Reconstructing Seen Images from fMRI Using Machine Learning

Reconstructing Seen Images from fMRI Using Machine Learning

David Lucha

The following thesis was submitted in partial fulfilment of the requirements for the Bachelor of Psychological Science (Honours) at the University of Queensland

PSYC4071: Individual Research Thesis

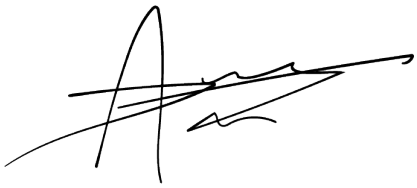
5th October 2022

Assessable Word Count: 7,990

Total Word Count: 15,512

Originality Statement

I, David Lucha, hereby certify that this thesis is my own original work. Any ideas, text or research referred to in this thesis which are not my own have been appropriately referenced. I also declare that this thesis has not been previously submitted for assessment.



David Lucha

October 5, 2022

Acknowledgements

First and foremost, I wish to express my gratitude to my supervisor, Alex Puckett. From the beginning of the year, your confidence in me to develop this project was incredibly motivating. Thank you for guiding me to create a project I am truly proud of. I am especially grateful for your flexibility, feedback, and for making sense in light of my frequent states of chaos and confusion. Next, I would like to acknowledge Clinton Condon. You were an unbelievably supportive point of contact. Your willingness to be involved, your kindness and understanding mean more to me than I can express. To Ashley York and Fernanda Ribeiro, I am honoured to have been able to work around you. I appreciate you both sharing your expertise.

I would also like to show my appreciation to Maria Podguzova for having established much of the Python implementation for the network and for answering several of my technical questions. And to Guy Gaziv for being willing to discuss and clarify the details of their research.

To Ryan and Esto, it was a privilege to go through this year alongside you; it was reassuring knowing that I was not alone on this wild journey. To Christine, thank you for being a dependable friend and for providing invaluable feedback on such short notice. To my other friends and family, you kept me grounded and were reliable sources of encouragement and often, much-needed distraction. Thank you, particularly to my mum, who has invested so much into my education and supported me beyond what I would have ever asked.

Most importantly, to my partner, Emily: none of this would have been possible without you. Your patience, kindness and understanding kept me afloat. This has not been an easy year, yet you loved and supported me without restraint, and I am eternally thankful for that. Thank you for listening to me for hours on end, for reading over my work multiple times, and for the many extra dog days you took on. I love you.

Abstract

Deep learning approaches have been increasingly successful in reconstructing seen images from brain data measured using functional magnetic resonance imaging (fMRI). This brain decoding research has been vital in understanding how visual information is encoded in low- and high-order visual areas. However, much of the current work is limited due to the lack of high-quality, large-scale neuroimaging datasets. The Natural Scenes Dataset (NSD) was developed to address this limitation and is the most comprehensive fMRI dataset of its kind, providing greater spatial resolution (i.e., voxel resolution) than previous datasets. However, there is currently no available research investigating the viability of the dataset for image reconstruction tasks. This thesis aimed to evaluate whether machine learning approaches could successfully use the NSD to generate high-fidelity image reconstructions from fMRI. Additionally, the thesis aimed to explore the role of voxel resolution on the quality of reconstructions and to use the NSD to explore the supplementary role of high-order visual areas in perception. Across three studies, a series of artificial neural networks were trained to reconstruct seen images from novel human brain data provided by the NSD. Reconstruction quality was assessed using three measures of identification accuracy computed using two leading image similarity metrics. In study one, it was found that neural networks consistently generated reconstructions of seen images with above-chance accuracy. In study two, finer voxel resolutions resulted in higher quality reconstructions than coarser voxel resolutions. Finally, study three revealed that, alone, low-order visual areas were most important for reconstructing seen images. However, high-order visual areas provided a supplementary effect, significantly boosting reconstruction quality when added to lower visual areas. These findings suggest that the NSD is a promising tool for brain decoding research and sheds new light on how high-order visual areas provide contextual and semantic information required for visual perception.

