed group-level statistical inference for two types of decoding schemes, by showing the group means and confidence intervals of decoding accuracy and identification accuracy. Each data point represents the mean accuracy corresponding to a unique pair of subjects. Similar to previous subsections, we applied bootstrapping to these dyadic data points due to dyadic dependency to calculate the group mean and 95% confidence interval.

8.3 Results

8.3.1 Decoding the converted brain activity using different schemes

To confirm whether the converted brain activity contains generalizable representations rather than those tailored for a specific readout (the VGG19 decoder), we examined the feasibility of decoding the converted brain activity using a different scheme. A different feature decoder was trained to predict CLIP-ViT features (Radford et al., 2021; see Subsection 8.2.1) of the stimuli from the target subject’s fMRI activity patterns. This decoder was then tested on the converted brain activity obtained from VGG19-based converters. We further reconstructed images from the decoded CLIP-ViT features using the same reconstruction method.

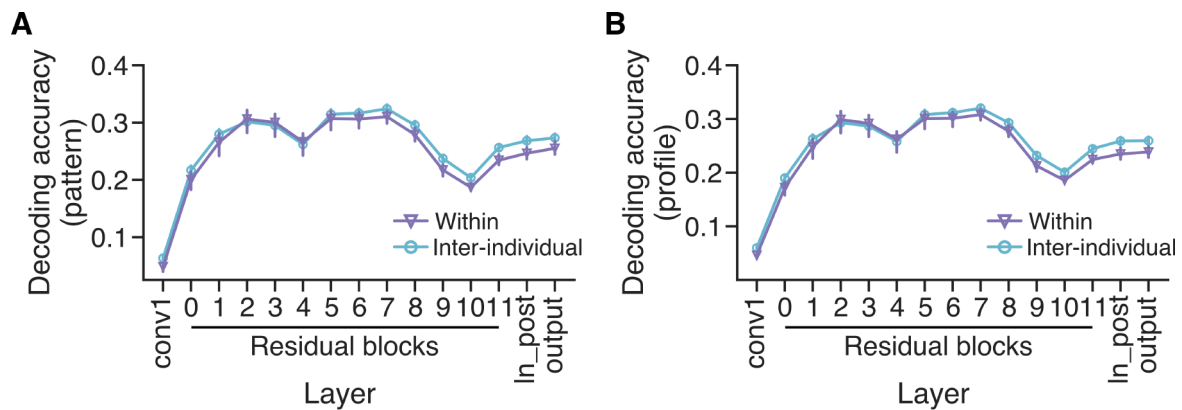


Fig. 8.1 DNN feature decoding performance using the CLIP-ViT decoder. (A) Pattern correlation. The DNN features from various layers of the CLIP-ViT model are used for analysis, including the conv1 layer, transformer blocks 0-11, the In_post layer, and the model output layer. The mean pattern correlation for each layer is shown for both the Within condition and the Inter-individual condition (VC; error bars, 95% C.I. from five subjects for the Within condition and from 20 pairs for the Inter-individual condition). (B) Profile correlation. The two conditions are compared as in (A) but with profile correlation coefficients.

Decoding CLIP-ViT features from converted brain activity demonstrates accuracies similar to those of the within-individual decoding across all evaluated DNN layers (Figure 8.1). Figure 8.2A shows examples of reconstructed images from these decoded CLIP-ViT features (see Figure 8.3 for the reconstructions of all pairs). While CLIP-ViT features were not used during the converter's training, the reconstructions with the CLIP-ViT decoder accurately reflected the visual object in the presented images. The reconstructed contents resemble those obtained with the VGG19 decoder, although the details differ slightly. The quantitative evaluation showed that reconstructions with the CLIP-ViT decoder yielded identification accuracy approaching that of reconstructions with the VGG19 decoder across all evaluations (Figure 8.2B). These results suggest that the brain activity converted using our method is not tailored for the specific decoder and can be read out by different decoding schemes.

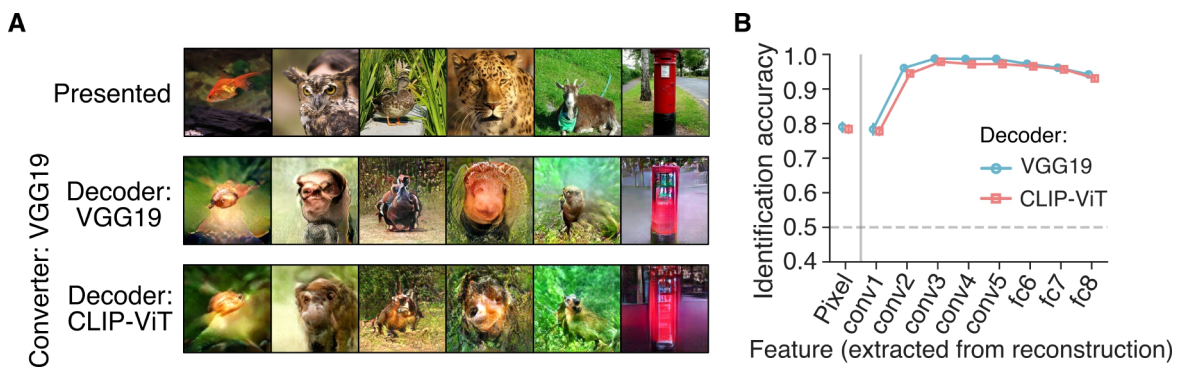


Fig. 8.2 Reconstruction by converter and decoder trained with different DNNs. (A) Reconstructed images using the VGG19 decoder and CLIP-ViT decoder. The source brain activity was converted into the target's space using VGG19-based converters and then evaluated using the VGG19 decoder and CLIP-ViT decoder of the target subject. The reconstructions were generated from the same pair (VC; source: Subject 1, target: Subject 2; see Figure 8.3 for reconstructions of all pairs). Some stimulus images have been replaced with visually similar alternatives for copyright reasons. (B) Identification accuracy. Pairwise identification was performed using the pixel values and the extracted DNN feature values from the reconstructions obtained via the VGG19 decoder or CLIP-ViT decoder. The DNN features were extracted from 5 convolutional (conv) layers and three fully connected (fc) layers of the AlexNet model. The mean identification accuracy was calculated across all individual pairs (error bars, 95% C.I. from 20 pairs; dashed lines, 50% chance level).

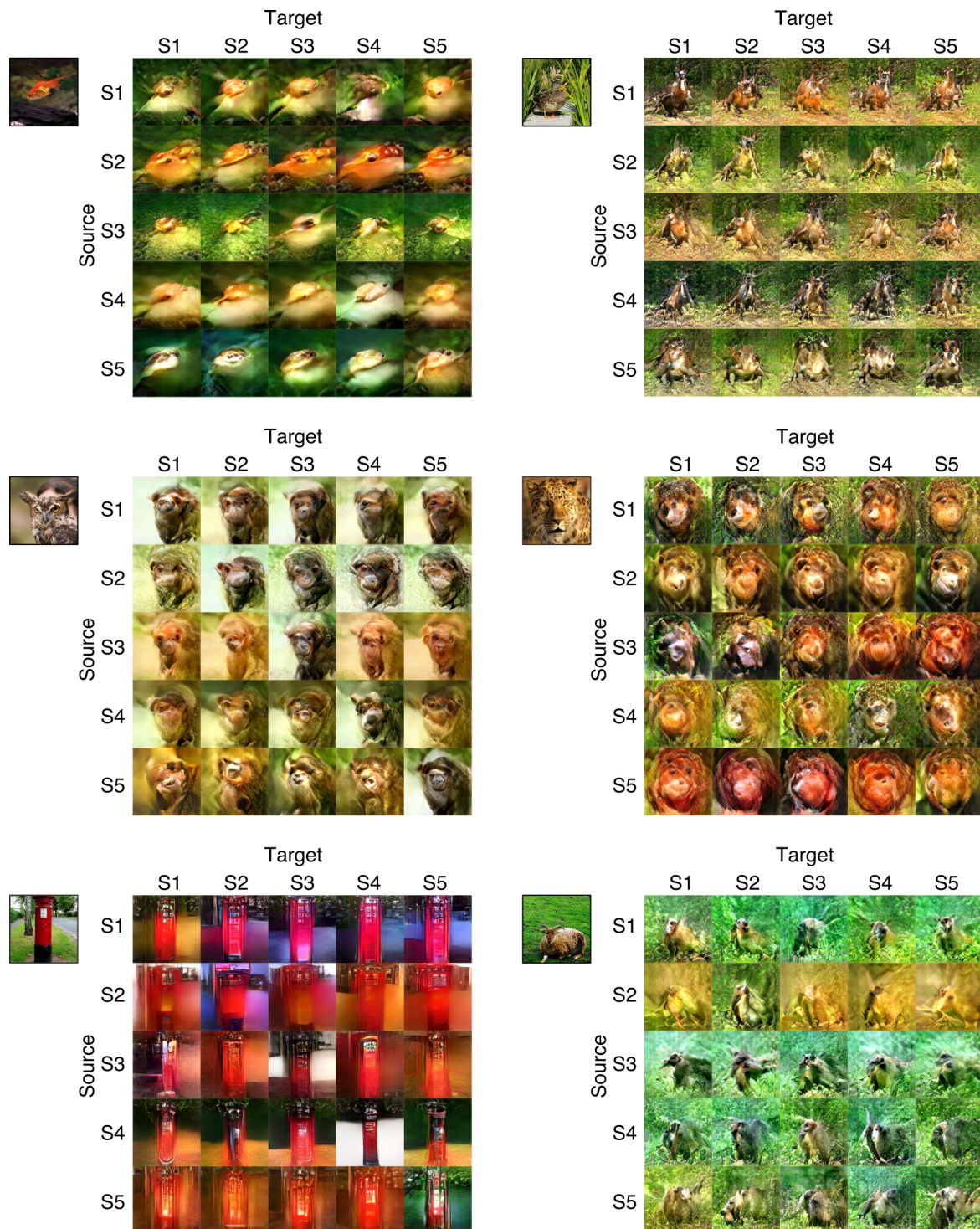


Fig. 8.3 Reconstructed images with the CLIP-ViT decoder for all pairs. The reconstructions for each test stimulus are shown for 20 pairs using the CLIP-ViT decoder. The diagonal images in each block are reconstructed in the Within condition. All reconstructions were generated using the whole visual cortex (VC). For copyright reasons, some stimulus images have been replaced with visually similar alternatives.

8.3.2 Varying the number of training data for conversion

We investigated the effect of training sample size for converters on image reconstruction quality. We varied the number of training samples for converters (trained with VGG19 features) at various levels: 300, 600, 900, 1,200, 2,400, 3,600, 4,800, and 6,000 training samples, with data collection time ranging from 40 minutes to approximately 13 hours. To evaluate the decoding from converted brain activity, we consistently used 6,000 samples from the target subject to train the VGG19 and CLIP-ViT feature decoders.

Figure 8.4 shows the reconstructed images with the VGG19 decoder, using converters trained with varying sample sizes. Though the visual quality of the reconstructions diminished with smaller training sample sizes, converters trained with as few as 300 or 600 samples (around one hour) still yielded perceptible images. The identification accuracy increased with the number of training samples, gradually approaching the accuracy observed in the Within condition (Figure 8.5). Similar results were obtained for the reconstruction using the CLIP-ViT decoder (Figures 8.6 and 8.7). These results indicate the feasibility of inter-individual image reconstruction with converters trained on limited data, achieving modest performance while decreasing the dependence on extensive fMRI data collection.

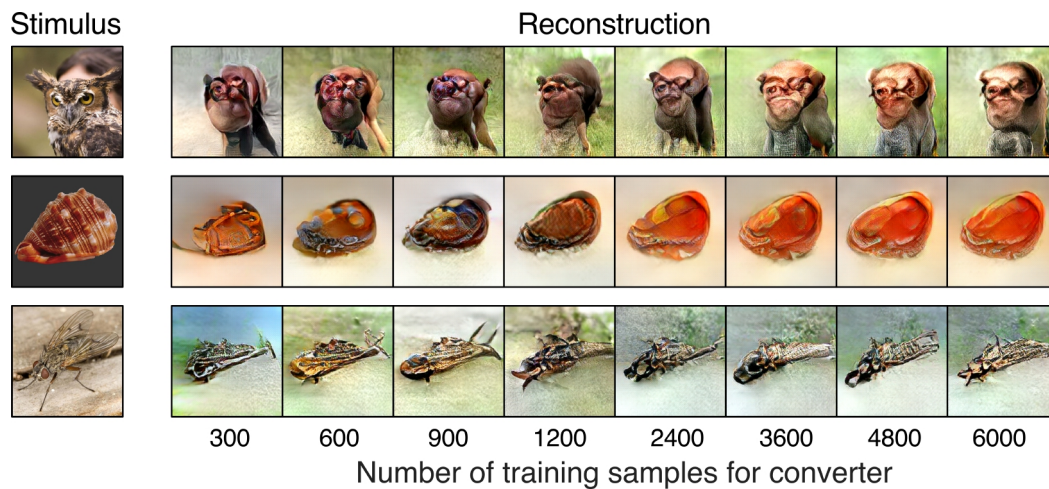


Fig. 8.4 Reconstructed images using the VGG19 decoder. Reconstructed images were generated with a varying number of training data from the same pair (VC; source: Subject 1, target: Subject 2). For copyright reasons, some stimulus images have been replaced with visually similar alternatives.

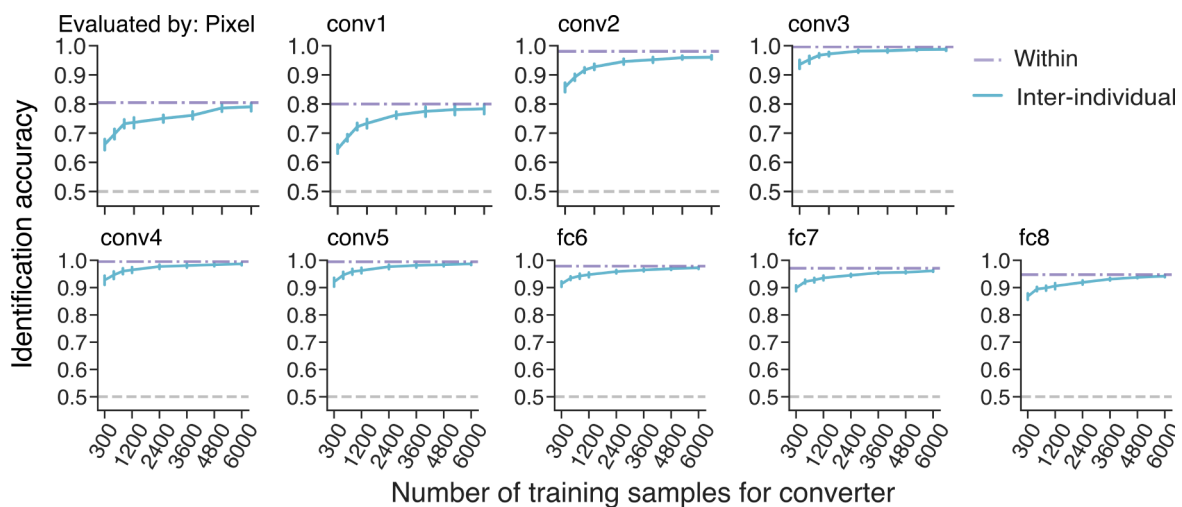


Fig. 8.5 Identification accuracy of reconstructed images using the VGG19 decoder. Pairwise identification was performed using the pixel values and the extracted DNN feature values from the reconstructions with a varying number of training data. The mean identification accuracy was calculated across all individual pairs (error bars, 95% C.I. from 20 pairs; dashed lines, 50% chance level). The results are shown with those from the Within condition (6,000 training samples).

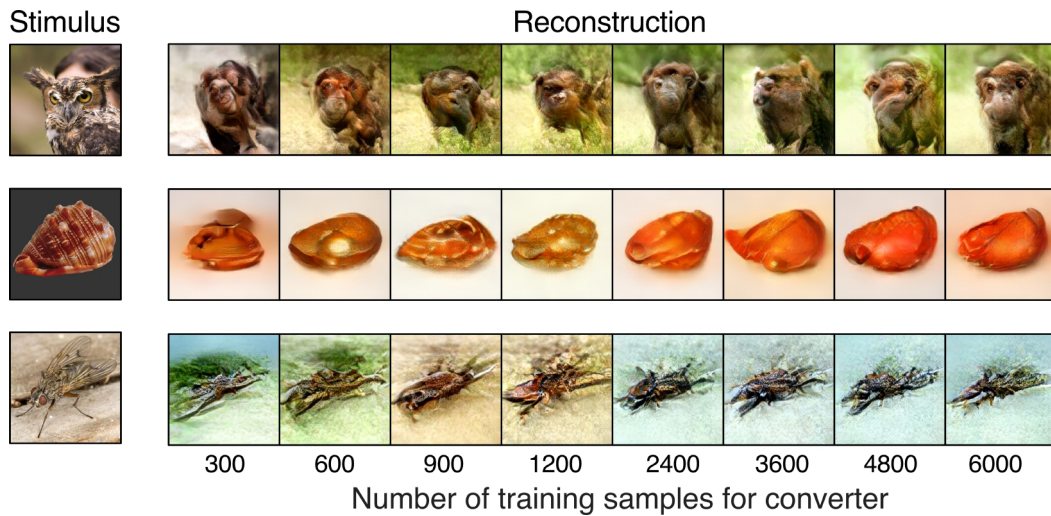


Fig. 8.6 The effect of training sample size for conversion on generalization. The source brain activity was converted into the target's space using VGG19-based converters and was used to reconstruct images using the target subject's CLIP-ViT decoder. Reconstructions were generated with varying training data from the same pair (VC; source: Subject 1, target: Subject 2). Some stimulus images have been replaced with visually similar alternatives for copyright reasons.

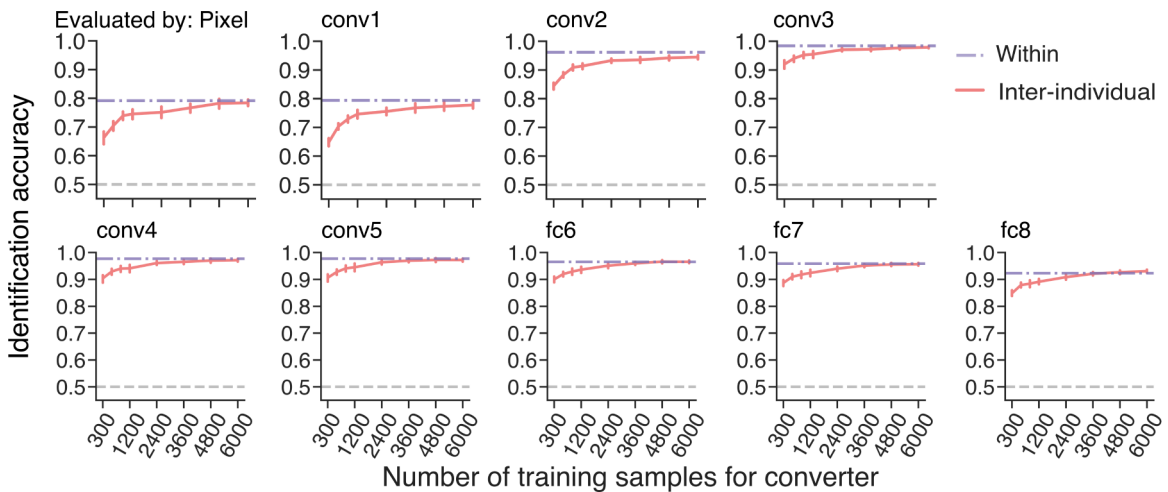


Fig. 8.7 The effect of training sample size for conversion on identification accuracy. The source brain activity was converted into the target's space using VGG19-based converters and was used to reconstruct images using the target subject's CLIP-ViT decoder. Pairwise identification was performed using the pixel values and the extracted DNN feature values (AlexNet) from the reconstructions with a varying number of training data. The mean identification accuracy was calculated across all individual pairs (VC; error bars, 95% C.I. from 20 pairs; dashed lines, 50% chance level). The results are shown with those from the Within condition.

8.4 Discussion

This chapter investigated the generalizability of inter-individual neural code conversion using content loss-based converters, focusing on whether the converted brain activity retains meaningful information that can be decoded by different models and with limited training data. Although the converters were trained using DNN features from the VGG19 model, we demonstrated that the converted brain activity could also be decoded by a CLIP-ViT decoder, which was trained independently of the converter, yielding reconstructions that were qualitatively and quantitatively similar to those obtained using the VGG19-based decoder. Even with limited training samples, both VGG19- and CLIP-ViT-based reconstructions achieved modest performance.

In our approach, brain activity patterns were indirectly aligned via the loss of latent representations of the visual content. This raises a concern that the converted brain activity could be specifically tailored to the particular form of content representation. However, we have found that although the converter was trained with VGG19 features (Simonyan & Zisserman, 2014), the converted brain activity could be decoded into CLIP-ViT features (Radford et al., 2021), which differ significantly from VGG19 features, resulting in faithfully reconstructed images (Figure 8.2). This indicates that our converter does not simply convert brain activity to a format specially tailored for a specific DNN representation. Rather, it learns a generalizable brain activity representation interpretable by other decoding schemes. Although our current study focuses on image reconstruction, the converted brain activity could be used with other decoding schemes and other types of information and tasks, to examine the versatility of the conversion based on DNN representations.

The content loss-based converter used a training approach similar to feature decoders, where it constituted additional layers to the target's decoder. However, our results suggest that it uniquely specializes in learning the statistical relationship between the source and target brain spaces in several key aspects. First, we have demonstrated that the converter effectively converted brain activity patterns across subjects, achieving conversion accuracy comparable to converters optimized by brain loss in Chapter 5, highlighting its capability to align brain activity patterns. Second, as discussed above, the converted brain activity could be interpreted by another decoding scheme (Figures 8.1 and 8.2), indicating that the converter learns a more generalizable mapping between brain spaces. Further, converter training required a substantially smaller amount of data for successful image reconstruction than feature decoder training (Figures 8.4-8.7), suggesting a separate underlying mapping structure distinct from feature decoding.

These results suggest that the converter captures generalizable representations beyond the specific readout. This generalizability enables more flexible and scalable decoding pipelines,

potentially allowing pre-trained decoders to be reused across individuals and tasks. However, our choice of VGG19 features as a latent representation of visual content may not fully capture the diverse aspects of visual experience and neural representations. While VGG19 primarily models the ventral visual system's 2D perception processes, it may inadequately represent motion, 3D structure, and global perception. Future studies may explore whether this approach can generalize beyond static image perception to other perceptions, such as video, 3D objects, or dynamic scenes, and whether representations derived from alternative models can be decoded from converted brain activity.