Table of Contents

List of Tables

List of Figures

Reconstructing Seen Images from fMRI Using Machine Learning

Mind-reading is often relegated to the spheres of science-fiction; however, there are now sophisticated methods to read out the contents of one's mind based solely on patterns of brain activity evoked by sensory stimuli. Specifically, advancements in machine learning have allowed researchers to reconstruct seen images from novel human brain activity measured using functional magnetic resonance imaging (fMRI; Rakhimberdina et al., 2021). This research has been used to understand the nature of visual perception (Chen et al., 2014; Gaziv et al., 2022), which has important implications for the study of imagery, dreams, and brain-computer interfaces (Du et al., 2022; Horikawa & Kamitani, 2017a). However, these machine learning approaches require considerably more brain imaging data than previously available (Allen et al., 2022). In response to this, the Natural Scenes Dataset (NSD) was developed to expand machine learning applications in neuroscience. To date, it is the most comprehensive dataset of its kind and leverages ultra-high field fMRI, which provides the ability to capture brain activity with greater spatial detail (Allen et al., 2022). Thus, the NSD has critical potential for advancing complex brain decoding research; however, given its recent development, the NSD has yet to be used for reconstructing seen images from fMRI. Therefore, this thesis aims to establish the viability of the NSD for natural image reconstruction, evaluate the effect of its high fidelity, and finally, use the NSD to better understand the nature of visual perception.

Imaging and Neuroscience

Modern neuroscience has been propelled by advancements in fMRI, which provides the ability to record brain activity with a high spatial resolution (Poldrack & Farah, 2015). fMRI indirectly measures brain activity via changes in the blood oxygen level-dependent (BOLD) signal (Ogawa et al., 1990). Briefly, the BOLD signal captures changes in the magnetic field of the brain induced by the metabolic demands associated with increased

neuronal activity (Heeger & Ress, 2002). For example, by measuring the BOLD signal during the presentation of a sensory stimulus, a pattern of activity can be extracted, representing activity changes elicited in response to the stimulus (Mazaika, 2009; Monti, 2011). The ability of fMRI to capture these changes in the magnetic field as a proxy for neuronal activity has played an essential role in establishing how sensory information is represented in cortical areas (Poldrack & Farah, 2015).

fMRI Reveals the Structure of Visual Cortex

fMRI has helped establish how the complex structure of the visual system integrates distinct visual features into a single stream of perception (Courtney & Ungerleider, 1997; Yacoub et al., 2008). For example, fMRI has been used to identify ocular dominance columns, which are small ensembles of neurons that share the same functional preference for encoding certain types of visual information (Yacoub et al., 2008). Furthermore, fMRI has been critical for mapping the retinotopic organization of early visual areas (Wandell & Winawer, 2011). Retinotopy refers to the notion that two adjacent points in the visual field – and, thus, the retina – are represented by adjacent points in the cortex (Grill-Spector & Malach, 2004). Because of this, the spatial relationships within a seen image are preserved in their associated cortical representations (Wandell & Winawer, 2011). By contrast, high-order visual areas are less defined in their retinotopy and are instead increasingly specialized for processing contextual information and complex stimuli such as faces (Gauthier et al., 2000; Kanwisher et al., 1997), places (Epstein & Kanwisher, 1998), and bodies (Peelen & Downing, 2005).

The structured organization of the visual cortex results in visual information being encoded in stable patterns of brain activity. For example, early research demonstrated the viability of accurately mapping simple, high-contrast visual stimuli to reliable patterns of activity in the brain (Miyawaki et al., 2008; Thirion et al., 2006). Significantly, researchers

have since developed methods to map more complex stimuli, such as images of natural scenes, to their response in visual areas (Kay et al., 2008; Poldrack & Farah, 2015). This ability to capture reliable patterns of brain activity has driven a pursuit to decode the sensory stimuli which gave rise to them. The idea being, if visual stimuli evoke stable brain responses, then these responses should reflect stimulus-related features that can subsequently be decoded (Chen et al., 2014; Du et al., 2019).