Chapter 9

Inter-site neural code conversion

9.1 Introduction

Neural code conversion aims to flexibly convert brain activity patterns across individuals while preserving the fine-grained information. Previous chapters have demonstrated that such conversions can preserve fine-grained visual representations, enabling accurate inter-individual decoding and image reconstruction even without shared stimuli. However, these analyses have been limited to data collected within a single site, under controlled experimental protocols. In practical applications, brain data are often collected from different sites using distinct scanning hardware, experimental designs, or stimulus sets, raising an important question regarding the feasibility of neural code conversion across sites.

Inter-site neural code conversion aims to enable functional alignment across brain data acquired under heterogeneous conditions. Unlike conventional functional alignment methods that require shared stimuli or anatomical correspondences, our approach uses hierarchical DNN features as latent representations of visual content to optimize the converter. This framework does not rely on share stimuli and thus allows brain activity from a source subject in one dataset to be converted into the representational space of a target subject in another dataset.

In this chapter, we evaluate the effectiveness of inter-site neural code conversion using three publicly available fMRI datasets: Deeprecon (Ho et al., 2023; Horikawa & Kamitani, 2022; Shen et al., 2019b), NSD (Allen et al., 2022), and THINGS (Hebart et al., 2023). We examine whether visual representations can be preserved through inter-site conversion by decoding DNN features and reconstructing perceived images across subject pairs from different datasets. We further compare decoding and reconstruction performance with standard within-individual baselines to assess the quality of inter-site conversions. Through

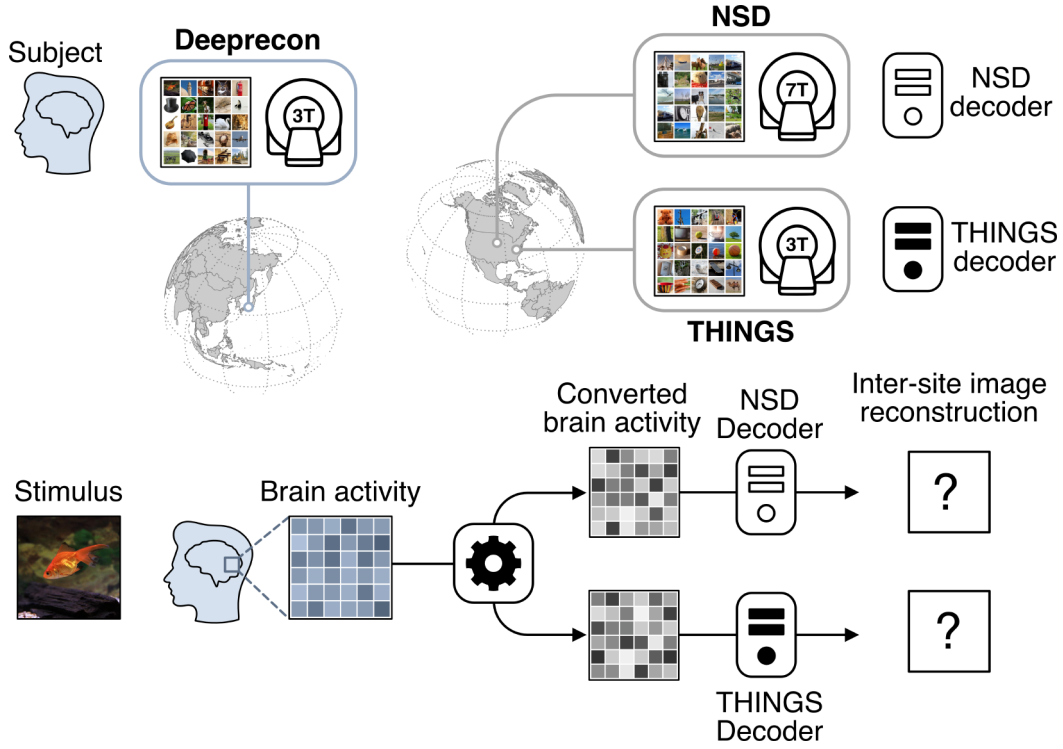


Fig. 9.1 Inter-site neural code conversion and image reconstruction. The trained converter functions across different datasets, which may have different stimuli and scanners with different resolutions. When presented with a novel stimulus, the subject's brain activity pattern from one dataset is converted into the space of another dataset. The converted pattern is then decoded into image feature representations by a feature decoder within the target dataset, enabling inter-site image reconstruction as perceived by the source subject.

this analysis, we explore the scalability of content-based neural code conversion across diverse experimental environments and data sources.

9.2 Methods

9.2.1 Inter-site neural code conversion

The inter-site neural code converter is designed to enable functional alignment across different measurement sites using a content loss-based optimization framework. For each subject pair across sites, we designate a source subject (from one dataset/site) and a target subject (from another). First, we train a decoder in the target dataset to predict latent visual features from the target subject's fMRI activity. Then, an inter-site neural code converter is trained to transform the source subject's brain activity into the target subject's brain space by minimizing the

content loss between the latent features decoded from the converted activity and those of the original stimulus.

The inter-site neural code converter is based on the content loss that was described in Chapter 5, and this method does not require paired brain activity between the source and target individuals. Therefore, the source and target can come from different sites or datasets, even when stimuli, scanning protocols, or resolutions vary. During testing, novel stimuli are presented to the source subject, and the measured brain activity is converted into the target space. The converted activity is decoded using the target site’s decoder to obtain the latent visual features, which are then passed to a pre-trained image generator to reconstruct the perceived image (Figure 9.1). The inter-site conversion was evaluated using three datasets: the Deeprecon (Ho et al., 2023; Horikawa & Kamitani, 2022; Shen et al., 2019b), NSD (Allen et al., 2022), and THINGS (Hebart et al., 2023). There were 94 subject pairs included in the inter-site analyses (see Chapter 2 for more details). As in Chapter 5, we used the hierarchical DNN features from the VGG19 model (Simonyan & Zisserman, 2014) as the latent representation of visual stimulus content.

9.2.2 Group statistics

We performed group-level statistical inference by showing the group means and confidence intervals for decoding accuracy and identification accuracy. Each data point represents the mean accuracy corresponding to an inter-site pair of subjects. As in previous chapters, we applied bootstrapping to these dyadic data points due to dyadic dependency to calculate the group mean and 95% confidence interval.

9.3 Results

9.3.1 Inter-site DNN feature decoding

We extended our analysis to investigate the feasibility of inter-site neural code conversion, where converters were trained between source and target subjects from distinct sites. For this analysis, we used the Deeprecon dataset and two additional datasets: the NSD and the THINGS dataset. The source subject was selected from one of these three datasets, while the target subject was selected from a different dataset, resulting in 94 individual pairs with no stimuli shared between any pairs. In this inter-site analysis, due to the absence of ground truth (measured fMRI activity) for evaluating conversion accuracy in the brain space, we evaluate the inter-site neural code conversion through the inter-site DNN feature decoding.

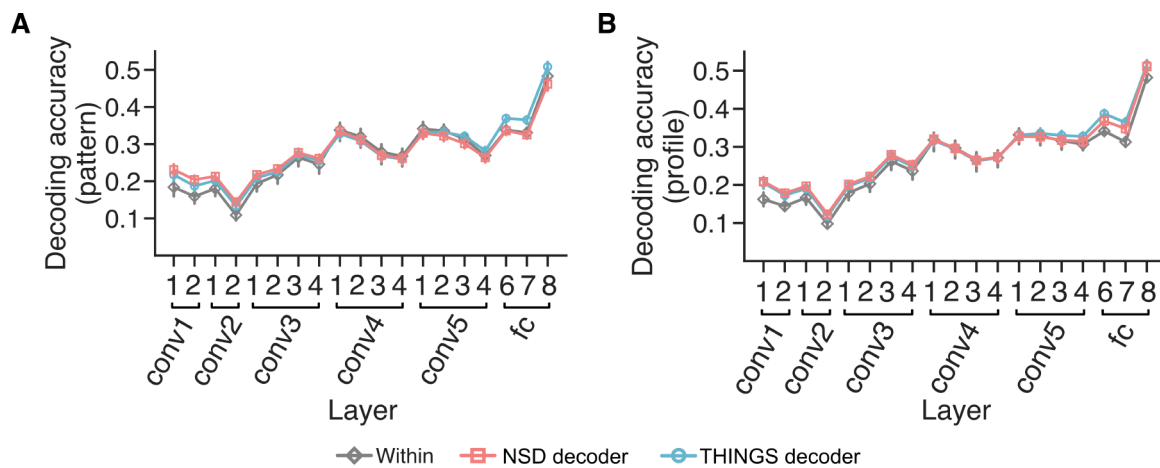


Fig. 9.2 Feature decoding accuracy for Deeprecon test data. (A) Pattern correlation. The mean pattern correlation for each layer of the VGG19 model is shown for the Within and Inter-site conditions (VC; error bars, 95% C.I. from five subjects for the Within condition, from 20 pairs for the NSD decoder condition and from 15 pairs for the THINGS decoder condition). (B) Profile correlation. The three conditions are compared as in (A) but with profile correlation coefficients.

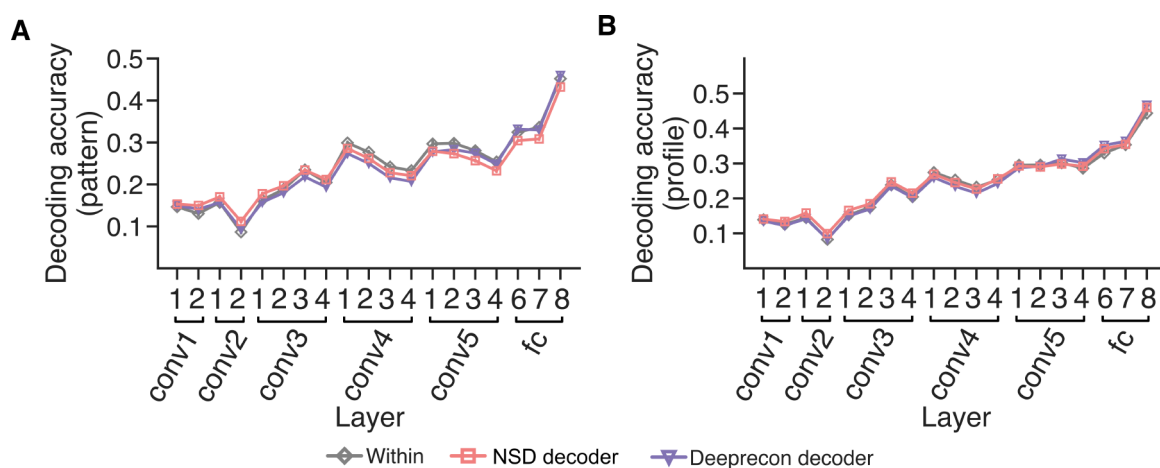


Fig. 9.3 Feature decoding accuracy for THINGS test data. (A) Pattern correlation. The mean pattern correlation for each layer of the VGG19 model is shown for the Within and Inter-site conditions (VC; error bars, 95% C.I. from three subjects for the Within condition, from 12 pairs for the NSD decoder condition, and from 15 pairs for the Deeprecon decoder condition). (B) Profile correlation. The three conditions are compared as in (A) but with profile correlation coefficients.

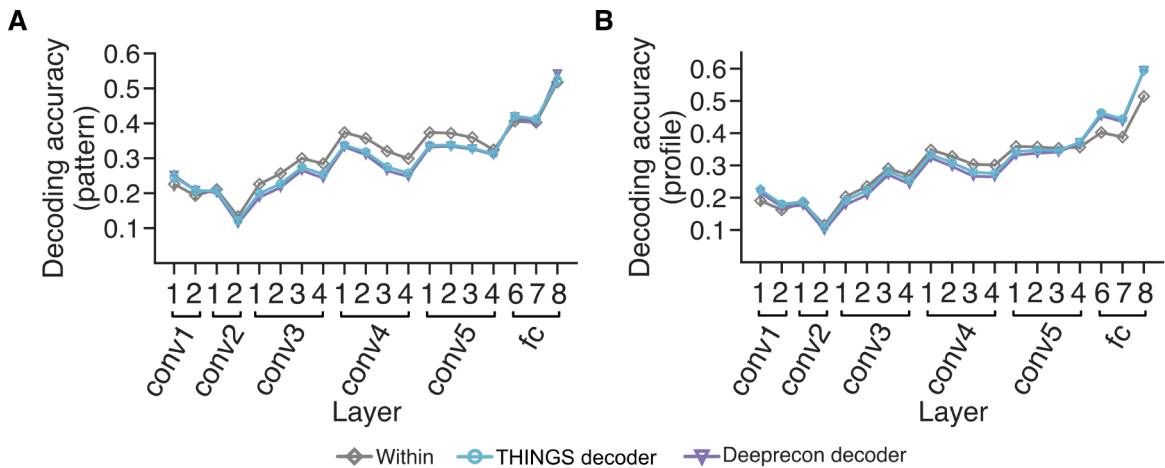


Fig. 9.4 Feature decoding accuracy for NSD test data. (A) Pattern correlation. The mean pattern correlation for each layer of the VGG19 model is shown for the Within and Inter-site conditions (VC; error bars, 95% C.I. from four subjects for the Within condition, from 12 pairs for the NSD decoder condition, and from 20 pairs for the Deeprecon decoder condition). (B) Profile correlation. The three conditions are compared as in (A) but with profile correlation coefficients.

Figure 9.2 presents the results of inter-site decoding accuracy, where the source subject originates from the Deeprecon dataset, and the target subjects are drawn from the THINGS and NSD datasets. The decoding performance is evaluated using both pattern correlation and profile correlation across different layers of the VGG19 model. The results show that inter-site decoding yields a consistent pattern of accuracy for both correlation metrics. For all three decoding conditions, Within, NSD decoder, and THINGS decoder, the decoding accuracy is relatively low in the early layers of the VGG19 model, where features represent low-level visual information such as edges and textures. As the layers progress toward higher levels, which capture more abstract and semantic features, decoding accuracy increases steadily. This layer-wise trend is consistent across all decoding conditions, indicating that the hierarchical DNN representations are similarly captured regardless of whether the decoder is trained and tested within the same individual or across sites. Furthermore, a comparison of decoding accuracy across the three methods reveals that their performance is comparable at each layer, with overlapping confidence intervals. The results closely parallel the results reported in Chapter 6, where inter-individual decoding within a single dataset showed similar decoding patterns and accuracy levels. Similar decoding tendencies are also observed for test data from both the THINGS and NSD datasets in the inter-site setting (Figures 9.3 and 9.4). These results suggest that inter-site neural code conversion preserves hierarchical DNN features.

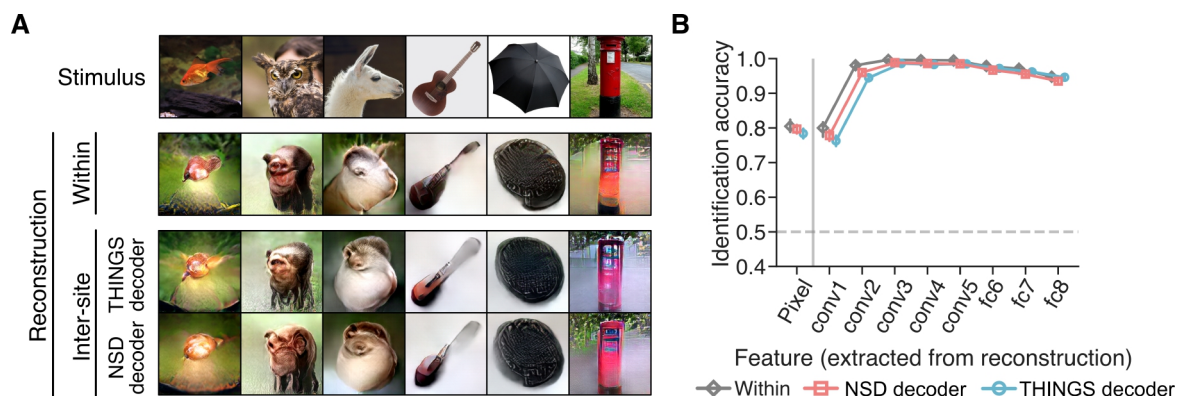


Fig. 9.5 Inter-site reconstruction for Deeprecon test data. (A) Reconstructions. Within-individual reconstructions were obtained using decoders from the same subject, while inter-site reconstructions were obtained using the target’s decoders from other datasets (VC; source: Subject1, target: Subject 6, 9; see Figure 9.6 for reconstructions of all pairs). For copyright reasons, some stimulus images have been replaced with visually similar alternatives. (B) Identification accuracy. Pairwise identification was performed using the pixel values and the extracted DNN feature values from the reconstructions. The DNN features were extracted from eight layers of the AlexNet model, including five convolutional (conv) layers and three fully connected (fc) layers. The mean identification accuracy was calculated across all subjects or individual pairs (error bars, 95% C.I. from five subjects or 20 pairs; dashed lines, 50% chance level).

9.3.2 Inter-site image reconstruction

To further evaluate the effectiveness of inter-site neural code conversion, we performed inter-site image reconstruction by reconstructing images using the DNN features decoded from the converted brain activities. Consistent with the procedure described in Chapter 6, we applied the same image reconstruction algorithm and identification analysis to assess the quality of the reconstructions.

Figure 9.5A presents examples of the inter-site image reconstruction for Deeprecon test data. (see Figure 9.6 for the reconstructions of 20 individual pairs). The inter-site reconstructions captured the core characteristics of the presented images, including their shape, color, and textures, showcasing visual content similar to that of the within-individual reconstructions. A quantitative evaluation revealed that the identification accuracy of inter-site reconstructions was slightly lower but still comparable to within-individual reconstructions (Figure 9.5B). Similar image reconstruction results are also observed for test data from both the THINGS and NSD datasets in the inter-site setting (Figures 9.7-9.10). These results demonstrated that brain activity patterns can be converted across subjects from different

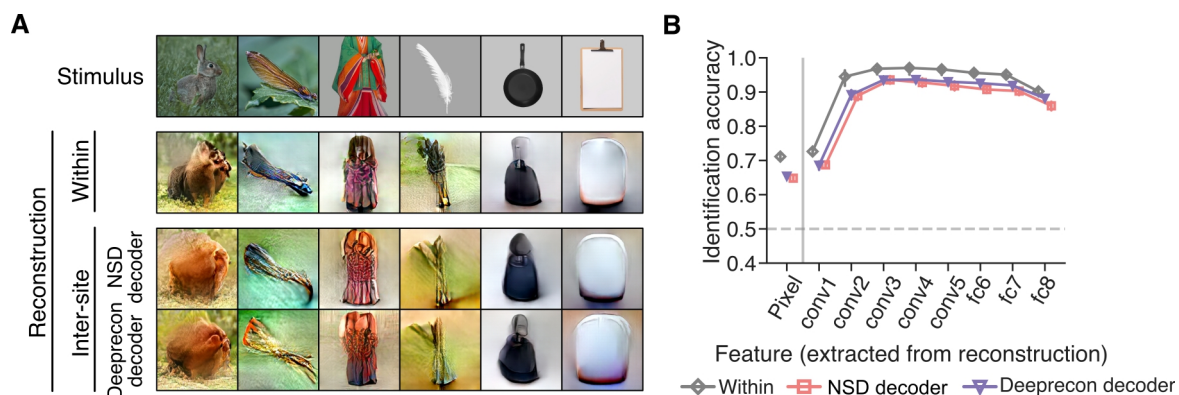


Fig. 9.7 Inter-site reconstruction for THINGS test data. (A) Reconstructions. Examples for source Subject 6 and target Subjects 1 and 9 are shown (VC; see Figure 9.8 for reconstructions of all pairs; some stimulus images were replaced for copyright reasons). (B) Identification accuracy. The mean identification accuracy was calculated across all subjects or individual pairs (error bars, 95% C.I. from three subjects, 12 pairs, or 15 pairs; dashed lines, 50% chance level).

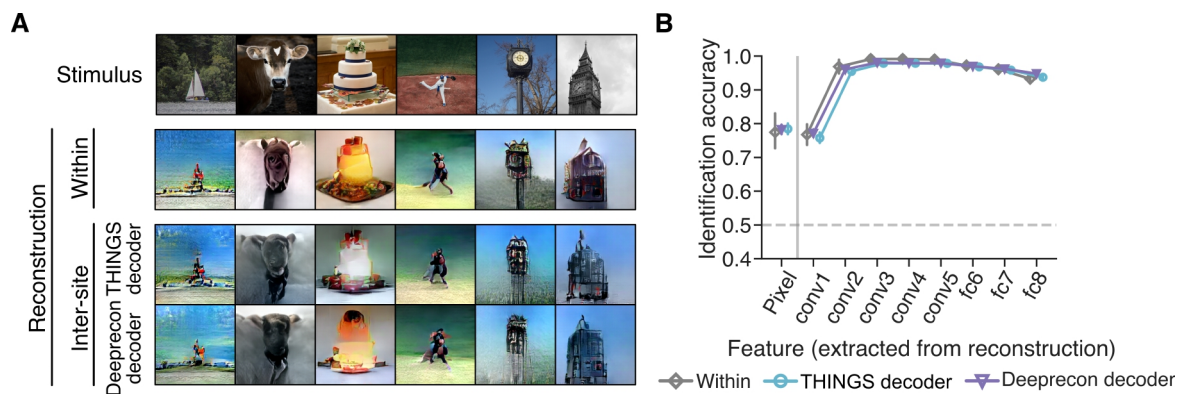


Fig. 9.8 Inter-site reconstructions for NSD test data. (A) Reconstructions. Examples for source Subject 9 and target Subjects 1 and 6 are shown (VC; see Figure 9.10 for reconstructions of all pairs; some stimulus images were replaced for copyright reasons). (B) Identification accuracy. The mean identification accuracy was calculated across all subjects or individual pairs (error bars, 95% C.I. from four subjects, 12 pairs, or 20 pairs; dashed lines, 50% chance level).

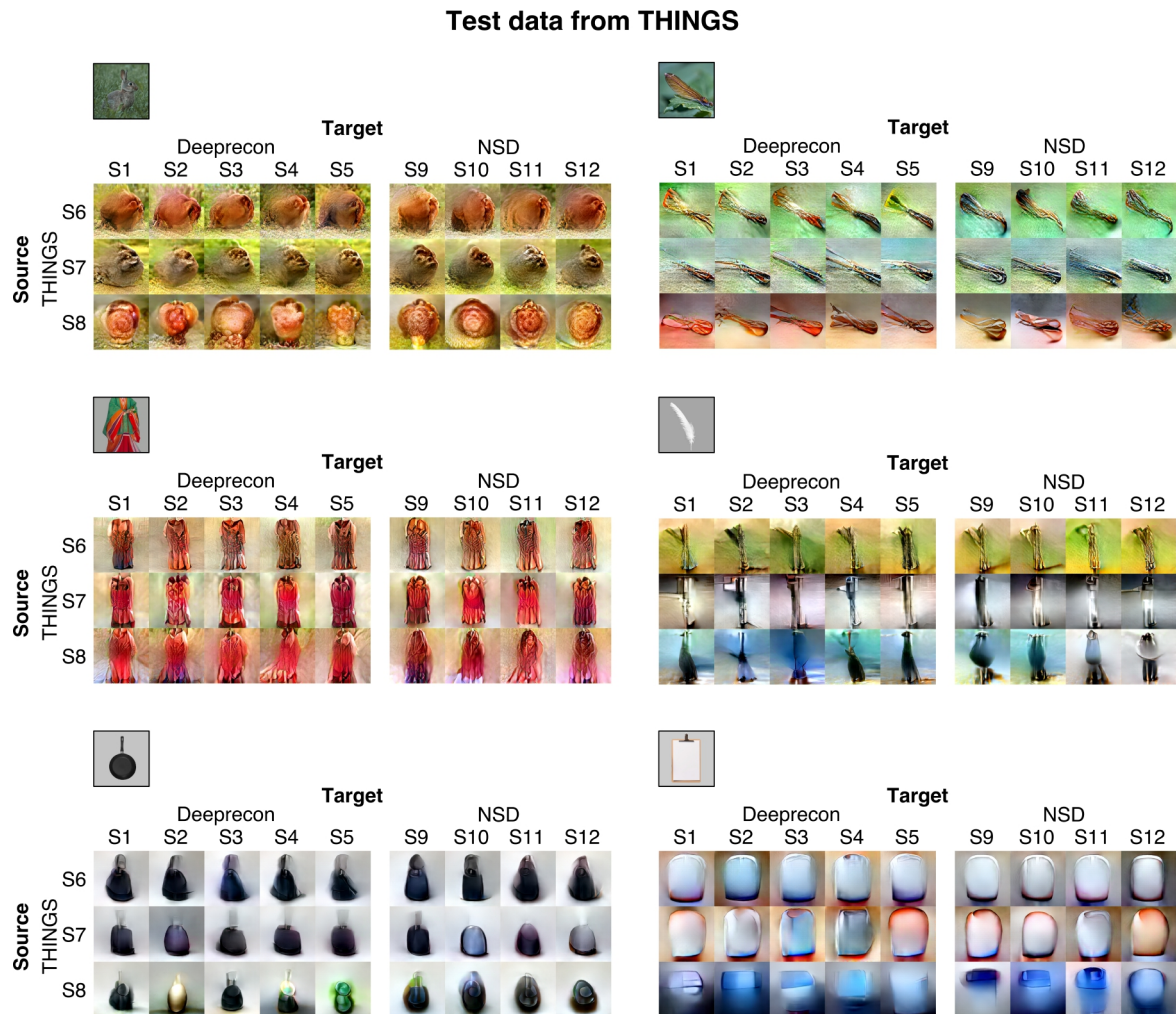


Fig. 9.9 Inter-site image reconstructions using THINGS test data for all pairs. All reconstructions were generated using the whole visual cortex (VC). For copyright reasons, some stimulus images have been replaced with visually similar alternatives..

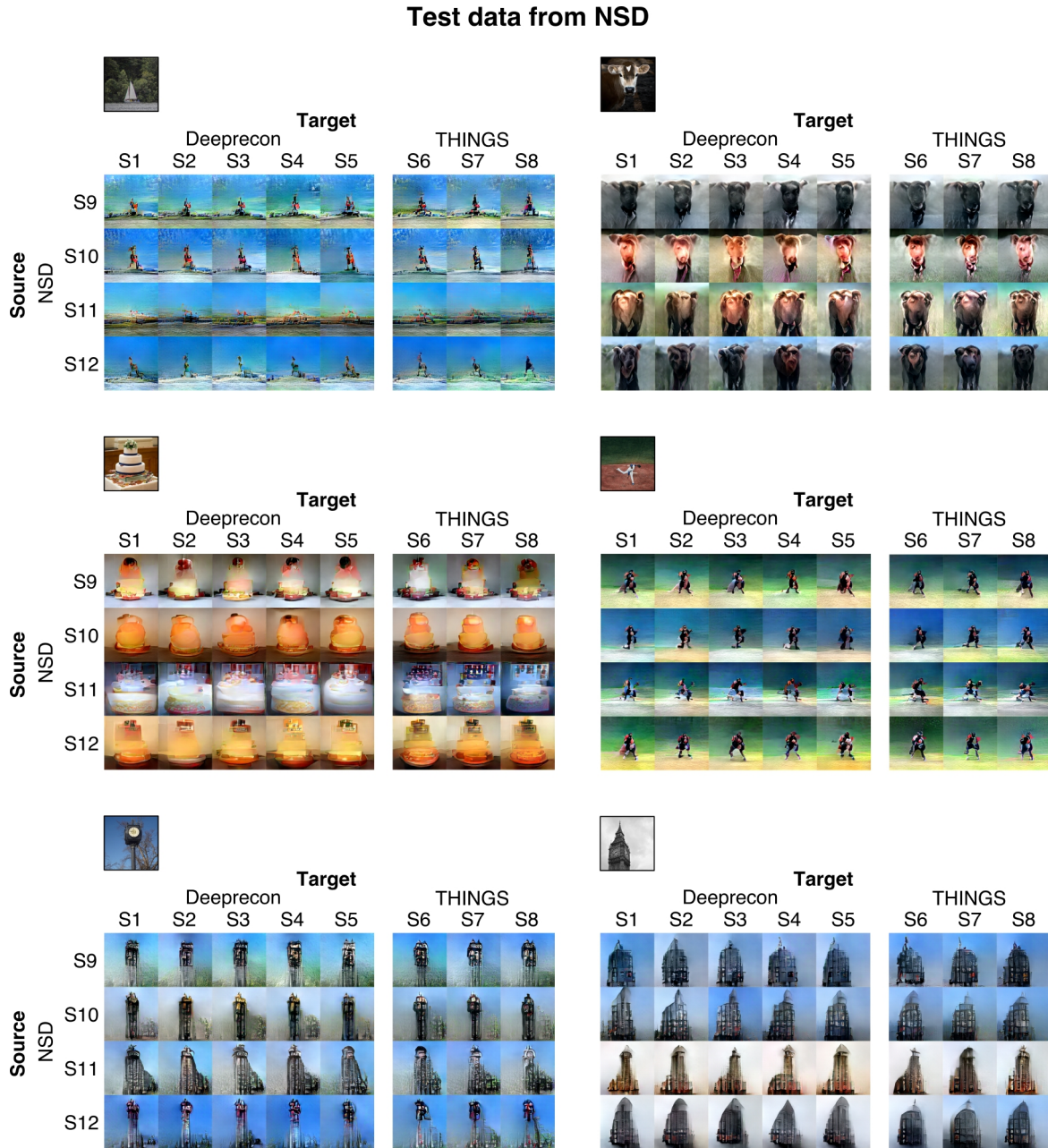


Fig. 9.10 Inter-site image reconstructions using NSD test data for all pairs. All reconstructions were generated using the whole visual cortex (VC). For copyright reasons, some stimulus images have been replaced with visually similar alternatives.

datasets, preserving the fine-grained feature representations that enable inter-site image reconstruction.

9.4 Discussion

This chapter investigated the feasibility of inter-site neural code conversion by evaluating both DNN feature decoding and image reconstruction across different fMRI datasets. By using data from the Deeprecon, THINGS, and NSD datasets, we demonstrated that it is possible to convert brain activity patterns from a source subject in one dataset to a target subject in a different dataset, without shared stimuli. The decoded DNN features from these inter-site conversions preserved sufficient visual information to reconstruct images that closely resembled the original stimuli. These findings provide evidence that hierarchical fine-grained visual representations in the brain can be generalized and decoded across individuals and recording environments.

We have demonstrated inter-site image reconstruction using the Deeprecon, THINGS, and NSD datasets. While all combinations of these datasets for conversion yielded reliable reconstructions, it should be noted that there are different characteristics among these datasets, particularly in the diversity of the stimuli. The Deeprecon dataset includes fMRI brain responses to both natural and artificial images and is carefully designed to prevent overlap in object categories between training and testing datasets. The NSD and THINGS datasets provide more extensive stimuli of natural scenes and object images. However, the diversity of NSD stimuli is limited in semantic and visual contents, with considerable overlap of these contents between training and test sets (Shirakawa et al., 2024). This overlap might introduce bias in inter-site reconstruction results, particularly with the NSD test set, because semantic and visual contents from the training data that overlaps with the test data could be incorporated during the training of the converter. The use of multi-subject data in decoding may increase the risk of spurious prediction of “novel” content due to the growing overlap of semantic and visual contents between the training and test sets (Bao et al., 2024; Gong et al., 2024; Quan et al., 2024; Scotti et al., 2024).

Our method provides a promising tool for scalable neural data analyses. Conventional within-site analyses (Bao et al., 2024; Bilenko & Gallant, 2016; Gong et al., 2024; Haxby et al., 2011; Ho et al., 2023; Quan et al., 2024; Scotti et al., 2024; Van Uden et al., 2018; Yamada et al., 2015) are often conducted under homogeneous conditions, such as the same scanner, identical experimental design, and subjects from a single site, which may restrict the generalizability of the results. In contrast, our method enabled the conversion of brain data across different datasets and demonstrated that inter-site converted brain data can

achieve image reconstructions with performance comparable to within-site reconstructions (Figures 9.5-9.10). This approach potentially expands the reuse of brain data from various publicly available datasets, and enables scalable data analysis and modeling that transcend institutional and geographical barriers.

In inter-site neural code conversion, the data across different sites may not only involve different subjects or recording environments but also varying modalities. Our method is adaptable for use with brain data from various modalities. While we primarily analyze brain activity patterns from fMRI signals, the content loss approach can be extended to other modalities, such as EEG or MEG, which offer high temporal resolution for capturing rapid neural dynamics. Furthermore, this method supports inter-modality neural code conversion, such as transforming high spatial resolution fMRI data into EEG or MEG representations, enabling robust neural data analysis and modeling across both sites and modalities.

Chapter 10

Neural code conversion with subjective experience

10.1 Introduction

Previous chapters demonstrated that neural code converters trained on basic image perception tasks perform well when tested on similar tasks, successfully preserving visual representations and enabling inter-individual decoding and visual image reconstruction. However, beyond basic image perception, various experimental paradigms in the visual domain have revealed how subjective factors influence neural representations. Visual illusions reflect the dissociation between sensory input and perception, with differential cortical activity correlating with the subjective experience of illusory attribute (Cheng et al., 2023; Cornelissen et al., 2006; Gerardin et al., 2018; Ho & Schwarzkopf, 2022; Hong & Tong, 2017; Knebel & Murray, 2012; Kok & de Lange, 2014; Kok et al., 2016; Sasaki & Watanabe, 2004; Seghier & Vuilleumier, 2006). Attention-related tasks demonstrate that top-down cognitive processes can modulate visual cortical responses, enhancing the representation of relevant features (Çukur et al., 2013; Horikawa & Kamitani, 2022; Kamitani & Tong, 2006; Klein et al., 2014; Niazi et al., 2014; Reddy et al., 2009; Serences et al., 2004; Störmer et al., 2009; Yantis & Serences, 2003). In addition, visual imagery produces perception-like experiences without external input, activating visual areas similar to those engaged during veridical perception (Dijkstra et al., 2019; Horikawa & Kamitani, 2017; Kosslyn et al., 2001). These tasks involve distinct subjective experiences, raising the question of whether a converter trained on the basic image perception task can generalize to different visual tasks that engage internally driven processes and subjective contents.

In this chapter, we aim to investigate the feasibility of converting neural representations of subjective experience across individuals, including perceptual illusions, selective attention, and visual imagery. As in previous chapters, we trained two types of neural code converters using content loss-based optimization and brain loss-based optimization, based on brain activity evoked by perceived natural image stimuli. In the test stage, the source subject's brain activity during various forms of subjective experience was converted into the brain space of the target subject. The converted brain activity can be decoded into latent representations, which could then be converted into an image, reflecting the extent to which subjective content is preserved through inter-individual neural code conversion (Figure 10.1). We first examined whether illusory perceptions, including line and color illusions, can be transferred across individuals by comparing inter-individual reconstructions with within-individual and stimulus-based reconstructions. Next, we evaluated whether brain activity during selective attention can be aligned across individuals by reconstructing the attended images. Finally, we tested whether brain activity during visual imagery can also be converted across individuals by assessing the quality of the reconstructed imagined images.

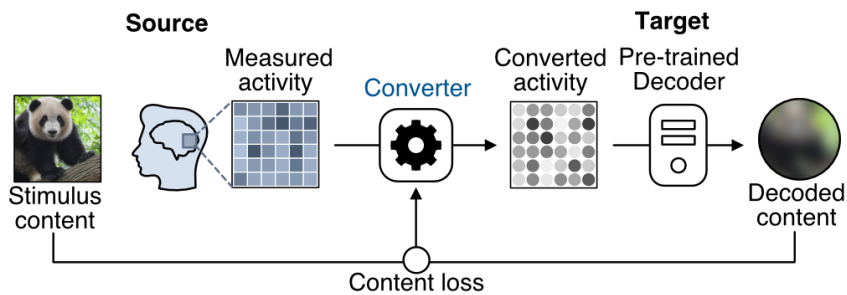
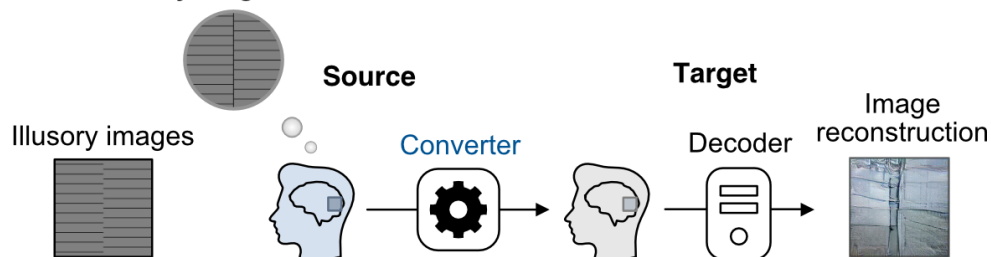
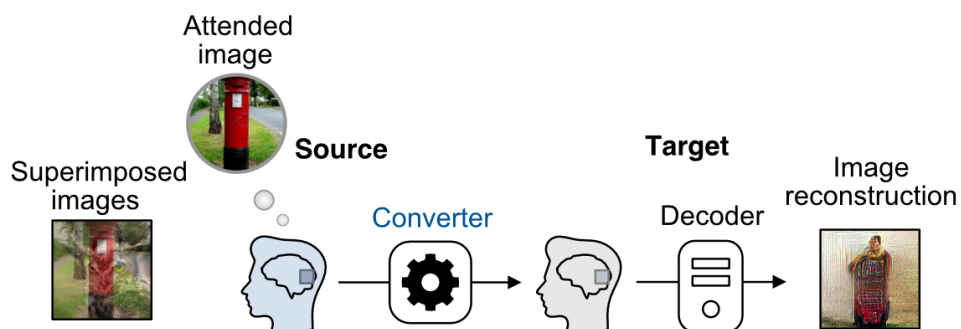
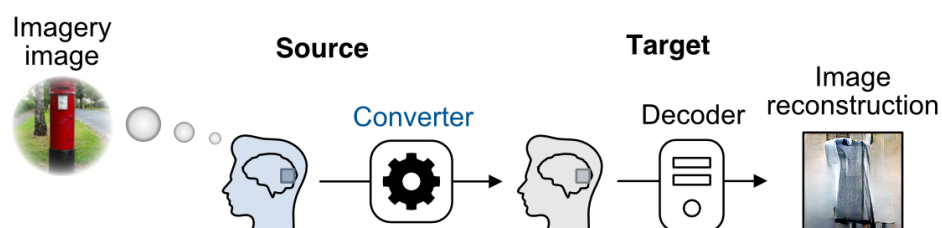
A Training: Natural images**B Test: Illusory images****C Test: Attended images****D Test: Imagery images**

Fig. 10.1 Neural code conversion with subjective experience. (A) Training the neural code converter using brain response to natural images. The training approach is the same as in Chapter 5. (B) Neural code conversion for illusion. When presented with an illusion stimulus, the source subject's brain activity pattern is converted into the space of target subject. The converted pattern is then decoded into image feature representations by a feature decoder within the target subject, enabling inter-individual image reconstruction as perceived by the source subject. (C) Neural code conversion during attention. When presented with a superimposed stimulus, the source subject was asked to attend to one the images. The brain activity during attention was converted across subjects and used for image reconstruction. (D) Neural code conversion during visual imagery. There is no veridical perception image, the source subject was asked to imagine the target images. The brain activity during imagery was converted across subjects, and used for image reconstruction.

10.2 Methods

10.2.1 DNN feature decoding

Similar to Chapter 6, the feature decoding analysis followed the methods described in Horikawa and Kamitani (2017) and Shen et al. (2019b). For the illusion task, the converted brain activity patterns were decoded into rectified fc6 features of AlexNet model (Krizhevsky et al., 2012). For the attention and imagery tasks, the converted brain activity patterns were decoded into hierarchical features of the VGG19 DNN model (Simonyan & Zisserman, 2014), ranging from conv1_1 to fc8. More details of the DNN features can be found in Subsection 6.2.1.

10.2.2 Visual image reconstruction

GAN

For reconstructing illusory images, we followed Cheng et al. (2023) and used a generator network originally developed by Dosovitskiy and Brox (2016), pre-trained using a Generative Adversarial Network (GAN) framework (available at: <https://lmb.informatik.uni-freiburg.de/resources/binaries>). This generator transforms the rectified outputs of fc6 layer features from the CaffeNet model into natural images using ImageNet training data (Russakovsky et al., 2015). In our implementation, for illusion reconstruction, decoded fc6 features from brain activity were fed into the generator to produce reconstructed images (VanRullen & Reddy, 2019).

Pixel optimization (iCNN)

For reconstructing attended and imagined images, we used a pixel optimization method (Horikawa & Kamitani, 2022; Shen et al., 2019b) that is the same reconstruction method in Chapter 6, which optimizes pixel values of an input image so that its deep neural network (DNN) features match those decoded from brain activity. We adopted the VGG19 model (Simonyan & Zisserman, 2014) implemented in Caffe as the DNN feature extractor, and DNN features before the rectification operation from eight layers (conv1-5 and all fully connected layers) were used to guide the optimization. In our implementation, for both attended images and imagined images, decoded VGG19 features from brain activity were fed into the iCNN reconstruction framework to produce reconstructed images.

10.2.3 Evaluation of reconstruction

Illusory line

We evaluated reconstructed images of the line illusion, both globally (whole image) and locally (specific regions). For the local analysis, we defined two types of circular regions: illusory regions (where illusory lines were expected) and non-illusory regions (where only inducer lines appeared), based on the stimulus geometry. Four circular regions were extracted from each stimulus image, with two located in illusory areas and two in non-illusory areas. These regions were defined by selecting the largest disks tangent to the central lines of the image.

We detected the principal orientation in each image or region using the Radon transform, following a method from texture analysis (Jafari-Khouzani & Soltanian-Zadeh, 2005). Each image was converted to grayscale, and the largest disk-shaped area within the region was analyzed. It was projected to a line space by summing the pixel intensities along each line within A :

$$R(r, \theta) = \sum_{(x,y) \in A} I(x,y) \delta(r - x \cos \theta - y \sin \theta) \quad (10.1)$$

where each line is parametrized by the distance from the center r and the orientation θ . The intensity of the pixel located at (x,y) is denoted by $I(x,y)$. If (x,y) is on the line, $\delta = 1$; otherwise, $\delta = 0$.

For each orientation, pixel intensities along parallel lines were summed, and the variance of these projections was calculated. The orientation with the highest variance was defined as the principal orientation. To focus on the illusory region, only lines passing within five pixels of the disk center were included.

Illusory color

We performed a regression analysis to evaluate the color filling-in effect, where a reconstructed image was approximated by the superposition of the illusory surface and the inducing stimulus with an additive error term.

Prior to the regression analysis, we created redness maps from the reconstructions, the illusory/real surfaces, and the inducing stimulus. A redness map was generated by converting the RGB image to the HSV color space and extracting the saturation (S) values of the red pixels. Non-red pixels were assigned zero. Red pixels were defined as those with hue (H) values in the range 0° – 10° or 160° – 180° .

Given the real color surface in the positive control image, we performed another regression analysis that excluded the inducing stimulus regressor to compare the illusory color with the real color. The following regression models were used for each reconstruction:

Illusion vs. control:

$$y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i \quad (10.2)$$

Illusion vs. positive control:

$$y_i = \beta_1 x_{i,1} + \varepsilon_i \quad (10.3)$$

Here, y_i is the redness value at the i -th pixel of the reconstructed image (size $h \times w$), ε_i is the residual error, and $x_{i,1}$, $x_{i,2}$ are the redness values from the illusory (or real) surface and the inducing stimulus, respectively. The coefficients β_1 and β_2 reflect the relative contributions of each component and were estimated by minimizing the sum of squared errors across all pixels in each reconstruction.

Attended images

We performed identification analysis separately for reconstructions from attention trials and single-image trials. For attended image reconstruction, we calculated correlation coefficients between the pixels or features extracted from the reconstruction and those of the two candidate images (attended vs. unattended). The image with the higher correlation was selected as the predicted attended image. Identification accuracy was defined as the proportion of correctly identified attended images across all reconstructions. For reconstructions from single-image viewing, each reconstructed image was compared with the presented image and the false alternative of another image. The false alternative is drawn from the remaining test images. This pairwise identification was repeated nine times for each reconstruction, using a different false image each time. The identification accuracy for a reconstructed image is defined as the ratio of correct identifications. For each subject (Within) or each individual pair, we calculated the mean identification accuracy by averaging the identification accuracies of all reconstructed images.

Imagined images

We performed identification analysis separately for imagined artificial shapes and natural images. For each reconstructed image, an identification was considered correct when the

correlation coefficient of the reconstructed image's feature vector was greater with the true feature vector than with the false alternative. The false alternative was selected from the remaining test images of the same type. For each reconstructed artificial shape, identification was repeated against all 14 candidate artificial images. Similarly, for each reconstructed natural image, identification was repeated against all nine candidate natural images. The identification accuracy for a reconstructed image is defined as the ratio of correct identifications. For each subject (Within) or each individual pair, we calculated the mean identification accuracy by averaging the identification accuracies of all reconstructed images.