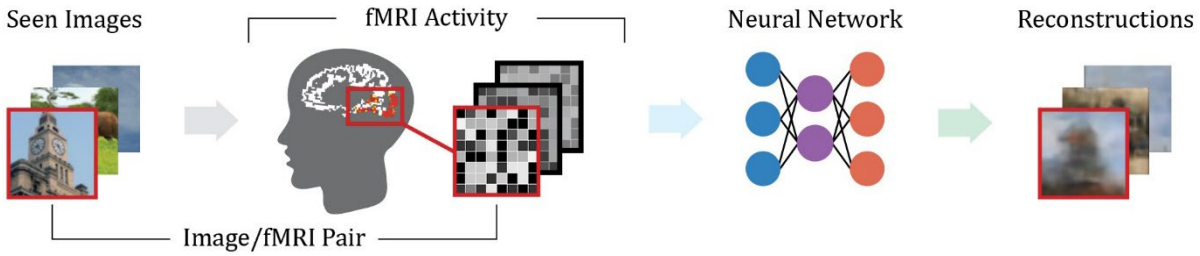
Early models were successful in decoding these patterns of brain activity to reconstruct low-level visual features such as oriented stimuli, patterns, and letters (Haynes & Rees, 2005; Kamitani & Tong, 2005, 2006; Thirion et al., 2006). However, these models were not viable for reconstructing complex natural scenes (i.e., natural image reconstruction). In contrast, advancements in deep neural networks have allowed researchers to leverage the rich, structured information encoded in the visual cortex to make significant progress toward high-fidelity natural image reconstruction (Beliy et al., 2019; Fang et al., 2020).

Machine Learning Approaches for Brain Decoding

Deep neural networks are a subset of machine learning techniques that involve multilayered neural network architectures (Zou et al., 2009). These biologically-inspired networks resemble the structure and function of the human brain and act as self-improving predictive algorithms (Cox & Dean, 2014; Shen et al., 2019b). This homology between the architecture of deep neural networks and representations in the visual cortex allows them to accurately capture the complex mappings between stimulus features and brain activity (see Figure 1; Cox & Dean, 2014; Wen et al., 2018).

Figure 1

Diagram of Natural Image Reconstruction Using Machine Learning

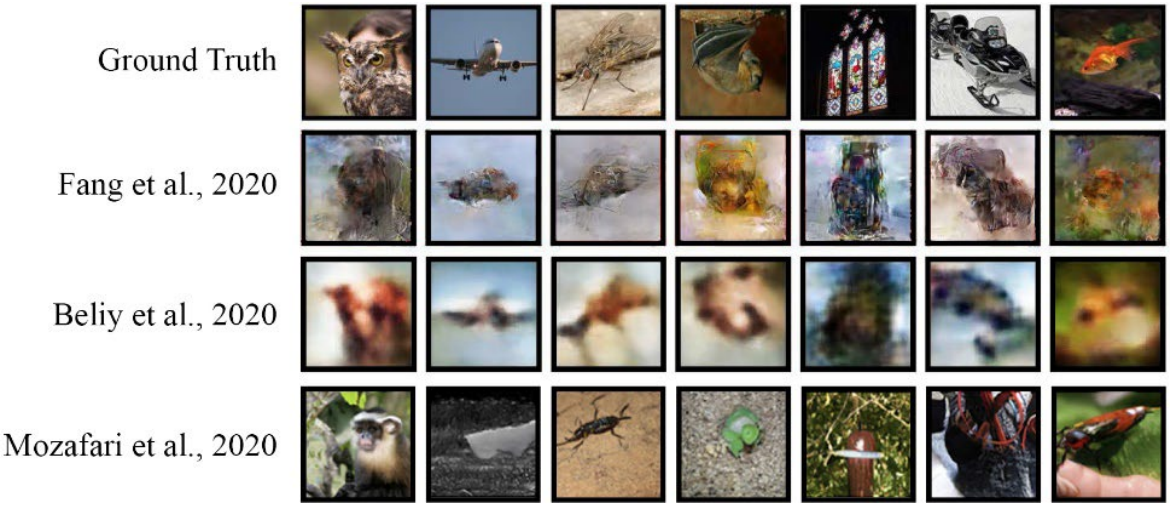


Note. Image reconstruction datasets consist of a large set of images and their evoked activity patterns measured using fMRI. Through multiple training iterations, the network learns the statistical associations between seen images and their associated patterns of brain activity (see Appendix A for a detailed explanation of how neural networks learn). By learning how component visual features are represented in the brain, these networks can subsequently decode novel brain activity to reconstruct seen images (Ren et al., 2021).

Neural networks are particularly suited to complex image reconstruction tasks due to their ability to extract minute statistical patterns from noisy data (such as fMRI), and they can effectively generalize to novel data with appropriate training (Lemm et al., 2011; Poldrack & Farah, 2015). As such, deep neural networks are the current state-of-the-art for fMRI-based natural image reconstruction and can generate reconstructions that maintain the colour and structure of seen images (see Figure 2; Rakhimberdina et al., 2021).