10.2.4 Group statistics

Group-level analyses were performed by presenting group means with individual data distributions in the figures. In all quantitative analyses of image reconstruction, each data point represented the mean accuracy for a subject or individual pair. These included the proportion of illusory orientation detection for line illusions, regression coefficients for color illusions, and identification accuracy for attended and imagined images. Group means were calculated by averaging these mean accuracies across all subjects or all subject pairs.

10.3 Results

10.3.1 Neural code conversion for line illusion

We first investigated whether brain activity patterns evoked by line illusion could be aligned across individuals while preserving subjective perceptual content. We used representative illusory contours induced by offset gratings (Soriano et al., 1996; see Chapter 2 for more details). The stimuli consisted of six configurations, combining two inducer orientations (0° and 90°) with four illusory line orientations (0° , 45° , 90° , and 135°). Brain activity from the whole visual cortex (VC) in response to these images was converted from the source subject to the target subject using two types of neural code converters. The converted activity was then decoded into DNN feature representations and subsequently used to reconstruct images. As a baseline, we also performed within-subject reconstruction, in which DNN features were predicted using decoders trained on the same subject's data, referred to as the "Within" condition. To assess whether brain activity is essential for reconstructing illusory features, we also reconstructed the stimulus images directly from their stimulus-derived DNN features.

Figure 10.2 shows example reconstructions from the visual cortex (VC). Reconstructions based on stimulus-derived DNN features did not exhibit illusory lines, suggesting that the stimulus alone was insufficient to evoke the perceptual illusion. In contrast, reconstructions

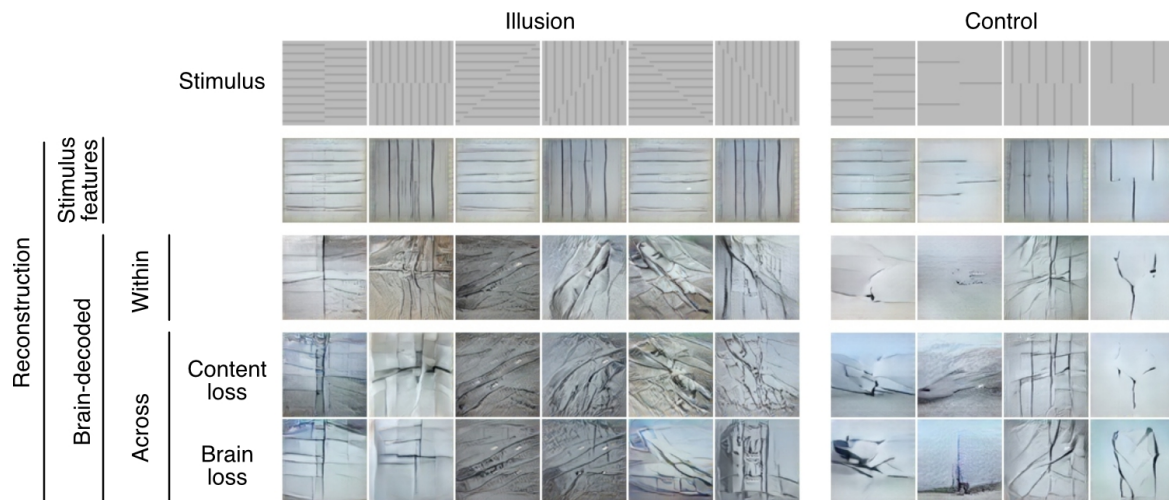


Fig. 10.2 Reconstructed illusory line images. The reconstructions are shown for the Stimulus features, Within, Brain loss, and Content loss conditions. Each stimulus image was all from the same source subject (VC; source: Subject 2, target: Subject 1)

based on DNN features decoded from brain activity, obtained either through the within-individual or across-individual approach, contained line components of the illusory and the inducer orientations. The illusory orientation appeared more prominent than the inducer orientation. In the control images, where fewer grating lines were presented to weaken the illusory percept, the line components corresponding to the illusory orientation were less prominent. When comparing the within-individual and across-individual conditions, both types of neural code conversion produced reconstructions that closely resembled those from the within-individual condition, successfully capturing the key features of the illusory orientation. The prominence and spatial distribution of the illusory lines were generally comparable across the two conditions.

To quantitatively evaluate the reconstructed illusory lines, we used the Radon transform (Jafari-Khouzani & Soltanian-Zadeh, 2005) to detect the most apparent line orientation in each reconstruction. We calculated Radon projections for line areas traversing the center of the image at each orientation and defined the orientation with the largest variance as the principal orientation (Figure 10.3A and 10.3B; see Subsection 10.2.3). While the illusory line is typically perceived at the abutting region of the inducer gratings, this global analysis does not capture the localized nature of the percept. Therefore, we also conducted a local analysis by detecting the principal orientation separately in (1) the illusory region, where the illusory line is expected, and (2) non-illusory regions, where only the inducer orientation is present (Figure 10.3C). We then calculated the proportion of reconstructions in which the detected orientation was closer to the illusory orientation in each region, allowing comparison between

local and global representations. Figure 10.3D presented closer-to-illusory proportions for the global, local illusory, and local non-illusory regions in Within-individual condition and two across-individual conditions. The proportions of reconstructions in which the detected orientation was closer to the illusory orientation were higher in the global and local illusory regions compared to the non-illusory regions. This pattern indicates that reconstructions captured the expected illusory percept more reliably in the illusory regions. This trend was consistently observed across all three conditions: the within-individual condition, and both types of across-individual conditions (content loss and brain loss). When comparing within- and across-individual conditions, the closer-to-illusory proportions in both the global and illusory regions were slightly higher in the within-individual condition. However, the across-individual conditions, particularly the content loss-based converter, also yielded comparable levels of illusory orientation detection. These results suggest that the neural code converters can preserve localized representations of subjective illusory content across individuals, although with slightly reduced strength compared to the within-individual case.

10.3.2 Neural code conversion for color illusion

We next investigated whether brain activity evoked by color illusions could be aligned across individuals. Neon color spreading is a phenomenon where color appears to extend beyond the physical stimulus, forming a transparent surface (Bressan et al., 1997; van Tuijl, 1975). We used two types of neon color spreading: the Ehrenstein configuration, where color in line regions appears to fill a circle (Ehrenstein, 1941), and the Varin configuration, where color in wedge regions appears to form a rectangle (Varin, 1971). For both illusions, we prepared control images that weakened the illusory percepts, and positive control images that physically mimicked the percept with real lines or a uniform color surface. As in the line illusion analysis, we applied two types of neural code converters to transform brain activity across individuals, followed by DNN decoding and image reconstruction. We also included within-individual and stimulus-based reconstructions for comparison.

Figure 10.4 shows examples of the color illusion reconstructions from the visual cortex (VC). Reconstructions based on stimulus-derived DNN features did not show perceptual color spreading, indicating that the stimulus alone was insufficient to produce the illusion. In the Ehrenstein configuration of neon color spreading, reconstructions based on brain activity showed more extensive color spreading compared to those from the control condition, where a line-width gap was introduced to weaken the illusion. In the Varin configuration, the control images were designed to suppress the color spreading while preserving the contour structure. Both the illusory and control reconstructions exhibited similar contour-like intensity profiles, but color spreading was more prominent in the illusory condition. This effect, which

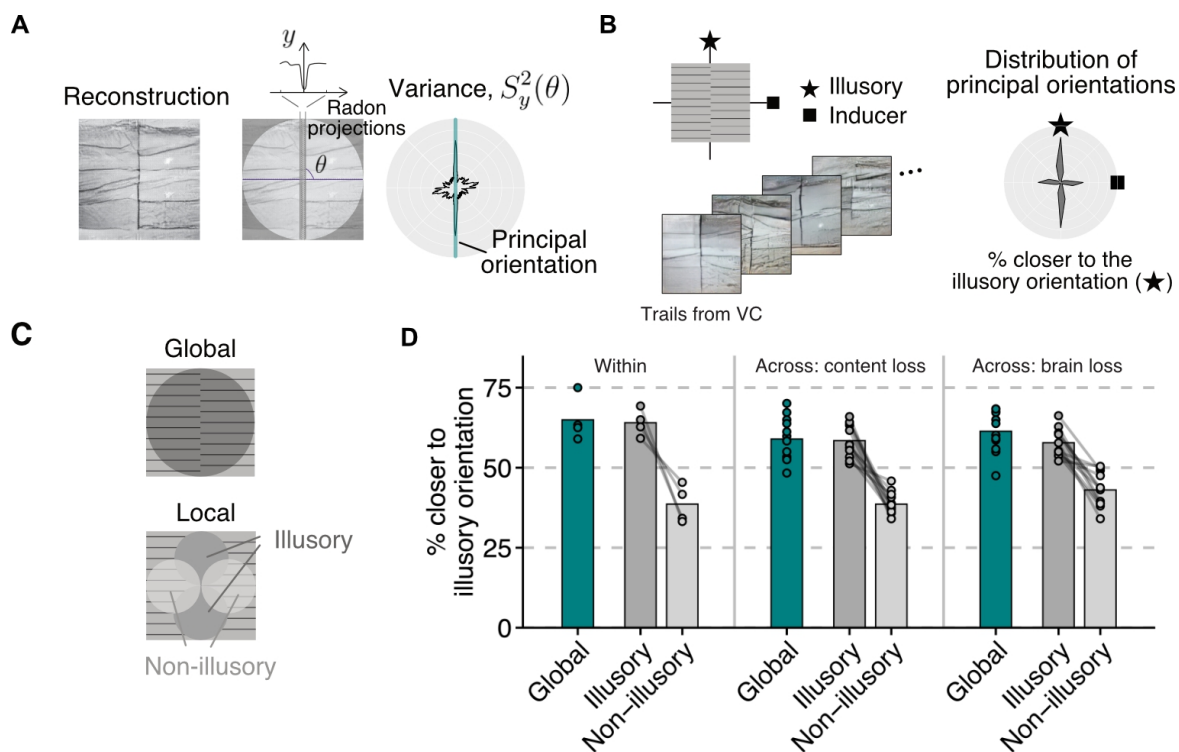


Fig. 10.3 Quantitative evaluation of reconstructed illusory line images. (A) Principal orientation detection. The orientation with the largest variance in Radon projections across line positions was identified as the principal orientation in an image. (B) Global and local presence of illusory orientation. (C) Distribution of principal orientations in reconstructions from VC. (D) Proportion of principal orientations biased toward the illusory orientation. For each reconstructed image, we determined whether the principal orientation was closer to the illusory orientation than to the inducer orientation. The mean proportion was first computed across all reconstructed images for each subject (within-individual) or subject pair (across-individual), and then averaged across subjects or pairs ($n = 4$ subjects; 12 individual pairs) to obtain the group-level mean. Bar graphs represent the mean proportions, and each dot indicates an individual subject or subject pair.

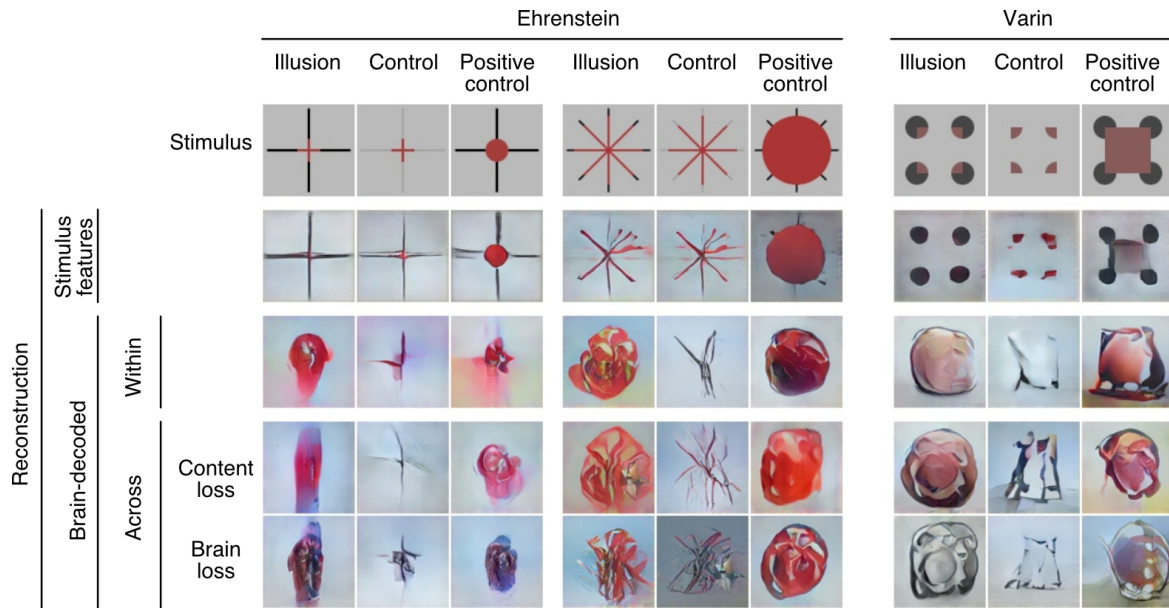


Fig. 10.4 Reconstructed neon color spreading images. The reconstructions are shown for the Stimulus features, Within, Brain loss, and Content loss conditions. Each stimulus image was all from the same source subject (VC; source: Subject 1, target: Subject 2).

resembled the positive control condition where physical color surfaces were present, was observed in both within- and across-individual conditions. Although the difference between the illusory and control conditions was more evident in the within-individual reconstructions, the across-individual condition using the content loss converter also captured key aspects of the illusory color surface.

To quantitatively evaluate color illusion reconstruction, we performed regression analysis on the redness (saturation) maps of each reconstructed image (Figure 10.5A and 10.5B; see Subsection 10.2.3). Redness maps were computed from the RGB values of the reconstructions, stimulus images, and predefined illusory surface regions. The redness profile of each reconstruction was regressed onto two predictors: the expected illusory surface map and the stimulus map, with corresponding regression coefficients. The illusory surface regressor was shared between the illusion and control conditions for comparison, and its coefficient was used as the measure of illusory color reconstruction.

Regression was performed for all individual reconstructions and pooled across configurations (e.g., line size and number in Ehrenstein) for each subject or subject pair. In both the Ehrenstein and Varin configurations, the illusory surface coefficient (β_1) was higher in the illusion condition compared to the control condition, indicating that reconstructed images reflected the spreading of color in the absence of physical color surfaces (Figure 10.5C and 10.5D). This difference was evident in the within-individual condition and similarly

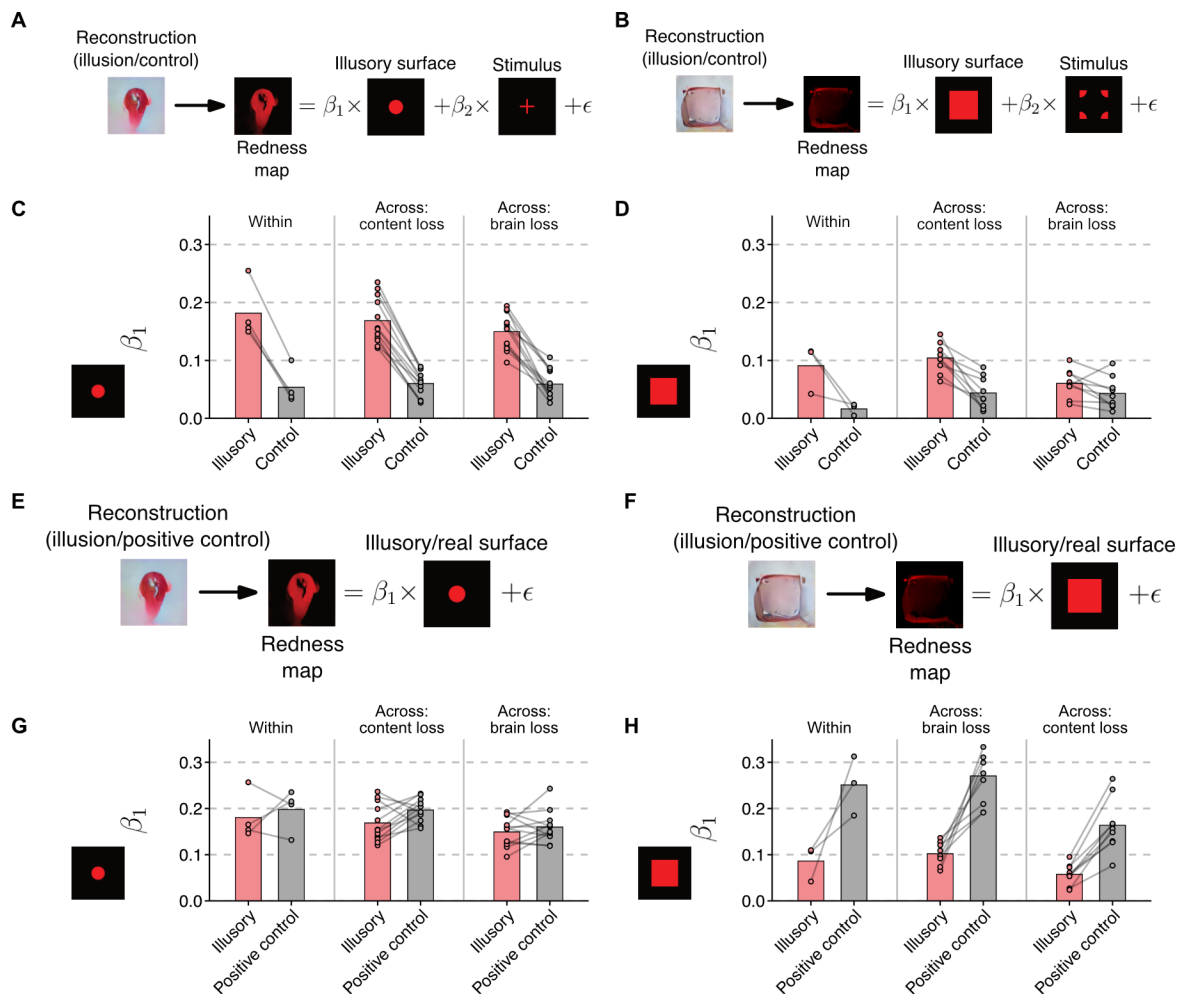


Fig. 10.5 Quantitative evaluation of reconstructed neon color spreading images. (A) Illustration of regression analysis for Ehrenstein configuration with the illusion and the control conditions. The redness map of a reconstructed image was fitted by those of the illusory surface (expected region of color filling-in) and the stimulus. (B) Illustration of regression analysis for Varin configuration with the illusion and the control conditions. (C) The illusory surface coefficient values for Ehrenstein configuration with the illusion and the control conditions. For each reconstructed image, we calculated the illusory surface coefficient value. The mean value was first computed across all reconstructed images for each subject (within-individual) or individual pair (across-individual), and then averaged across subjects or pairs ($n = 4$ subjects; 12 individual pairs) to obtain the group-level mean. Bar graphs represent the mean proportions, and each dot indicates an individual subject or subject pair. (D) The illusory surface coefficient values for Varin configuration with the illusion and the control conditions. (E) Illustration of regression analysis for Ehrenstein configuration with the illusion and the positive control conditions. The redness map of a reconstructed image was fitted by that of the illusory or real surface. (F) Illustration of regression analysis for Varin configuration with the illusion and the positive control conditions. (G) The illusory surface coefficient values for Ehrenstein configuration with the illusion and the positive control conditions. (H) The illusory surface coefficient values for Varin configuration with the illusion and the positive control conditions.

observed in both across-individual conditions, with stronger effects in the content loss-based conversion.

To further examine the similarity between reconstructed illusions and real percepts, we also regressed the profile of the redness map of reconstructions from the illusion and positive control conditions onto the illusory or real color surface maps (Figure 10.5E and 10.5F). For the within-individual analysis, in the Ehrenstein configuration, the illusory surface coefficients were slightly lower than those in the positive control condition (Figure 10.5G), indicating that the reconstructed illusory color percept closely resembled the physically presented surface. In contrast, in the Varin configuration, the difference between the illusion and positive control conditions was generally larger (Figure 10.5H). Similar patterns were observed in across-individual conditions. Reconstructions using the content loss-based converter produced illusory surface coefficients that were more comparable to the within-individual results than those obtained using the brain loss-based converter. These results indicate that subjective representations of color spreading illusions can be converted across individuals through neural code conversion.

10.3.3 Neural code conversion during selective attention

We next examined whether brain activity during selective attention could be aligned across individuals. In attention trials, subjects viewed a superposition of two natural images (45 combinations from 10 images) and were instructed to attend to one of them while ignoring the other. This task made the attended image more perceptually dominant. Brain activity from these trials was converted from source to target subjects using the same neural code converters and used to reconstruct images, as in previous tasks. To isolate the effect of attention, we also reconstructed images from brain activity recorded during single-image viewing, in which no attention was directed. Within-individual reconstructions were also performed for comparison.

Figure 10.6 show example reconstructions from the attention and single-image viewing trials. For the within-individual condition, the reconstructed images more closely resembled the attended images, capturing their overall shape, color, and fine details more strongly than those of the unattended images. Even when the visual input was identical, the reconstructions differed substantially depending on the direction of attention. The quality of reconstructions from the attention trials was slightly lower than that from single-image viewing trials, suggesting that attention effectively modulated the perceptual representation captured in brain activity.