Figure 2

Natural Image Reconstruction Examples Using Machine Learning Approaches



Note. A comparison of machine learning approaches for natural image reconstruction (Beliy et al., 2019; Fang et al., 2020; Mozafari et al., 2020). The seen image is displayed on the top row (ground truth). Some methods (see Mozafari et al., 2020) prioritise the realism of reconstructions; however, these can often be distorted due to their overreliance on semantic information (Rakhimberdina et al., 2021). Despite producing blurry reconstructions, other deep learning approaches nevertheless decode discernible low-level visual features, including colour, contrast, shape, and texture (see reconstructions by Beliy et al., 2019). Images compiled by Rakhimberdina et al. (2021).

Natural Image Reconstruction for Neuroscience Inquiry

As shown, neural networks can produce impressive natural image reconstructions from novel brain data. Additionally, with the use of perturbation methods, these networks have become vital scientific tools to explore the role of specific visual areas in perception. In machine learning, perturbation methods are used to alter the features of a neural network's input to determine the relevance of the altered feature on the outcome of interest (Ras et al., 2022; Ribeiro et al., 2022). This approach can be used to examine the relative importance of different visual areas for reconstruction quality by training networks on data from different

regions of interest (ROIs). For example, Ren et al. (2021) found that relative to high-order visual areas, networks trained on low-order visual areas (e.g., V1-V3) generated reconstructions that better maintained the visual structure of the original image. Similarly, using different datasets and neural network architectures, other researchers have found similar effects whereby networks trained on lower visual areas reliably generate higher fidelity reconstructions compared to higher visual areas (Gaziv et al., 2022; Han et al., 2019).

These findings suggest that low-order visual areas contain the most valuable information for image reconstruction tasks. Importantly, these findings are consistent with the functional specialization of the visual cortex. Namely, that early visual areas specialize in processing basic visual features such as orientation, edge detection and contrast, which are essential in generating recognizable reconstructions (Brouwer & Heeger, 2009). However, these previous studies do not seem to consider that higher visual areas largely encode contextual and semantic information (Courtney & Ungerleider, 1997; Huth et al., 2012). Therefore, directly comparing reconstructions from low and high visual areas may not be appropriate because semantic information alone might not translate to high-quality reconstructions of low-level features. In other words, high-order areas might only benefit reconstruction quality when combined with lower visual areas. To this point, some evidence suggests that high-order areas provide semantic information, which can be used to improve reconstructions relative to early visual areas alone (Gaziv et al., 2022). However, this work did not quantify these differences, nor did they control for simply providing more brain data to the network. Hence, it is difficult to conclude whether high-order areas provided semantic information to the network or if these benefits were driven by increased model complexity.

Natural Scenes Dataset

Thus far, machine learning approaches have led to significant progress in decoding brain activity and have offered new tools to investigate visual perception (Shen et al., 2019b),

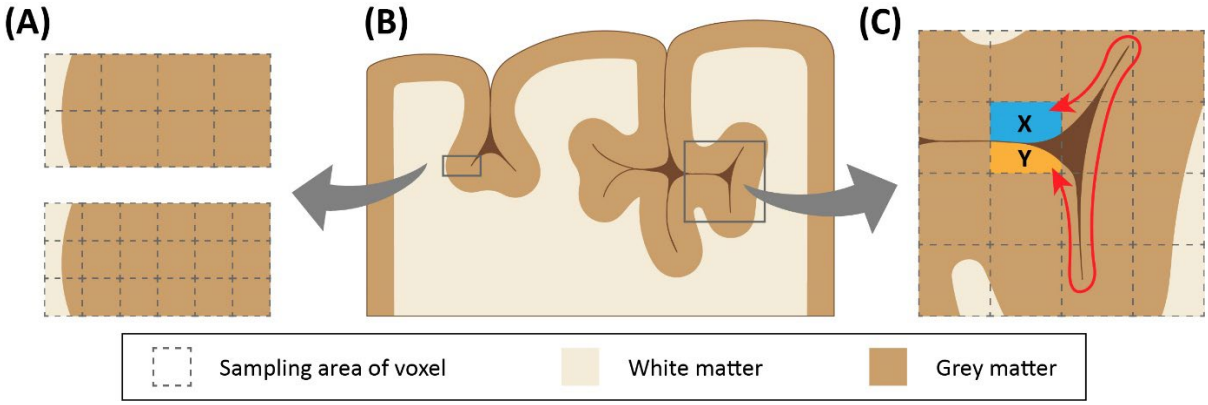
including the study of imagery and dreams (Horikawa & Kamitani, 2017b). However, this progress has been stifled by the lack of high-quality imaging data (Rakhimberdina et al., 2021). More specifically, deep neural networks for natural scene processing require tens of thousands of image samples for sufficient training (Allen et al., 2022). However, the most widely used datasets for natural image reconstruction only contain between one and five thousand image/fMRI pairs (Chang et al., 2019; Horikawa & Kamitani, 2017a). As such, many have cited the lack of paired data as one of the significant limitations of the current state of research (Du et al., 2018; Gaziv et al., 2022). Hence, the NSD was developed in response to this growing need for neuroimaging data and provides measurements of high-resolution fMRI responses to over 70,000 natural scene images across eight participants (Allen et al., 2022). Therefore, the unprecedented scale of the dataset is posited to expand the potential of machine learning applications in neuroscience (Allen et al., 2022).