For the across-individual condition, similar patterns were observed. Reconstructions based on converted brain activity also resembled the attended image more than the unattended

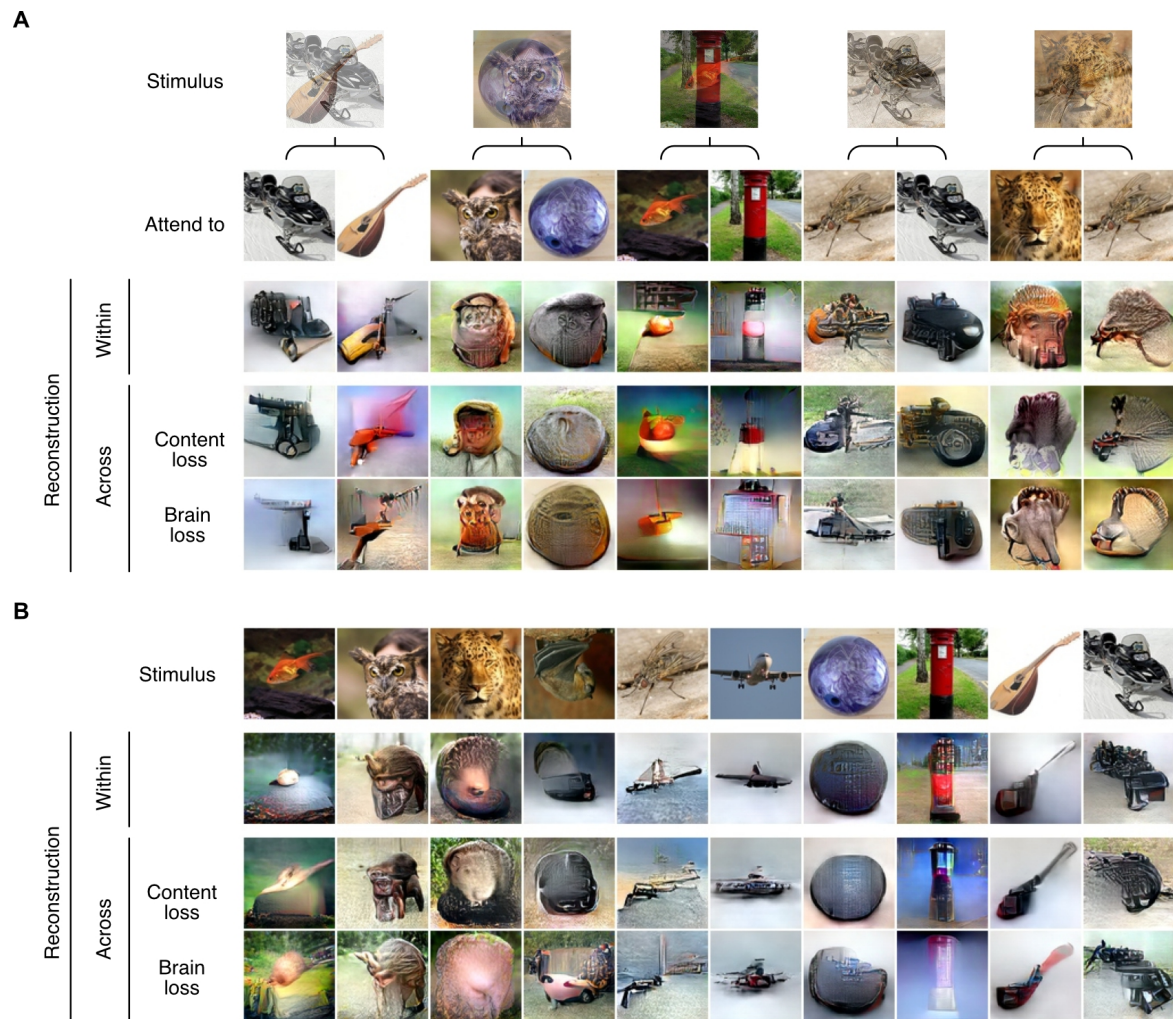


Fig. 10.6 Inter-individual reconstruction for attended images and single images. (A) Reconstructions from attention trials. The reconstructions under the three analytical conditions for each stimulus image were all from the same source subject (VC; source: Subject 1, target: Subject 2). (B) Reconstructions from single-image viewing trails. The reconstructions are shown as in (A) but with single-image trails.

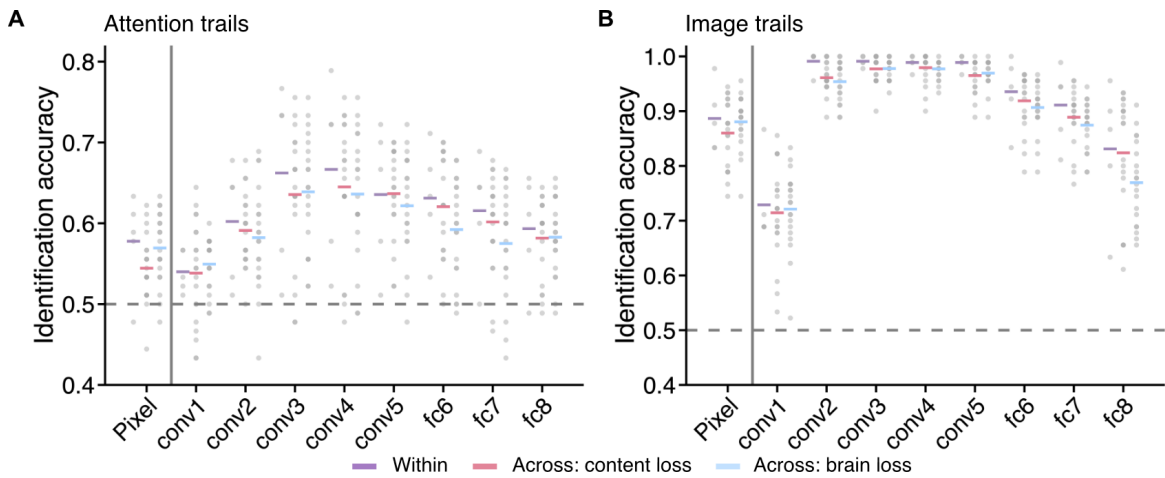


Fig. 10.7 Quantitative evaluation of reconstructed attended images and single images. (A) Identification accuracy of attended image reconstructions. Pairwise identification was performed using the pixel values and the extracted DNN feature values of the reconstructions. The DNN features were extracted from 5 convolutional (conv) layers and 3 fully connected (fc) layers of the AlexNet model. The mean identification accuracy was calculated over all reconstructed images for each subject or individual pair, and then averaged across all subjects or pairs to obtain the group-level mean ($n = 5$ subjects; 20 individual pairs; dashed lines, 50% chance level). (B) Identification accuracy of single-image reconstruction.

one, indicating that neural code conversion preserved attention-modulated representations across individuals. The difference between attended and unattended image content was particularly evident when using the content loss-based converter, while the brain loss-based converter showed weaker attention selectivity in some cases.

To quantitatively evaluate the reconstructions from attention and single-image viewing trials, we performed a pairwise identification analysis using both pixel values and DNN features extracted from the reconstructed images (see Subsection 10.2.3). For each reconstruction, we used its feature pattern to identify the corresponding stimulus image from two candidates by comparing feature correlations. We then calculated the mean identification accuracy across all reconstructions for each individual or individual pair. Figure 10.7 show the mean identification accuracies averaged across subjects, with each dot representing one subject (within-individual) or individual subject pairs (across-individual). For the within-individual condition, attended image identification showed lower accuracy at the earliest DNN layer (conv1), but improved at lower-to-middle layers, peaking around conv3-5 layers (Figure 10.7A). The identification accuracies in the attention condition were generally lower across all layers compared to those in the single-image viewing condition (Figure 10.7B).

The across-individual conditions exhibited similar profiles, with peak performance at middle-to-higher layers, although overall accuracy was slightly lower than that of the within-

individual condition in both attended and single-image image identification (Figure 10.7). Among the two converters, the content loss-based approach outperformed the brain loss-based approach across most layers. These results demonstrate that attention-modulated representations can be preserved across individuals.

10.3.4 Neural code conversion for visual imagery

Unlike perceptual tasks involving external stimuli, visual imagery requires internally generated representations. We similarly investigated whether the brain activity during visual imagery could also be aligned across individuals. In imagery trials, subjects imagined specific artificial or natural images. Brain activity collected during imagery was converted from source to target subjects using the same neural code converters as in previous tasks, and the converted activity was used to reconstruct the imagined images. Within-individual reconstructions were also performed for comparison. As in the attention analysis, we used identification analysis to quantitatively evaluate the reconstruction of imagined images.

Figure 10.8 show example reconstructions of imagined artificial shapes and natural images. In the within-individual condition, the reconstructed images captured coarse structural features of the imagined content, such as overall shape and dominant contours. Artificial shapes were more reliably reconstructed, often preserving clear geometric structure. In contrast, reconstructed natural images showed weaker resemblance to the target images, with only partial textures emerging and generally lower clarity compared to reconstructions from perception-based tasks. In the across-individual condition, both types of converters produced similar reconstructions that reflected the content of the imagined artificial and natural images, although the structural fidelity was slightly degraded compared to the within-individual condition.

The quantitative results showed that identification accuracy was consistently highest in the within-individual condition across image types and DNN layers (Figure 10.9). Between the two across-individual conditions, the content loss-based converter and brain loss-based converter yielded overall comparable performance. For artificial shapes, the content loss-based converter outperformed the brain loss-based converter in low-to-middle DNN layers but showed slightly lower accuracy in higher layers (Figure 10.9A). Natural images showed a different pattern, with the brain loss-based converter performing better in low-to-middle layers and the content loss-based converter showing higher accuracy in higher layers (Figure 10.9B). These results suggest that neural code conversion can preserve internally generated visual representations across individuals, with performance varying depending on image type.

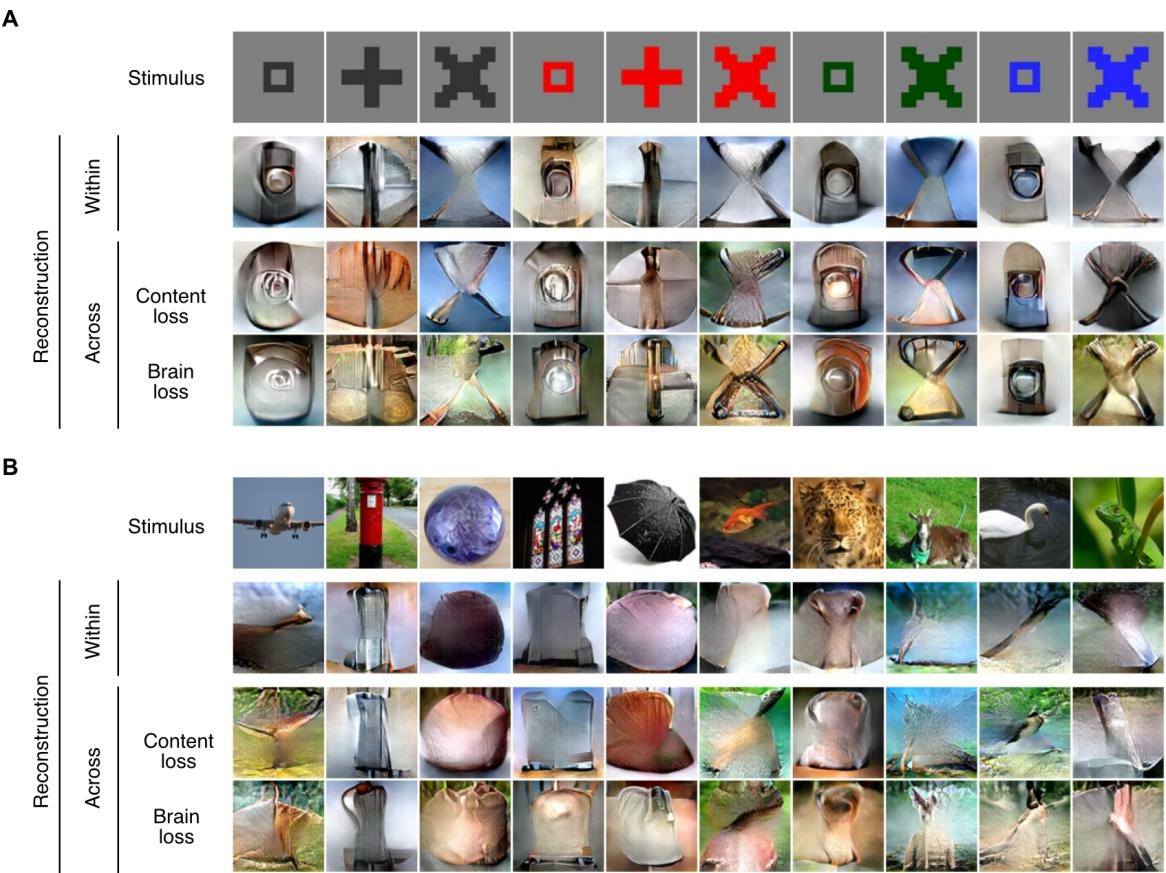


Fig. 10.8 Inter-individual reconstruction for imagined images. (A) Reconstructions of imagined artificial images. The reconstructions under the three analytical conditions for each stimulus image were all from the same source subject (VC; source: Subject 1, target: Subject 2). (B) Reconstructions of imagined natural images. The reconstructions are shown as in (A) but with natural images.

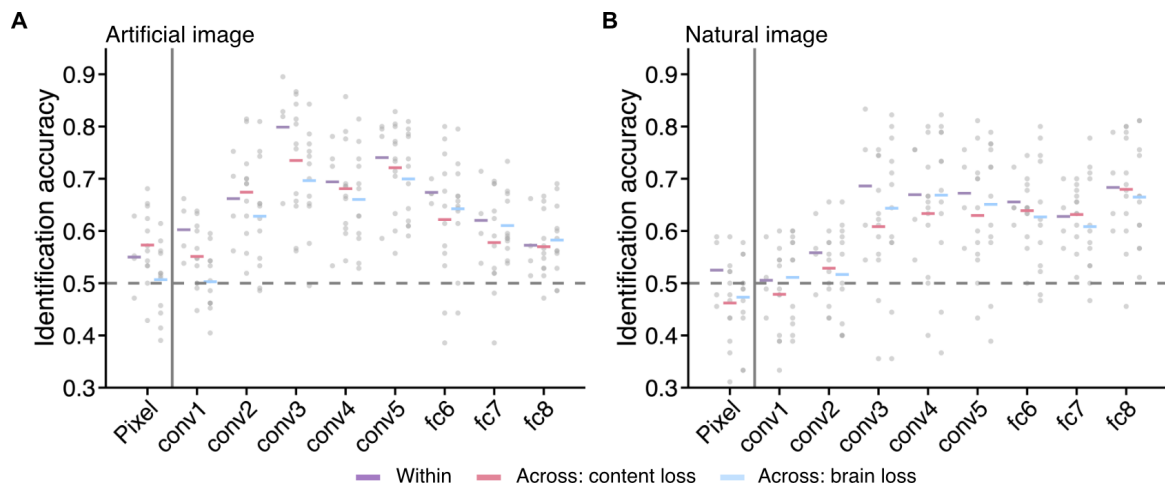


Fig. 10.9 Quantitative evaluation of reconstructed imagined images. (A) Identification accuracy of artificial image reconstructions. Pairwise identification was performed using the pixel values and the extracted DNN feature values of the reconstructions. The DNN features were extracted from 5 convolutional (conv) layers and 3 fully connected (fc) layers of the AlexNet model. The mean identification accuracy was calculated over all reconstructed images for each subject or individual pair, and then averaged across all subjects or pairs to obtain the group-level mean ($n = 4$ subjects; 12 individual pairs; dashed lines, 50% chance level). (B) Identification accuracy of natural image reconstruction.

10.4 Discussion

This chapter investigated whether subjective visual experiences such as illusory perception, selective attention, and visual imagery could be converted across individuals using neural code conversion and visual image reconstruction. Note that the converter, decoder, and reconstruction models had been trained on fMRI responses to natural images, which did not involve the images used in the test stage (attention, illusion, and imagery tasks). This design ensures that the models generalize to novel perceptual and cognitive conditions not seen during training. We found that both line and color illusions could be reliably converted across individuals, with inter-individual reconstructions capturing perceptual features such as illusory contours and color spreading that were not physically present in the stimulus. In the attention task, the reconstructed images from converted brain activity more closely resembled the attended images than the unattended ones. This demonstrates that attention-modulated representations in the visual cortex can be preserved across individuals through neural code conversion. In the imagery task, reconstructed images of imagined artificial shapes preserved geometric structure, and those of natural images retained partial semantic and textural content. Despite lower reconstruction fidelity compared to perception, identification analysis confirmed that internal visual representations could be converted across individuals.

These results suggest that subjective visual experiences may partially share representation across individuals.

Our study builds on the assumption that the brain uses a shared or partly shared representational space to encode both externally driven perception and internally generated subjective experiences (Kamitani & Tong, 2005). This shared space allows neural code converters trained on stimulus-driven neural responses to be applied to subjective experiences. While previous studies have decoded subjective states using stimulus-trained models within individuals (Cheng et al., 2023; Horikawa & Kamitani, 2022; Shen et al., 2019b), our approach introduces an additional requirement: the representational space should also be at least partly functionally aligned across individuals. When both within-individual and across-individual consistency are satisfied, neural code conversion makes it possible to convert and reconstruct subjective visual experiences such as illusions, attention, and imagery across different brains. Our findings suggest that certain aspects of subjective experience may be supported by shared neural representational structures that are at least partially consistent across individuals.

Chapter 11

Auditory neural code conversion

11.1 Introduction

Neural code conversion has proven effective for aligning brain activity patterns across individuals, enabling accurate decoding and reconstruction of visual stimuli. Prior chapters demonstrated that content loss-based converters trained on hierarchical DNN features can preserve fine-grained visual representations and generalize across individuals, datasets, decoding schemes, and experiment tasks. However, these findings were limited to the visual domain. Whether such conversion approach can be extended to other sensory domains remains unclear.

Our approach use latent representations from deep neural networks that are trained to perform the task of interest to align brain activity. This general principle allows the method to be extended to a wide range of sensory modalities or cognitive domains (Figure 11.1). Like vision, recent studies have shown that deep neural networks trained on natural sound recognition tasks can capture feature hierarchies that align with cortical processing stages (Park et al., 2023). This suggests that content-based neural code conversion may also be applicable to auditory stimuli, using DNN-derived auditory features as content representations for functional alignment across individuals. Task-relevant latent representations derived from neural networks trained on other domains such as language or decision-making may also serve as a basis for functional alignment across individuals. For example, representations from foundation models, including Generative Pre-trained Transformers (GPTs) for text processing and SpeechT5 (Ao et al., 2021) for spoken language, could be used to align brain activity during semantic tasks such as reading or listening (Nastase et al., 2021). Similarly, representations from models like Centaur (Binz et al., 2024), trained on psychological decision-making tasks, could be used to align neural activity during decision-making task.

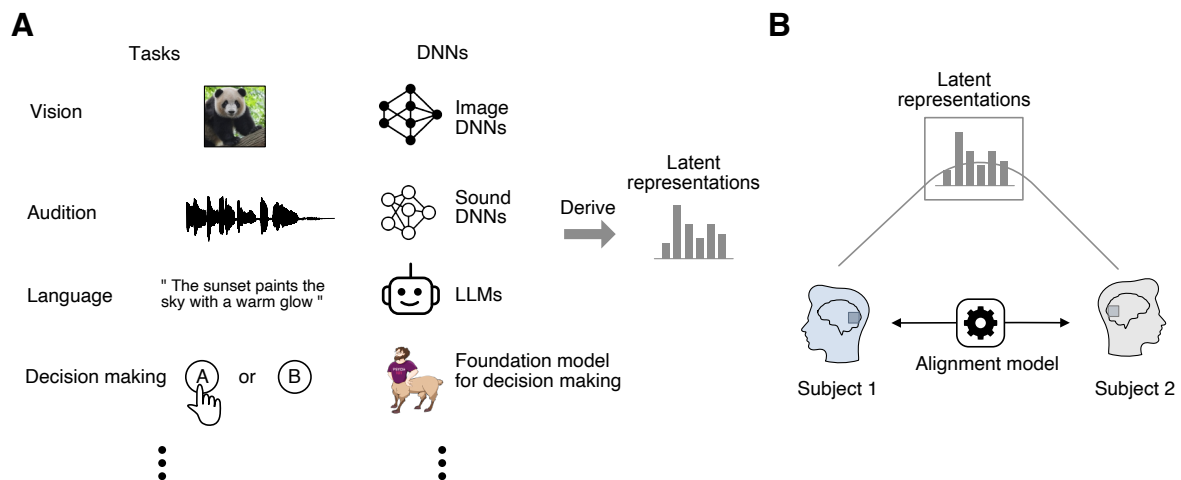


Fig. 11.1 Align brains via latent representations derived from DNNs for the task of interest.

In this chapter, we preliminarily examine this general approach by exploring the feasibility of auditory neural code conversion. We apply both content loss- and brain loss-based conversion methods to fMRI responses to natural sounds and evaluate performance in multiple auditory cortical regions on the DeepSoundRecon dataset. To assess whether auditory representations are preserved after conversion, we perform decoding analysis that decode auditory DNN features from the converted brain activity. By comparing the two conversion approaches and examining results across early and higher-order auditory areas, we aim to determine whether the content-loss based neural code conversion method can generalize beyond the visual domain.

11.2 Methods

11.2.1 Auditory fMRI data

We used the DeepSoundRecon dataset (See Chapter 2 for more details) for auditory neural code conversion analysis. The DeepSoundRecon dataset consists of 1,200 training audio clips and 50 test audio clips (8 s each) with corresponding fMRI responses (four repetitions for each training clip and eight repetitions for each test clip) from five subjects. Subject 1 (S1), a pilot subject, was excluded due to the exploratory nature of their data (Park et al., 2023). Whole-brain fMRI data were collected while subjects listened to auditory stimuli. The analysis used ROIs defined by the dataset, including the whole auditory cortex (AC), A1, LBelt, and PBelt from the early auditory cortex, as well as A4 and A5 from the auditory association cortex (Figure 11.2).

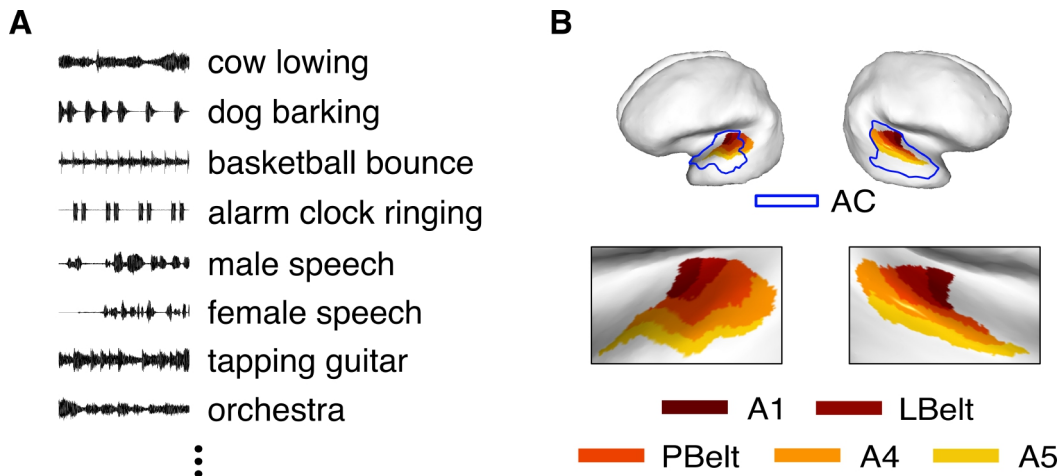


Fig. 11.2 Auditory stimuli and ROIs. (A) Example of sound stimuli from the Deep-SoundRecon dataset. (B) The ROIs of the auditory cortex. The analysis used the whole auditory cortex (AC), A1, LBelt, PBelt from the early auditory cortex, and A4 and A5 from the auditory association cortex.

11.2.2 Auditory content representations

In the auditory analysis, we used the VGGish-ish DNN model (Iashin & Rahtu, 2021) implemented using the PyTorch library (Paszke et al., 2019) as the content representations for converter training and DNN feature decoding analysis. This model is pre-trained for the recognition task of natural sounds (the pre-trained model is available from <https://github.com/v-iashin/SpecVQGAN>). The model consists of 13 convolutional layers and three fully connected layers, and was used to extract DNN features from the spectrogram of sounds. The number of units in each layer is as follows: conv1_1 and conv1_2, 1,720,320; conv2_1 and conv2_2, 860,160; conv3_1, conv3_2, and conv3_3, 430,080; conv4_1, conv4_2, and conv4_3, 215,040; conv5_1, conv5_2, and conv5_3, 53,760; fc1 and fc2, 4,096; and fc3, 309.

11.2.3 Auditory neural code conversion methods

Similar to the methods described in Chapter 5 for visual stimuli, we applied both content loss-based and brain loss-based neural code conversion approaches to auditory stimuli. In the content loss-based method, a nonlinear MLP model was trained to convert a source subject's brain activity pattern into the target space such that the decoded auditory contents from the converted pattern closely matched the contents of the original sound stimulus (Figure 11.3). The contents are the DNN features extracted from a pre-trained VGGish-ish model (see Subsection 11.2.2 for details). In the brain loss-based method, a ridge regression model was used to directly align the source subject's brain activity with the target subject's measured

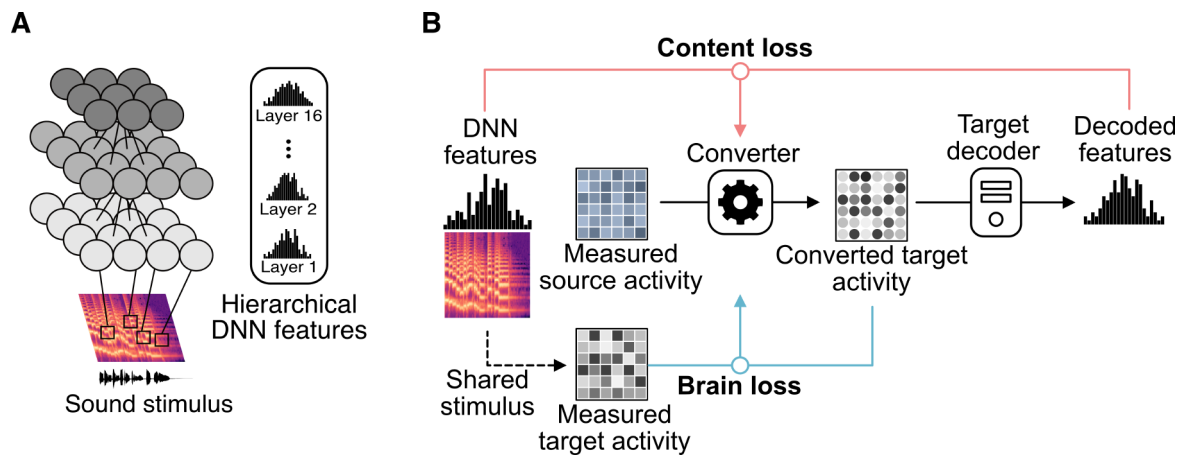


Fig. 11.3 The overview of auditory converter training using content loss-based optimization and brain loss-based optimization. (A) Hierarchical DNN features as auditory content. (B) The content loss-based optimization minimizes the loss between the DNN features decoded from the converted target activity and those extracted from the corresponding auditory stimulus. The brain loss-based optimization minimizes the loss between the converted target activity and the measured target activity.

brain responses. This approach minimizes the difference between the converted and measured target brain activity patterns (Figure 11.3B).

11.2.4 Conversion accuracy for auditory stimuli

To evaluate the performance of the auditory neural code converter, we used two correlation-based metrics: pattern correlation and profile correlation. Pattern correlation refers to the Pearson correlation coefficient between the converted and the measured voxel activity patterns for each test sound. Profile correlation calculates the Pearson correlation coefficient between the sequences of converted and measured responses of a single voxel to the 50 sound test stimuli.

As in the visual domain described in Chapter 5, to account for noise in fMRI measurements, we estimated the noise ceiling by averaging the correlation coefficients between repeated responses to identical auditory stimuli. All correlation coefficients were normalized by dividing the raw values by their corresponding noise ceilings. For each subject pair and ROI, the normalized pattern correlations across all sound stimuli were averaged to obtain the mean pattern accuracy, while the normalized profile correlations across all voxels were averaged to obtain the mean profile accuracy.

11.2.5 DNN feature decoding analysis

To decode auditory content features from brain activity, we trained a ridge regression model to predict DNN features extracted from auditory stimuli based on fMRI activity patterns. Before training, both DNN feature values and voxel responses were normalized. A voxel selection step was applied by calculating Pearson correlations between voxel response sequences and DNN feature values, and the 500 voxels with the highest correlations were retained for training (Horikawa & Kamitani, 2022, 2017). The ridge parameter was set to 100 to enhance model stability.

As in the visual domain described in Chapter 6, test samples were averaged across repetitions to improve the signal-to-noise ratio. Decoder performance was assessed using pattern correlation and profile correlation. For each subject or subject pair, pattern correlations were averaged across test stimuli, and profile correlations were averaged across DNN units to yield the final mean decoding accuracies, which were used in group-level analysis.

11.2.6 Group statistics

We performed group-level statistical inference by showing the group means and confidence intervals for conversion accuracy and decoding accuracy. Each data point represents the mean accuracy corresponding to an individual pair of subjects. Similar to the group statistics for visual neural code conversion analysis in Chapters 5 and 6, due to potential correlations between pairs involving the same subject (dyadic dependency), we applied bootstrapping to these dyadic data points to calculate the group mean and 95% confidence interval. We performed bootstrap sampling separately on the source and target IDs, obtaining data points of the original sample size while ensuring that pairs where the source and target IDs were the same were excluded. The obtained data points were then used to calculate the mean. This process was repeated 1000 times to generate bootstrap replicates, from which we calculated the 95% confidence interval.

11.3 Results

11.3.1 Auditory neural code conversion

To further evaluate the potential of our approach, we conducted preliminary analyses examining neural responses in the auditory cortex during sound stimulus presentation (Park et al., 2023). Using the same content loss-based framework, we analyzed neural code conversion performance across different regions of interest (ROIs). Conversion accuracy was assessed

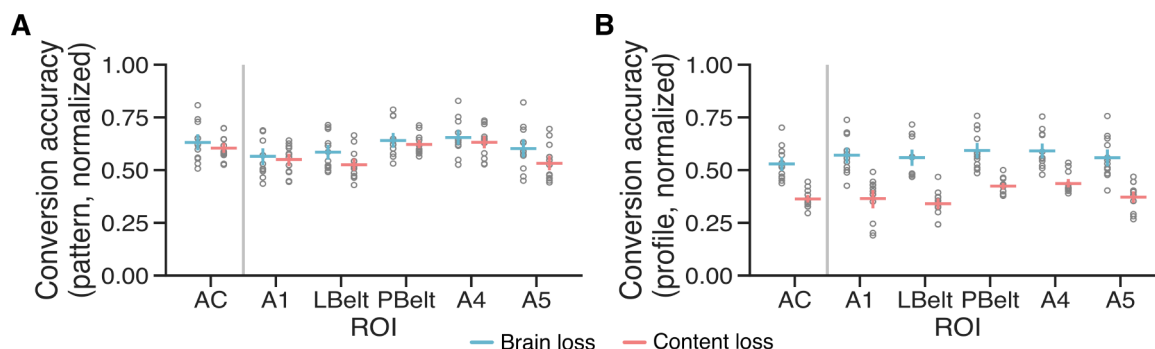


Fig. 11.4 Auditory conversion accuracy. (A) Conversion accuracies measured by pattern correlation. Distributions of the pattern correlation coefficients of 12 individual pairs are shown for the AC and auditory subareas. The horizontal bars represent the mean accuracies across the 12 pairs, with vertical bars representing the 95% C.I., and each dot indicates an individual pair's mean correlation coefficient over stimuli. (B) Conversion accuracies measured by profile correlation.

using both pattern correlation and profile correlation, normalized by their respective noise ceilings. For comparison, we also used the brain loss-based converter in the auditory domain.

Figure 11.4 presents the conversion accuracy under two conversion conditions. In terms of pattern correlation, both content loss-based and brain loss-based converters showed comparable conversion accuracy across all auditory ROIs, with particularly consistent performance in the early auditory areas such as A1. Profile correlation results revealed a distinct performance gap between the two methods. While the brain loss-based converter maintained moderate profile accuracy across ROIs, the content loss-based method showed a noticeable drop, especially in early auditory areas like A1 and PBelt. These results demonstrate that our neural code conversion framework is applicable to auditory neural response, with reasonable accuracy in converting brain responses across individuals. The trend aligns with results in the visual domain (Chapter 5), though there are relatively lower profile correlations in the auditory condition. The successful neural code conversion in both sensory systems suggests the potential versatility of our approach for probing shared neural representations across diverse sensory modalities.

11.3.2 Auditory inter-individual DNN feature decoding

Similar to the visual domain analysis in Chapter 6, we next examined whether fine-grained auditory representations are preserved in the converted brain activity by conducting a DNN feature decoding analysis. Feature decoders were trained to transform fMRI activity from the target subject's auditory cortex to the hierarchical feature representations of the sound stimuli extracted from a pre-trained VGGish-ish model. These decoders were then applied to the

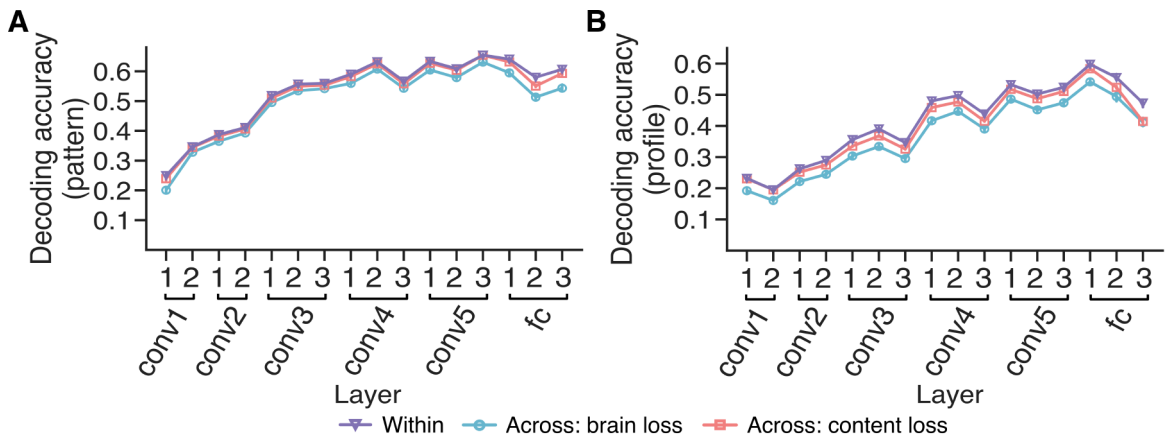


Fig. 11.5 Auditory decoding accuracy. (A) Auditory decoding accuracy measured by pattern correlation. The mean pattern correlation for each layer of the VGGish-ish model is shown for the Within, Brain loss, and Content loss conditions (AC; error bars, 95% C.I. from four subjects for the Within condition, and from 12 pairs for the Brain loss and Content loss conditions). (B) Auditory decoding accuracy measured by profile correlation. The three conditions are compared as in (A) but with profile correlation coefficients.

converted brain activity obtained from the source subject, enabling prediction of the auditory DNN features corresponding to the test stimuli. Decoding performance was quantified using two metrics. First, pattern correlation, defined as the Pearson correlation between the decoded feature vector and the ground-truth feature vector for each test sound. Second, profile correlation, which measures the correlation between decoded and true feature sequences across all test sounds for each individual DNN unit. For each layer, pattern correlations were averaged across all stimuli, and profile correlations were averaged across all units, yielding layer-wise mean decoding accuracies. As a baseline, we also evaluated decoding performance in a within-individual condition, where decoders were trained and tested on data from the same individual. This provided a reference for evaluating the effectiveness of inter-individual decoding via neural code conversion in the auditory domain.

Figure 11.5 presents the decoding performance across the hierarchical layers of the VGGish-ish model, comparing the within-subject condition with the two inter-individual conversion methods. Panel A shows decoding accuracy in terms of pattern correlation, while Panel B shows profile correlation results. In all three conditions, decoding accuracy progressively increased from low-level to high-level layers of the DNN. This trend was consistent across metrics, suggesting that the hierarchical structure of auditory representations is preserved across both within- and inter-individual decoding.

Among the three conditions, within-individual decoding yielded the highest accuracy across nearly all layers, followed closely by content loss-based conversion. Brain loss-

based conversion yielded the lowest performance. While slight performance drops were observed in the two inter-individual conditions, the overall trend demonstrates that fine-grained auditory features can be reliably preserved and decoded after neural code conversion. These results demonstrate the effectiveness of our conversion framework in aligning auditory representations across individuals, suggesting a potential extension of our method beyond the visual domain.

11.4 Discussion

This chapter investigated the preliminary application of our approach to the auditory domain, demonstrating its potential to generalize beyond visual perception. By converting brain activity patterns based on latent auditory representations derived from deep neural networks, we showed that accurate inter-individual conversion and decoding of hierarchical auditory features is feasible. These results highlight the effectiveness of using DNN-derived features as content representations within our framework. The decoded auditory features may also be applied to downstream tasks such as sound reconstruction (Park et al., 2023), further expanding the practical utility of the approach.

We observed that while the pattern correlation for auditory neural code conversion was comparable to that observed in the visual domain, the profile correlation was consistently lower. This discrepancy may be explained by the relatively smaller number of voxels in the auditory cortex compared to the visual cortex, based on typical ROI definitions. In our content loss-based framework, the optimization focuses on minimizing the difference between decoded and true stimulus features. This objective can often be achieved by optimizing only a limited subset of voxels that are strongly associated with auditory representations. As a result, other voxels within the auditory ROI may remain weakly optimized. Since profile correlation reflects the average correlation across all voxels, its value may be affected by these under-optimized voxels, leading to an overall lower score. These observations suggest that the effectiveness of content loss-based optimization may vary across brain regions depending on their spatial extent and voxel distribution. While slightly lower performance was observed in the auditory domain, the overall flexibility of our approach suggests potential for broader application across sensory and cognitive domains.

Chapter 12

General discussion

12.1 Summary

In this thesis, we aimed to develop a functional alignment method that (1) eliminates the need for identical sets of stimuli (shared stimuli) to be presented to individuals during training and (2) captures the fine-grained feature representations of stimuli, enabling downstream tasks such as decoding and image reconstruction with performance comparable to within-individual analyses. To achieve this, we proposed a flexible approach for training neural code converters using content loss-based optimization, which optimizes the converter parameters based on the stimulus content as represented by latent features of a deep neural network (DNN) model, rather than relying on brain activity patterns paired by the same set of stimuli.

In Chapters 5 and 6, we first demonstrated that converters optimized using visual content can accurately convert brain activity patterns across subjects and capture fine-grained visual features for inter-individual image reconstruction, with performance comparable to converters optimized using paired brain activity. We then showed that similar conversion accuracy and inter-individual image reconstruction quality can be obtained using either overlapping or non-overlapping stimuli between converter and decoder training, indicating that shared stimuli are unnecessary for our method. In Chapter 7, an analysis of converters trained using different levels of image representations as visual content supports the notion that the use of multiple hierarchical visual features contributes to accurate neural code conversion. In Chapter 8, we demonstrated that the converted brain activity is not specifically tailored to the specific readout and can be effectively read out by alternative decoding schemes. Also, we showed that images with recognizable object silhouettes can be reconstructed even when the converter is trained with limited amounts of data. In Chapter 9, we further constructed inter-site neural converters across various datasets and demonstrated that inter-site image reconstructions faithfully reflect the viewed images, with quality approaching that of within-

site reconstructions. In Chapter 10, we performed neural code conversion with subjective experience, and found that the converted activity preserved subjective content, enabling inter-individual reconstruction under illusion, attention, and imagery conditions. While our approach is validated in the vision domain, preliminary analysis in Chapter 11 suggests these principles can be extended to other domains. Our results demonstrated that the proposed model could perform functional alignment without the need for shared stimuli and can capture fine-grained features during inter-individual and inter-site neural code conversion.

12.2 Revisit the methodology

Content loss-based converter versus brain loss-based approaches

Throughout this study, we have explored neural code conversion to align brain activity patterns across individuals for inter-subject decoding and reconstruction. We compared two distinct training strategies for neural code converters: the proposed content loss-based approach, which optimizes the converter using stimulus-derived latent features from deep neural networks (DNNs), and the brain loss-based approach, which aligns measured brain activity patterns across individuals directly.

The brain loss-based converter aims to minimize the discrepancy between converted brain activity patterns and the measured activity in a target subject's brain space, typically requiring shared stimuli between subjects during training. While effective in aligning brain activity patterns, this method is inherently tied to brain space and is susceptible to noise and idiosyncratic variability in the response patterns of individual subjects. As shown in Chapters 5 and 6, although this approach achieves reasonable conversion and reconstruction accuracy, it may inadvertently preserve subject-specific noise and idiosyncratic representations.

In contrast, the content loss-based converter aligns representations at the level of stimulus content by minimizing the difference between decoded DNN features from converted brain activity and the true features extracted from the stimulus. This strategy does not require the need for shared stimuli, allowing converter training even when brain responses are collected independently across subjects and sites. The optimization occurs within the feature space of a DNN (e.g., VGG19), which is thought to approximate the hierarchical structure of visual cortical processing. As demonstrated in Chapters 5-10, the content loss-based approach not only enables accurate decoding and image reconstruction, but also generalizes across different decoders, datasets, and stimulus domains, including auditory perception. The empirical results in Chapters 5-11 also support the superiority of the content loss-based method in preserving fine-grained and hierarchical perceptual representations.

The comparison highlights the methodological advantages of content loss-based conversion in capturing shared content representations across individuals. By decoupling alignment from direct neural similarity and grounding it in perceptual feature space, this approach offers a flexible and generalizable framework for functional alignment and inter-individual decoding. Its scalability across subjects, datasets, and modalities suggests strong potential for future applications.

Inter-individual decoding without shared stimuli

Our primary goal was to convert one individual's brain patterns to another's innate brain space without shared stimuli, allowing for use in downstream tasks including decoding and visual image reconstruction as a secondary goal. However, one can achieve inter-individual decoding without aligning brain activity patterns in the innate space. A multi-subject shared decoder can be trained with an alignment layer by optimizing with specific tasks (Bao et al., 2024; Gong et al., 2024; Liu et al., 2024; Quan et al., 2024; Scotti et al., 2024; Zhou et al., 2024). This method can implicitly create a shared latent space for brain data. Similar approaches to handling non-shared stimuli have been applied to decoding in other neuroimaging modalities, such as magnetoencephalography (MEG) (Benchetrit et al., 2023; Défossez et al., 2023), electroencephalography (EEG) (Défossez et al., 2023), and electrophysiological neural recordings (Azabou et al., 2024).

While these inter-individual decoding studies have shown promising results for particular decoding tasks, the alignment of brain activity patterns is performed implicitly with decoder training, and alignment in the innate space is not their goal. Thus, the potential for converting brain activity for general purposes is not carefully examined. In contrast, our neural code conversion method explicitly generated the brain data in the subjects' innate brain space, allowing for explicit evaluation of the conversion across brain spaces. The successful decoding and image reconstruction in Chapters 6-11 obtained by applying the target's original model to the converted brain activity further underscore the effectiveness of our method. The flexibility in converting brain data and seamless compatibility with existing data analysis pipelines potentially pave the way for scalable data analyses.

Limitations

While the proposed content loss-based neural code conversion framework shows promising results across datasets and perception domains, several limitations should be noted.

First, our evaluation is subject to limitations inherent to the dataset. As mentioned in Chapter 9, the stimuli of NSD dataset exhibits limited diversity in semantic and visual

content, with considerable overlap between training and test sets (Shirakawa et al., 2024). Such overlap may introduce bias in inter-site reconstruction, as overlapped content same as test data may be used during converter training. Furthermore, when using multi-subject data for decoding, the likelihood of overlapping semantic content between subjects increases, potentially increasing the risk of spurious prediction of “novel” content (Bao et al., 2024; Gong et al., 2024; Quan et al., 2024; Scotti et al., 2024).

Second, in the visual domain, the model relies on VGG19-derived features as the latent representation of visual content. While effective for static image perception, VGG19 does not fully capture other aspects of visual experience, such as motion, depth, or global scene understanding. This reliance may limit the generalizability of the converter to different visual input.

Third, in the auditory domain, we adopted a similar setup to the visual experiments by using VGGish-ish features as the content representation and applying the same decoder architecture. However, this setup may not fully capture the nature of auditory processing. Therefore, direct transfer of visual-based setups to other sensory domains may require adaptation to domain-specific characteristics.

These limitations suggest that the effectiveness of content loss-based optimization may vary across brain regions and datasets, depending on spatial extent, voxel distribution, and the suitability of content representation.

Implications

The findings of this thesis may have several implications.

First, neural code conversion can improve the efficiency of data usage in brain decoding studies. By enabling inter-individual decoding with limited training data from the subject, this method reduces the need for extensive fMRI scanning, which lowers both economic cost and participant burden. This is particularly important for studies involving children, elderly individuals, or clinical populations with limited scanning tolerance. In addition, this approach facilitates population-level neuroscience analyses and supports the development of generalizable models of brain function that go beyond individual variability.

Second, neural code conversion holds potential for advancing objective diagnosis and treatment in psychiatry and neurology. By comparing converted brain activity across individuals, researchers could identify disease-specific neural signatures or decode pathological mental states such as hallucinations. This may enable novel neuroimaging-based diagnostic tools and inform new intervention strategies using targeted brain stimulation.

Finally, neural code conversion may help future brain-to-brain communication. As brain activity patterns are converted into the subject’s innate brain space while preserving fine-

grained information, they can be used to design high-resolution brain stimulation patterns (Humayun et al., 1999; Lozano et al., 2019; Polanía et al., 2018; Reutsky-Gefen et al., 2013) and to induce mental content similar to that of the source subject. This enables the transmission of mental content between two individuals, providing a foundation for brain-to-brain communication. Such a development could offer a new mode of communication beyond language, with potential applications in accessibility, mutual understanding, and creative collaboration across cultural and linguistic boundaries.

12.3 Future directions

Alternative representations of visual content

Future research could explore alternative representation of visual content. Different DNN architectures, with their unique designs and training paradigms, might capture distinct aspects of neural representations. Beyond DNNs, visual features derived from established models, such as 2D models like HMAX (Mutch & Lowe, 2008; Riesenhuber & Poggio, 1999; Serre et al., 2007), GIST (Oliva & Torralba, 2001), and SIFT (Lowe, 1999); motion models such as Motion Energy Model (Adelson and Bergen, 1985) and Optical Flow (Horn & Schunck, 1981; Lucas & Kanade, 1981); and 3D models like PFH and FPFH (Rusu et al., 2009) and Shape Contexts (Belongie et al., 2002), could serve as alternative latent representations. These models analyze visual content at multiple levels and scales, with some drawing inspiration from or aligning with principles observed in the visual cortex (Cadieu et al., 2014; Güçlü & Van Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014; Leeds et al., 2013; Rice et al., 2014; Riesenhuber & Poggio, 1999; Serre et al., 2007; Yamins et al., 2014). It should also be noted that content representation is not limited to model-derived latent features. When combined with reconstruction models, pixel-level representations of images or movies can directly define content loss. While these diverse approaches show promise, determining optimal features for visual content in neural code conversion requires further investigation.

Homology between less-understood brain regions

Future research could use neural code conversion to explore the homology between brain regions. Our approach provides a data-driven framework for exploring functional homology between brain areas through conversion accuracy. While we demonstrated this using well-characterized visual areas, the method can potentially extend to comparing neural representations between less understood brain regions or even across species. By training converters to predict activity patterns between different source and target regions, we can

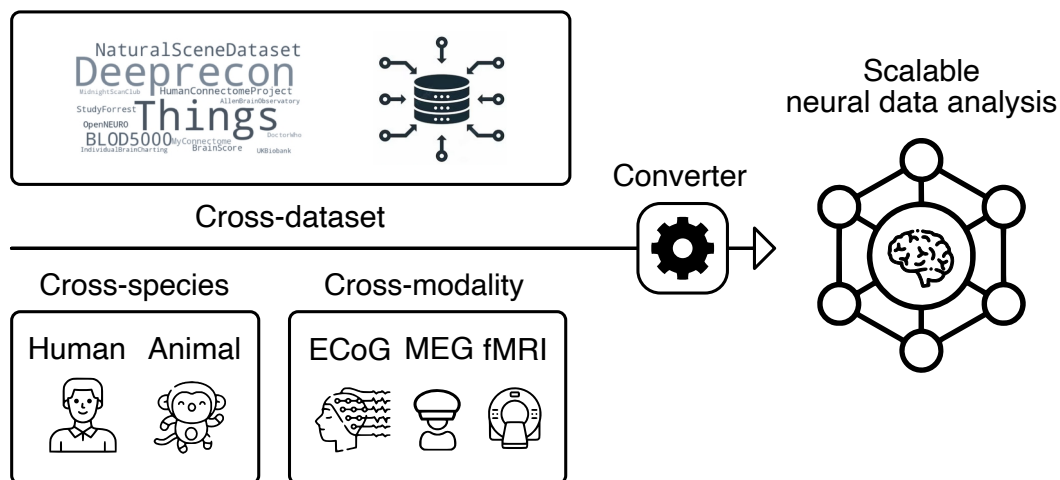


Fig. 12.1 Toward scalable neural data analysis.

quantitatively assess their functional correspondence without requiring prior assumptions about their roles. The conversion accuracy between regions serves as an objective measure of their representational similarity. In our analysis of the visual system, this revealed stronger conversion between areas at similar hierarchical levels (Figure 5.11), consistent with known functional organization. This suggests that our approach may help uncover organizational principles in other neural systems where the functional hierarchy is less clear. The framework may be particularly valuable for establishing cross-species homologies by identifying regions that process information in similar ways, even when their anatomical locations or evolutionary origins differ.

Scalable neural data analysis

Future research could explore a more scalable way to analyze brain data by expanding neural code conversion to broader contexts. Current frameworks can work across individuals and datasets, but are typically trained and evaluated within a single species or imaging modality, which limits their generalizability and scalability. A promising direction is to develop converters that can work not only across datasets, but also across a wide range of species, and different neural recording modalities. This may help pool data across species such as humans and animals, and incorporating multiple neural recording modalities (Figure 12.1). Handling such variability would support more scalable brain data analysis and contribute to a deeper understanding of shared neural representations across domains. Scalability in this direction would ultimately lead to a more scalable neuroscience.

References

- Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A*, 2(2), 284–299.
- Alexander-Bloch, A., Giedd, J. N., & Bullmore, E. (2013). Imaging structural co-variance between human brain regions. *Nature Reviews Neuroscience*, 14(5), 322–336.
- Alfaro-Almagro, F., Jenkinson, M., Bangerter, N. K., Andersson, J. L. R., Griffanti, L., Douaud, G., Sotiropoulos, S. N., Jbabdi, S., Hernandez-Fernandez, M., Vallee, E., Vidaurre, D., Webster, M., McCarthy, P., Rorden, C., Daducci, A., Alexander, D. C., Zhang, H., Dragonu, I., Matthews, P. M., . . . Smith, S. M. (2018). Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. *NeuroImage*, 166, 400–424.
- Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Dowdle, L. T., Caron, B., Pestilli, F., Charest, I., Hutchinson, J. B., Naselaris, T., & Kay, K. (2021). A massive 7T fmri dataset to bridge cognitive and computational neuroscience [Preprint]. *bioRxiv*. <https://www.biorxiv.org/content/10.1101/2021.02.22.432340v1>
- Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., Nau, M., Caron, B., Pestilli, F., Charest, I., et al. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1), 116–126.
- Andersson, J., Jenkinson, M., Smith, S., et al. (2007). Non-linear registration, aka Spatial normalisation FMRIB technical report TR07JA2. *FMRIB Analysis Group of the University of Oxford*, 2(1), e21.
- Andrews, T. J., Halpern, S. D., & Purves, D. (1997). Correlated size variations in human visual cortex, lateral geniculate nucleus, and optic tract. *The Journal of Neuroscience*, 17(8), 2859–2868.
- Ao, J., Wang, R., Zhou, L., Wang, C., Ren, S., Wu, Y., Liu, S., Ko, T., Li, Q., Zhang, Y., et al. (2021). SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing [Preprint]. *arXiv*. <https://arxiv.org/abs/2110.07205>
- 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(1), 26–41.

- Aytar, Y., Vondrick, C., & Torralba, A. (2016). SoundNet: Learning sound representations from unlabeled video. *Advances in Neural Information Processing Systems*, 29.
- Azabou, M., Arora, V., Ganesh, V., Mao, X., Nachimuthu, S., Mendelson, M., Richards, B., Perich, M., Lajoie, G., & Dyer, E. (2024). A unified, scalable framework for neural population decoding. *Advances in Neural Information Processing Systems*, 36.
- Bao, G., Gong, Z., Zhang, Q., Zhou, J., Fan, W., Yi, K., Naseem, U., Hu, L., & Miao, D. (2024). Wills Aligner: A robust multi-subject brain representation learner [Preprint]. *arXiv*. <https://arxiv.org/abs/2404.13282>
- Bazeille, T., Richard, H., Janati, H., & Thirion, B. (2019). Local optimal transport for functional brain template estimation. In A. C. S. Chung, J. C. Gee, P. A. Yushkevich, & S. Bao (Eds.), *Information Processing in Medical Imaging* (pp. 237–248, Vol. 11492). Springer International Publishing.
- 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.
- 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(5032), 716–719.
- Belongie, S., Malik, J., & Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4), 509–522.
- Benchetrit, Y., Banville, H., & King, J.-R. (2023). Brain decoding: Toward real-time reconstruction of visual perception [Preprint]. *arXiv*. <https://arxiv.org/abs/2310.19812>
- Benson, N. C., Jamison, K. W., Arcaro, M. J., Vu, A. T., Glasser, M. F., Coalson, T. S., Van Essen, D. C., Yacoub, E., Ugurbil, K., Winawer, J., & Kay, K. (2018). The Human Connectome Project 7 Tesla retinotopy dataset: Description and population receptive field analysis. *Journal of Vision*, 18(13), 23.
- Benson, N. C., Yoon, J. M. D., Forenzo, D., Engel, S. A., Kay, K. N., & Winawer, J. (2022). Variability of the surface area of the V1, V2, and V3 maps in a large sample of human observers. *The Journal of Neuroscience*, 42(46), 8629–8646.
- Beyer, L., Zhai, X., & Kolesnikov, A. (2022). Better plain ViT baselines for ImageNet-1k [Preprint]. *arXiv*. <https://arxiv.org/abs/2205.01580>
- Bilenko, N. Y., & Gallant, J. L. (2016). Pyrcca: Regularized kernel canonical correlation analysis in python and its applications to neuroimaging. *Frontiers in Neuroinformatics*, 10, 49.

- Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., Dayan, P., Demircan, C., Eckstein, M. K., Éltető, N., et al. (2024). Centaur: A foundation model of human cognition [Preprint]. *arXiv*. <https://arxiv.org/abs/2410.20268>
- Bressan, P., Mingolla, E., Spillmann, L., & Watanabe, T. (1997). Neon color spreading: A review. *Perception*, 26(11), 1353–1366.
- Cadieu, C. F., Hong, H., Yamins, D. L., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., & DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Computational Biology*, 10(12), e1003963.
- Chen, H., Xie, W., Vedaldi, A., & Zisserman, A. (2020). VGGSound: A large-scale audio-visual dataset. *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 721–725.
- Chen, P.-H., Guntupalli, J. S., Haxby, J. V., & Ramadge, P. J. (2014). Joint SVD-Hyperalignment for multi-subject fMRI data alignment. *2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 1–6.
- Chen, P.-H., Zhu, X., Zhang, H., Turek, J. S., Chen, J., Willke, T. L., Hasson, U., & Ramadge, P. J. (2016). A convolutional autoencoder for multi-subject fMRI data aggregation [Preprint]. *arXiv*. <https://arxiv.org/abs/1608.04846>
- Chen, P.-H. C., Chen, J., Yeshurun, Y., Hasson, U., Haxby, J., & Ramadge, P. J. (2015). A reduced-dimension fMRI shared response model. *Advances in Neural Information Processing Systems*, 28.
- Chen, S., He, Z., Han, X., He, X., Li, R., Zhu, H., Zhao, D., Dai, C., Zhang, Y., Lu, Z., Chi, X., & Niu, B. (2019). How big data and high-performance computing drive brain science. *Genomics, Proteomics & Bioinformatics*, 17(4), 381–392.
- Chen, Z., Qing, J., Xiang, T., Yue, W. L., & Zhou, J. H. (2023a). Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22710–22720.
- Chen, Z., Qing, J., & Zhou, J. H. (2023b). Cinematic mindscapes: High-quality video reconstruction from brain activity [Preprint]. *arXiv*. <https://arxiv.org/abs/2305.11675>
- Cheng, F. L., Horikawa, T., Majima, K., Tanaka, M., Abdelhack, M., Aoki, S. C., Hirano, J., & Kamitani, Y. (2023). Reconstructing visual illusory experiences from human brain activity. *Science Advances*, 9(46), ead3906.
- Churchland, P. M. (1998). Conceptual similarity across sensory and neural diversity: The Fodor/Lepore challenge answered. *The Journal of Philosophy*, 95(1), 5–32.
- Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6(1), 27755.

- Ciric, R., Thompson, W. H., Lorenz, R., Goncalves, M., MacNicol, E. E., Markiewicz, C. J., Halchenko, Y. O., Ghosh, S. S., Gorgolewski, K. J., Poldrack, R. A., et al. (2022). TemplateFlow: FAIR-sharing of multi-scale, multi-species brain models. *Nature Methods*, 19(12), 1568–1571.
- Connolly, A. C., Gobbini, M. I., & Haxby, J. V. (2012a). Three virtues of similarity-based multivariate pattern analysis: An example from the human object vision pathway. *Understanding Visual Population Codes: Toward a Common Multivariate Framework for Cell Recording and Functional Imaging*, 335–55.
- Connolly, A. C., Guntupalli, J. S., Gors, J., Hanke, M., Halchenko, Y. O., Wu, Y.-C., Abdi, H., & Haxby, J. V. (2012b). The representation of biological classes in the human brain. *Journal of Neuroscience*, 32(8), 2608–2618.
- Connolly, A. C., Sha, L., Guntupalli, J. S., Oosterhof, N., Halchenko, Y. O., Nastase, S. A., Castello, M. V. d. O., Abdi, H., Jobst, B. C., Gobbini, M. I., & Haxby, J. V. (2016). How the human brain represents perceived dangerousness or “predacity” of animals. *Journal of Neuroscience*, 36(19), 5373–5384.
- Cornelissen, F. W., Wade, A. R., Vladusich, T., Dougherty, R. F., & Wandell, B. A. (2006). No functional magnetic resonance imaging evidence for brightness and color filling-in in early human visual cortex. *The Journal of Neuroscience*, 26(14), 3634–3641.
- Cox, D. D., & Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI) “brain reading”: Detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage*, 19(2), 261–270.
- Çukur, T., Nishimoto, S., Huth, A. G., & Gallant, J. L. (2013). Attention during natural vision warps semantic representation across the human brain. *Nature Neuroscience*, 16(6), 763–770.
- Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis: I. Segmentation and surface reconstruction. *NeuroImage*, 9(2), 179–194.
- Défossez, A., Caucheteux, C., Rapin, J., Kabeli, O., & King, J.-R. (2023). Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5(10), 1097–1107.
- 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.
- Diedrichsen, J., & Kriegeskorte, N. (2017). Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLOS Computational Biology*, 13(4), e1005508.
- Dijkstra, N., Bosch, S. E., & Van Gerven, M. A. (2019). Shared neural mechanisms of visual perception and imagery. *Trends in Cognitive Sciences*, 23(5), 423–434.
- Dohmatob, E., Varoquaux, G., & Thirion, B. (2018). Inter-subject registration of functional images: Do we need anatomical images? *Frontiers in Neuroscience*, 12, 64.

- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale [Preprint]. *arXiv*. <https://arxiv.org/abs/2010.11929>
- Dosovitskiy, A., & Brox, T. (2016). Generating images with perceptual similarity metrics based on deep networks [Preprint]. *arXiv*. <https://arxiv.org/abs/1602.02644>
- Ehrenstein, W. (1941). Über abwandlungen der l. hermannschen helligkeitserscheinung. *Zeitschrift für Psychologie: Organ der Deutschen Gesellschaft für Psychologie*.
- Engel, S. A., Rumelhart, D. E., Wandell, B. A., Lee, A. T., Glover, G. H., Chichilnisky, E.-J., Shadlen, M. N., et al. (1994). fMRI of human visual cortex. *Nature*, 369(6481), 525–525.
- Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, 392(6676), 598–601.
- 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., et al. (2019). fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods*, 16(1), 111–116.
- Feilong, M., Nastase, S. A., Jiahui, G., Halchenko, Y. O., Gobbini, M. I., & Haxby, J. V. (2023). The Individualized Neural Tuning Model: Precise and generalizable cartography of functional architecture in individual brains. *Imaging Neuroscience*, 1, 1–34.
- Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1(1), 1–47.
- Ferrante, M., Boccato, T., Ozcelik, F., VanRullen, R., & Toschi, N. (2024). Through their eyes: Multi-subject brain decoding with simple alignment techniques. *Imaging Neuroscience*, 2, 1–21.
- 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(8), 1973–1980.
- Fischl, B., Sereno, M. I., & Dale, A. M. (1999). Cortical surface-based analysis: II: inflation, flattening, and a surface-based coordinate system. *NeuroImage*, 9(2), 195–207.
- Fonov, V., Evans, A. C., Botteron, K., Almli, C. R., McKinstry, R. C., Collins, D. L., Group, B. D. C., et al. (2011). Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage*, 54(1), 313–327.
- Formisano, E., Kim, D.-S., Di Salle, F., Van De Moortele, P.-F., Ugurbil, K., & Goebel, R. (2003). Mirror-symmetric tonotopic maps in human primary auditory cortex. *Neuron*, 40(4), 859–869.

- Frost, M. A., & Goebel, R. (2012). Measuring structural-functional correspondence: Spatial variability of specialised brain regions after macro-anatomical alignment. *NeuroImage*, 59(2), 1369–1381.
- Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4), 193–202.
- Gao, S., Mishne, G., & Scheinost, D. (2021). Nonlinear manifold learning in functional magnetic resonance imaging uncovers a low-dimensional space of brain dynamics. *Human Brain Mapping*, 42(14), 4510–4524.
- Gaziv, G., Belyi, R., Granot, N., Hoogi, A., Strappini, F., Golan, T., & Irani, M. (2022). Self-supervised natural image reconstruction and large-scale semantic classification from brain activity. *NeuroImage*, 254, 119121.
- Gerardin, P., Abbatecola, C., Devinck, F., Kennedy, H., Dojat, M., & Knoblauch, K. (2018). Neural circuits for long-range color filling-in. *NeuroImage*, 181, 30–43.
- Gillis, G., Bhalerao, G., Blane, J., Mitchell, R., Pretorius, P. M., McCracken, C., Nichols, T. E., Smith, S. M., Miller, K. L., Alfaro-Almagro, F., Raymont, V., Martos, L., Mackay, C. E., & Griffanti, L. (2025). From big data to the clinic: Methodological and statistical enhancements to implement the UK biobank imaging framework in a memory clinic. *Human Brain Mapping*, 46(3), e70151.
- Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., Jenkinson, M., et al. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615), 171–178.
- Gong, Y., Chung, Y.-A., & Glass, J. (2021). AST: Audio Spectrogram Transformer [Preprint]. *arXiv*. <https://arxiv.org/abs/2104.01778>
- Gong, Z., Zhang, Q., Bao, G., Zhu, L., Liu, K., Hu, L., & Miao, D. (2024). MindTuner: Cross-subject visual decoding with visual fingerprint and semantic correction [Preprint]. *arXiv*. <https://arxiv.org/abs/2404.12630>
- Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, 48(1), 63–72.
- Griffiths, T. D., & Warren, J. D. (2002). The planum temporale as a computational hub. *Trends in Neurosciences*, 25(7), 348–353.
- Güçlü, U., & Van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27), 10005–10014.
- 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. *Advances in Neural Information Processing Systems*, 30.

- 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(6), 2919–2934.
- Harrison, S. A., & Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature*, 458(7238), 632–635.
- Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding neural representational spaces using multivariate pattern analysis. *Annual Review of Neuroscience*, 37(1), 435–456.
- Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539), 2425–2430.
- 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(2), 404–416.
- Haxby, J. V., Guntupalli, J. S., Nastase, S. A., & Feilong, M. (2020). Hyperalignment: Modeling shared information encoded in idiosyncratic cortical topographies. *eLife*, 9, e56601.
- Hayashi, S., Caron, B. A., Heinsfeld, A. S., Vinci-Booher, S., McPherson, B., Bullock, D. N., Bertò, G., Niso, G., Hanekamp, S., Levitas, D., Ray, K., MacKenzie, A., Avesani, P., Kitchell, L., Leong, J. K., Nascimento-Silva, F., Koudoro, S., Willis, H., Jolly, J. K., ... Pestilli, F. (2024). brainlife.io: A decentralized and open-source cloud platform to support neuroscience research. *Nature Methods*, 21(5), 809–813.
- Haynes, J.-D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7(7), 523–534.
- Haynes, J.-D., Sakai, K., Rees, G., Gilbert, S., Frith, C., & Passingham, R. E. (2007). Reading hidden intentions in the human brain. *Current Biology*, 17(4), 323–328.
- He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2021). Masked autoencoders are scalable vision learners [Preprint]. *arXiv*. <https://arxiv.org/abs/2111.06377>
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.
- Hebart, M. N., Contier, O., Teichmann, L., Rockter, A. H., Zheng, C. Y., Kidder, A., Corriveau, A., Vaziri-Pashkam, M., & Baker, C. I. (2023). THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *eLife*, 12, e82580.
- Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Van Wicklin, C., & Baker, C. I. (2019). THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLOS ONE*, 14(10), e0223792.

- Hegd , J., & Van Essen, D. C. (2000). Selectivity for complex shapes in primate visual area V2. *The Journal of Neuroscience*, 20(5), RC61.
- Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R. J., & Wilson, K. (2017). Cnn architectures for large-scale audio classification [Preprint]. *arXiv*. <https://arxiv.org/abs/1609.09430>
- Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(5), 393–402.
- Himmelberg, M. M., Winawer, J., & Carrasco, M. (2022). Linking individual differences in human primary visual cortex to contrast sensitivity around the visual field. *Nature Communications*, 13(1), 3309.
- 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.
- Ho, M.-L., & Schwarzkopf, D. S. (2022). The human primary visual cortex (V1) encodes the perceived position of static but not moving objects. *Communications Biology*, 5(1), 1–8.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.
- Hong, S. W., & Tong, F. (2017). Neural representation of form-contingent color filling-in in the early visual cortex. *Journal of Vision*, 17(13), 10.
- Horikawa, T., Tamaki, M., Miyawaki, Y., & Kamitani, Y. (2013). Neural decoding of visual imagery during sleep. *Science*, 340(6132), 639–642.
- Horikawa, T., & Kamitani, Y. (2022). Attention modulates neural representation to render reconstructions according to subjective appearance. *Communications Biology*, 5(1), 34.
- Horikawa, T., & Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, 8(1), 15037.
- Horn, B. K., & Schunck, B. G. (1981). Determining optical flow. *Artificial Intelligence*, 17(1-3), 185–203.
- 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(2), 91–109.
- Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2018). Densely connected convolutional networks [Preprint]. *arXiv*. <https://arxiv.org/abs/1608.06993>

- Humayun, M. S., de Juan Jr, E., Weiland, J. D., Dagnelie, G., Katona, S., Greenberg, R., & Suzuki, S. (1999). Pattern electrical stimulation of the human retina. *Vision Research*, 39(15), 2569–2576.
- Huo, J., Wang, Y., Qian, X., Wang, Y., Li, C., Feng, J., & Fu, Y. (2024). NeuroPictor: Refining fMRI-to-image reconstruction via multi-individual pretraining and multi-level modulation [Preprint]. *arXiv*. <https://arxiv.org/abs/2403.18211>
- 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(7600), 453–458.
- Hutton, C., Bork, A., Josephs, O., Deichmann, R., Ashburner, J., & Turner, R. (2002). Image distortion correction in fMRI: a quantitative evaluation. *NeuroImage*, 16(1), 217–240.
- Iashin, V., & Rahtu, E. (2021). Taming visually guided sound generation [Preprint]. *arXiv*. <https://arxiv.org/abs/2110.08791>
- Jafari-Khouzani, K., & Soltanian-Zadeh, H. (2005). Radon transform orientation estimation for rotation invariant texture analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6), 1004–1008.
- Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., & Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. *Proceedings of the 22nd ACM International Conference on Multimedia*, 675–678.
- Jiahui, G., Feilong, M., Visconti di Oleggio Castello, M., Nastase, S. A., Haxby, J. V., & Gobbini, M. I. (2023). Modeling naturalistic face processing in humans with deep convolutional neural networks. *Proceedings of the National Academy of Sciences*, 120(43), e2304085120.
- Kamitani, Y., & Tong, F. (2006). Decoding seen and attended motion directions from activity in the human visual cortex. *Current Biology*, 16(11), 1096–1102.
- Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, 8(5), 679–685.
- Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17(11), 4302–4311.
- Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, 452(7185), 352–355.
- Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, 10(11), e1003915.
- Klein, B. P., Harvey, B. M., & Dumoulin, S. O. (2014). Attraction of position preference by spatial attention throughout human visual cortex. *Neuron*, 84(1), 227–237.

- Knebel, J.-F., & Murray, M. M. (2012). Towards a resolution of conflicting models of illusory contour processing in humans. *NeuroImage*, 59(3), 2808–2817.
- Kok, P., Bains, L. J., van Mourik, T., Norris, D. G., & de Lange, F. P. (2016). Selective activation of the deep layers of the human primary visual cortex by top-down feedback. *Current Biology*, 26(3), 371–376.
- Kok, P., & de Lange, F. P. (2014). Shape perception simultaneously up- and downregulates neural activity in the primary visual cortex. *Current Biology*, 24(13), 1531–1535.
- Kosslyn, S. M., Ganis, G., & Thompson, W. L. (2001). Neural foundations of imagery. *Nature Reviews Neuroscience*, 2(9), 635–642.
- Kourtzi, Z., & Kanwisher, N. (2000). Cortical regions involved in perceiving object shape. *Journal of Neuroscience*, 20(9), 3310–3318.
- Kriegeskorte, N. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*.
- Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, 103(10), 3863–3868.
- Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, 17(8), 401–412.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25.
- Kruger, N., Janssen, P., Kalkan, S., Lappe, M., Leonardis, A., Piater, J., Rodriguez-Sanchez, A. J., & Wiskott, L. (2013). Deep hierarchies in the primate visual cortex: What can we learn for computer vision? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), 1847–1871.
- Laakso, A., & Cottrell, G. W. (2005). Churchland on connectionism. In B. L. Keeley (Ed.), *Paul churchland* (pp. 113–153). Cambridge University Press.
- LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4), 541–551.
- LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324.
- Leeds, D. D., Seibert, D. A., Pyles, J. A., & Tarr, M. J. (2013). Comparing visual representations across human fMRI and computational vision. *Journal of Vision*, 13(13), 25–25.
- Leipold, S., Greber, M., & Elmer, S. (2019). Perception and cognition in absolute pitch: Distinct yet inseparable. *The Journal of Neuroscience*, 39(30), 5839–5841.

- Lescroart, M. D., & Gallant, J. L. (2019). Human scene-selective areas represent 3D configurations of surfaces. *Neuron*, 101(1), 178–192.
- Li, M., Liu, F., Jiang, H., Lee, T. S., & Tang, S. (2017). Long-term two-photon imaging in awake macaque monkey. *Neuron*, 93(5), 1049–1057.e3.
- Liang, P., Shi, L., Chen, N., Luo, Y., Wang, X., Liu, K., Mok, V. C., Chu, W. C., Wang, D., & Li, K. (2015). Construction of brain atlases based on a multi-center MRI dataset of 2020 Chinese adults. *Scientific Reports*, 5(1), 18216.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. *Proceedings of the 13th European Conference on Computer Vision*, 740–755.
- Liu, Y., Ma, Y., Zhu, G., Jing, H., & Zheng, N. (2024). See through their minds: Learning transferable neural representation from cross-subject fMRI [Preprint]. *arXiv*. <https://arxiv.org/abs/2403.06361>
- Llera, A., Wolfers, T., Mulders, P., & Beckmann, C. F. (2019). Inter-individual differences in human brain structure and morphology link to variation in demographics and behavior. *eLife*, 8, e44443.
- Lowe, D. G. (1999). Object recognition from local scale-invariant features. *Proceedings of the Seventh IEEE International Conference on Computer Vision*, 2, 1150–1157.
- Lozano, A. M., Lipsman, N., Bergman, H., Brown, P., Chabardes, S., Chang, J. W., Matthews, K., McIntyre, C. C., Schlaepfer, T. E., Schulder, M., et al. (2019). Deep brain stimulation: Current challenges and future directions. *Nature Reviews Neurology*, 15(3), 148–160.
- Lu, Z., & Golomb, J. D. (2023). Generate your neural signals from mine: Individual-to-individual EEG converters [Preprint]. *arXiv*. <https://arxiv.org/abs/2304.10736>
- Lucas, B. D., & Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, 2, 674–679.
- Madan, C. R. (2022). Scan once, analyse many: Using large open-access neuroimaging datasets to understand the brain. *Neuroinformatics*, 20(1), 109–137.
- Marie, D., Jobard, G., Crivello, F., Perchey, G., Petit, L., Mellet, E., Joliot, M., Zago, L., Mazoyer, B., & Tzourio-Mazoyer, N. (2015). Descriptive anatomy of Heschl’s gyri in 430 healthy volunteers, including 198 left-handers. *Brain Structure and Function*, 220, 729–743.
- Mazziotta, J., Toga, A., Evans, A., Fox, P., Lancaster, J., Zilles, K., Woods, R., Paus, T., Simpson, G., Pike, B., et al. (2001). A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM). *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 356(1412), 1293–1322.

- Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision: Two cortical pathways. *Trends in Neurosciences*, 6, 414–417.
- Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M.-a., Morito, Y., Tanabe, H. C., Sadato, N., & Kamitani, Y. (2008). Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, 60(5), 915–929.
- Muraki, Y. (2024). *Improving visual image reconstruction from brain activity using texture and structure similarity losses* [Master's thesis, Kyoto University].
- Mutch, J., & Lowe, D. G. (2008). Object class recognition and localization using sparse features with limited receptive fields. *International Journal of Computer Vision*, 80, 45–57.
- Naselaris, T., Allen, E., & Kay, K. (2021). Extensive sampling for complete models of individual brains. *Current Opinion in Behavioral Sciences*, 40, 45–51.
- Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., & Gallant, J. L. (2009). Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63(6), 902–915.
- Nastase, S. A., Gazzola, V., Hasson, U., & Keysers, C. (2019). Measuring shared responses across subjects using intersubject correlation. *Social Cognitive and Affective Neuroscience*, 14(6), 667–685.
- Nastase, S. A., Liu, Y.-F., Hillman, H., Zadbood, A., Hasenfratz, L., Keshavarzian, N., Chen, J., Honey, C. J., Yeshurun, Y., Regev, M., et al. (2021). The “Narratives” fMRI dataset for evaluating models of naturalistic language comprehension. *Scientific Data*, 8(1), 250.
- Nenning, K.-H., Kollndorfer, K., Schöpf, V., Prayer, D., & Langs, G. (2015). Multi-subject manifold alignment of functional network structures via joint diagonalization. In S. Ourselin, D. C. Alexander, C.-F. Westin, & M. J. Cardoso (Eds.), *Information Processing in Medical Imaging* (pp. 462–473, Vol. 9123). Springer International Publishing.
- Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., & Clune, J. (2016). Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *Advances in Neural Information Processing Systems*, 29.
- Niazi, A. M., van den Broek, P. L. C., Klanke, S., Barth, M., Poel, M., Desain, P., & van Gerven, M. A. J. (2014). Online decoding of object-based attention using real-time fMRI. *European Journal of Neuroscience*, 39(2), 319–329.
- Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21(19), 1641–1646.
- Nonaka, S., Majima, K., Aoki, S. C., & Kamitani, Y. (2021). Brain hierarchy score: Which deep neural networks are hierarchically brain-like? *iScience*, 24(9), 103013.

- Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, 10(9), 424–430.
- Ogawa, S., Lee, T. M., Kay, A. R., & Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences*, 87(24), 9868–9872.
- Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42, 145–175.
- Park, J.-Y., Tsukamoto, M., Tanaka, M., & Kamitani, Y. (2023). Sound reconstruction from human brain activity via a generative model with brain-like auditory features [Preprint]. *arXiv*. <https://arxiv.org/abs/2306.11629>
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32.
- Paulk, A. C., Kfir, Y., Khanna, A. R., Mustroph, M. L., Trautmann, E. M., Soper, D. J., Stavisky, S. D., Welkenhuysen, M., Dutta, B., Shenoy, K. V., Hochberg, L. R., Richardson, R. M., Williams, Z. M., & Cash, S. S. (2022). Large-scale neural recordings with single neuron resolution using Neuropixels probes in human cortex. *Nature Neuroscience*, 25(2), 252–263.
- Penny, W. D., Friston, K. J., Ashburner, J. T., Kiebel, S. J., & Nichols, T. E. (2011). *Statistical parametric mapping: The analysis of functional brain images*. Elsevier.
- Polanía, R., Nitsche, M. A., & Ruff, C. C. (2018). Studying and modifying brain function with non-invasive brain stimulation. *Nature Neuroscience*, 21(2), 174–187.
- Poldrack, R. A., Mumford, J. A., & Nichols, T. E. (2011). *Handbook of functional MRI data analysis*. Cambridge University Press.
- Qian, X., Wang, Y., Huo, J., Feng, J., & Fu, Y. (2023). fMRI-PTE: A large-scale fMRI pretrained transformer encoder for multi-subject brain activity decoding [Preprint]. *arXiv*. <https://arxiv.org/abs/2311.00342>
- Quan, R., Wang, W., Tian, Z., Ma, F., & Yang, Y. (2024). Psychometry: An omnifit model for image reconstruction from human brain activity. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 233–243.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, 8748–8763.
- Reddy, L., Kanwisher, N. G., & VanRullen, R. (2009). Attention and biased competition in multi-voxel object representations. *Proceedings of the National Academy of Sciences*, 106(50), 21447–21452.

- Ren, Z., Li, J., Xue, X., Li, X., Yang, F., Jiao, Z., & Gao, X. (2021). Reconstructing seen image from brain activity by visually-guided cognitive representation and adversarial learning. *NeuroImage*, 228, 117602.
- Reutsky-Gefen, I., Golan, L., Farah, N., Schejter, A., Tsur, L., Brosh, I., & Shoham, S. (2013). Holographic optogenetic stimulation of patterned neuronal activity for vision restoration. *Nature Communications*, 4(1), 1509.
- Riccelli, R., Toschi, N., Nigro, S., Terracciano, A., & Passamonti, L. (2017). Surface-based morphometry reveals the neuroanatomical basis of the five-factor model of personality. *Social Cognitive and Affective Neuroscience*, nsw175.
- Rice, G. E., Watson, D. M., Hartley, T., & Andrews, T. J. (2014). Low-level image properties of visual objects predict patterns of neural response across category-selective regions of the ventral visual pathway. *Journal of Neuroscience*, 34(26), 8837–8844.
- Richard, H., Martin, L., Pinho, A. L., Pillow, J., & Thirion, B. (2019). Fast shared response model for fMRI data [Preprint]. *arXiv*. <https://arxiv.org/abs/1909.12537>
- Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11), 1019–1025.
- Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6), 386–408.
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533–536.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3), 211–252.
- Rusu, R. B., Blodow, N., & Beetz, M. (2009). Fast point feature histograms (FPFH) for 3d registration. *2009 IEEE International Conference on Robotics and Automation*, 3212–3217.
- Sabuncu, M. R., Singer, B. D., Conroy, B., Bryan, R. E., Ramadge, P. J., & Haxby, J. V. (2010). Function-based intersubject alignment of human cortical anatomy. *Cerebral Cortex*, 20(1), 130–140.
- Sasaki, Y., & Watanabe, T. (2004). The primary visual cortex fills in color. *Proceedings of the National Academy of Sciences*, 101(52), 18251–18256.
- Schönemann, P. H. (1966). A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1), 1–10.
- Scotti, P. S., Banerjee, A., Goode, J., Shabalin, S., Nguyen, A., Cohen, E., Dempster, A. J., Verlinde, N., Yundler, E., Weisberg, D., Norman, K. A., & Abraham, T. M. (2023). Reconstructing the mind's eye: fMRI-to-image with contrastive learning and diffusion priors [Preprint]. *arXiv*. <https://arxiv.org/abs/2305.18274>

- Scotti, P. S., Tripathy, M., Villanueva, C. K. T., Kneeland, R., Chen, T., Narang, A., Santhirasegaran, C., Xu, J., Naselaris, T., Norman, K. A., et al. (2024). MindEye2: Shared-subject models enable fMRI-to-image with 1 hour of data [Preprint]. *arXiv*. <https://arxiv.org/abs/2403.11207>
- Seeliger, K., Güçlü, U., Ambrogioni, L., Güçlütürk, Y., & van Gerven, M. A. J. (2018). Generative adversarial networks for reconstructing natural images from brain activity. *NeuroImage*, 181, 775–785.
- Seghier, M. L., & Vuilleumier, P. (2006). Functional neuroimaging findings on the human perception of illusory contours. *Neuroscience & Biobehavioral Reviews*, 30(5), 595–612.
- Serences, J. T., Schwarzbach, J., Courtney, S. M., Golay, X., & Yantis, S. (2004). Control of object-based attention in human cortex. *Cerebral Cortex*, 14(12), 1346–1357.
- Sereno, M. I., Dale, A., Reppas, J., Kwong, K., Belliveau, J., Brady, T., Rosen, B., & Tootell, R. (1995). Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science*, 268(5212), 889–893.
- Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3), 411–426.
- Sharan, L., Rosenholtz, R., & Adelson, E. H. (2014). Accuracy and speed of material categorization in real-world images. *Journal of Vision*, 14(9), 12–12.
- Shen, G., Dwivedi, K., Majima, K., Horikawa, T., & Kamitani, Y. (2019a). End-to-end deep image reconstruction from human brain activity. *Frontiers in Computational Neuroscience*, 13, 432276.
- Shen, G., Horikawa, T., Majima, K., & Kamitani, Y. (2019b). Deep image reconstruction from human brain activity. *PLoS Computational Biology*, 15(1), e1006633.
- Shirakawa, K., Nagano, Y., Tanaka, M., Aoki, S. C., Majima, K., Muraki, Y., & Kamitani, Y. (2024). Spurious reconstruction from brain activity [Preprint]. *arXiv*. <https://arxiv.org/abs/2405.10078>
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition [Preprint]. *arXiv*. <https://arxiv.org/abs/1409.1556>
- Soon, C. S., Brass, M., Heinze, H.-J., & Haynes, J.-D. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*, 11(5), 543–545.
- Soriano, M., Spillmann, L., & Bach, M. (1996). The abutting grating illusion. *Vision Research*, 36(1), 109–116.
- Steinmetz, N. A., Aydin, C., Lebedeva, A., Okun, M., Pachitariu, M., Bauza, M., Beau, M., Bhagat, J., Böhm, C., Broux, M., Chen, S., Colonell, J., Gardner, R. J., Karsh, B., Kloosterman, F., Kostadinov, D., Mora-Lopez, C., O’Callaghan, J., Park, J., . . .

- Harris, T. D. (2021). Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. *Science*, 372(6539), eabf4588.
- Störmer, V. S., McDonald, J. J., & Hillyard, S. A. (2009). Cross-modal cueing of attention alters appearance and early cortical processing of visual stimuli. *Proceedings of the National Academy of Sciences*, 106(52), 22456–22461.
- Strotzer, M. (2009). One century of brain mapping using brodmann areas. *Clinical Neuroanatomy*, 19(3), 179.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2014). Going deeper with convolutions [Preprint]. *arXiv*. <https://arxiv.org/abs/1409.4842>
- Takagi, Y., & Nishimoto, S. (2023). High-resolution image reconstruction with latent diffusion models from human brain activity. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14453–14463.
- Takahashi, T., Sasabayashi, D., Yücel, M., Whittle, S., Lorenzetti, V., Walterfang, M., Suzuki, M., Pantelis, C., Malhi, G. S., & Allen, N. B. (2022). Different frequency of heschl's gyrus duplication patterns in neuropsychiatric disorders: An MRI study in bipolar and major depressive disorders. *Frontiers in Human Neuroscience*, 16, 917270.
- Talairach, J., & Tournoux, P. (1988). *Co-planar stereotaxic atlas of the human brain: 3-dimensional proportional system: An approach to cerebral imaging*. Georg Thieme Verlag.
- Tan, M., & Le, Q. V. (2020). EfficientNet: Rethinking model scaling for convolutional neural networks [Preprint]. *arXiv*. <https://arxiv.org/abs/1905.11946>
- Tang, J., & Huth, A. G. (2025). Semantic language decoding across participants and stimulus modalities. *Current Biology*, S0960982225000545.
- Tang, J., LeBel, A., Jain, S., & Huth, A. G. (2023). Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5), 858–866.
- Ungerleider, S. K., & G., L. (2000). Mechanisms of visual attention in the human cortex. *Annual Review of Neuroscience*, 23(1), 315–341.
- Van Essen, D. C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T. E. J., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S. W., Della Penna, S., Feinberg, D., Glasser, M. F., Harel, N., Heath, A. C., Larson-Prior, L., Marcus, D., Michalareas, G., Moeller, S., ... Yacoub, E. (2012). The Human Connectome Project: A data acquisition perspective. *NeuroImage*, 62(4), 2222–2231.
- Van Essen, D. C. (2005). A Population-Average, Landmark-and Surface-based (PALS) atlas of human cerebral cortex. *NeuroImage*, 28(3), 635–662.
- Van Essen, D. C. (2004). Surface-based approaches to spatial localization and registration in primate cerebral cortex. *NeuroImage*, 23, S97–S107.

- Van Horn, J. D. (2021). Bridging the brain and data sciences. *Big Data*, 9(3), 153–187.
- Van Horn, J. D., & Toga, A. W. (2014). Human neuroimaging as a “Big Data” science. *Brain Imaging and Behavior*, 8(2), 323–331.
- van Tuijl, H. F. J. M. (1975). A new visual illusion: Neonlike color spreading and complementary color induction between subjective contours. *Acta Psychologica*, 39(6), 441–IN1.
- 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, 378029.
- VanRullen, R., & Reddy, L. (2019). Reconstructing faces from fMRI patterns using deep generative neural networks. *Communications Biology*, 2(1), 1–10.
- Varin, D. (1971). Fenomini di contrasto e diffusione cromatica nell’organizzazione spaziale del campo percettivo. *Rivista di Psicologia*, 65, 101.
- Vogelstein, J. T., Perlman, E., Falk, B., Baden, A., Gray Roncal, W., Chandrashekhar, V., Collman, F., Seshamani, S., Patsolic, J. L., Lillaney, K., Kazhdan, M., Hider, R., Pryor, D., Matelsky, J., Gion, T., Manavalan, P., Wester, B., Chevillet, M., Trautman, E. T., ... Burns, R. (2018). A community-developed open-source computational ecosystem for big neuro data. *Nature Methods*, 15(11), 846–847.
- Wasserman, N., Beliy, R., Urbach, R., & Irani, M. (2024). Functional brain-to-brain transformation with no shared data [Preprint]. *arXiv*. <https://arxiv.org/abs/2404.11143>
- 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(2), 79–94.
- Weiner, K. S., Golarai, G., Caspers, J., Chuapoco, M. R., Mohlberg, H., Zilles, K., Amunts, K., & Grill-Spector, K. (2014). The mid-fusiform sulcus: A landmark identifying both cytoarchitectonic and functional divisions of human ventral temporal cortex. *NeuroImage*, 84, 453–465.
- Xu, T., Yousefnezhad, M., & Zhang, D. (2018). Gradient hyperalignment for multi-subject fmri data alignment. *PRICAI 2018: Trends in Artificial Intelligence*, 11012, 1058–1068.
- Yamada, K., Miyawaki, Y., & Kamitani, Y. (2015). Inter-subject neural code converter for visual image representation. *NeuroImage*, 113, 289–297.
- Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23), 8619–8624.
- Yantis, S., & Serences, J. T. (2003). Cortical mechanisms of space-based and object-based attentional control. *Current Opinion in Neurobiology*, 13(2), 187–193.

- Yousefnezhad, M., & Zhang, D. (2017). Deep hyperalignment [Preprint]. *arXiv*. <https://arxiv.org/abs/1710.03923>
- Zatorre, R. J., Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex: Music and speech. *Trends in Cognitive Sciences*, 6(1), 37–46.
- Zhou, Q., Du, C., Wang, S., & He, H. (2024). CLIP-MUSED: CLIP-guided multi-subject visual neural information semantic decoding [Preprint]. *arXiv*. <https://arxiv.org/abs/2402.08994>
- Zong, W., Obenhaus, H. A., Skytøen, E. R., Eneqvist, H., De Jong, N. L., Vale, R., Jorge, M. R., Moser, M.-B., & Moser, E. I. (2022). Large-scale two-photon calcium imaging in freely moving mice. *Cell*, 185(7), 1240–1256.e30.

Appendix A

Publications

The current work has yielded the following manuscripts and presentation:

A.1 Manuscript

- Wang, H., Ho, J. K., Cheng, F. L., Aoki, S. C., Muraki, Y., Tanaka, M., Park, J.-Y., & Kamitani, Y. (2025). Inter-individual and inter-site neural code conversion without shared stimuli. *Nature Computational Science*, 5, 534-546.

A.2 Manuscript in preparation

- Wang, H., Cheng, F. L., Aoki, S. C., Tanaka, M., Nagano, Y., & Kamitani, Y. (in preparation). Inter-individual subjective experience reconstruction.

A.3 Presentation

- Wang, H., Ho, J. K., Cheng, F., Aoki, S. C., & Kamitani, Y. (2023). Inter-individual neural code conversion without paired stimuli. Poster presentation for *2023 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 and inter-site neural code conversion and image reconstruction and evaluation: <https://github.com/KamitaniLab/InterSiteNeuralCodeConversion>
- Code for illusion reconstruction and evaluation: <https://github.com/KamitaniLab/IllusionReconstruction>