Beyond the sheer scale of the dataset, the NSD also leveraged ultra-high field fMRI to provide data at finer voxel resolutions than previous image reconstruction datasets (Allen et al., 2022). Voxel resolution defines the smallest volume of sampled brain tissue using fMRI and relates to how much spatial detail is captured by a single scan (Kriegeskorte & Bandettini, 2007). The Generic Object Decoding dataset, commonly used in image reconstruction research, was collected at a voxel resolution of 3 mm (Horikawa & Kamitani, 2017a). At this resolution, fMRI voxels capture activity that span the entire cortical thickness and cover multiple cortical columns (Goense et al., 2016). This spatial limitation reduces an individual voxel's sensitivity to information encoded at smaller scales and increases partial volume effects (see Figure 3; Yacoub et al., 2008). This refers to when an individual voxel samples from an area containing multiple tissue types (e.g., mixing the signal from grey and white matter; Gardumi et al., 2016). By contrast, the NSD was collected with a finer voxel resolution of 1.8 mm, which reduces these partial volume effects (Allen et al., 2022;

Viessmann & Polimeni, 2021). Additionally, the NSD's use of ultra-high field imaging enhances sensitivity to BOLD signal and improves statistical power (Gardumi et al., 2016; Sengupta et al., 2017), which allows for detection of smaller effects in smaller anatomical regions (Cai et al., 2021; Torrisi et al., 2018).

Figure 3

The Effect of Voxel Resolution on Partial Volume Effects



Note. (A) Demonstrates the difference in partial volume effects as a function of voxel resolution. With a smaller voxel resolution, voxels are less likely to sample two issue types (e.g., white and grey matter). (B) Shows an example cross-section of cortical tissue. (C) Large voxel sizes also increase the likelihood of partial volume effects whereby a single voxel samples grey matter from two opposing sides of a sulcus (Gardumi et al., 2016). This can be problematic because points X and Y are contiguous using larger voxels. However, the red line shows the actual cortical distance between these two points, suggesting that mixing these signals misrepresents the underlying activity in this area.

Limitations of High-Resolution fMRI

The NSD has already been used to successfully predict brain activity in response to novel visual stimuli with state-of-the-art accuracy (Gu et al., 2022; Khosla & Wehbe, 2022). This research suggests that the technological improvements afforded by the dataset are of practical significance for encoding models; however, there are some inherent challenges in

using ultra-high field fMRI. For example, smaller voxel resolutions can lead to issues with motion and distortion and can therefore increase noise at the level of the voxel (Goense et al., 2016). Moreover, evidence suggests that improving voxel resolution may not always lead to increased decoding performance. For example, Gardumi et al. (2016) attempted to decode speech content (i.e., vowels) and speaker identity from ultra-high field fMRI. However, they found that speech content decoding, but not speaker identity, benefited from smaller voxel resolutions (Gardumi et al., 2016). This evidence suggests that the benefits of ultra-high field imaging may be task-dependent. That is, the suitability of the NSD for encoding models (i.e., stimulus-to-activation prediction) may not necessarily transfer to meaningful improvements in brain decoding tasks (i.e., activation-to-stimulus) such as natural image reconstruction.

The Current Study

The NSD addresses the scarcity of paired data for deep learning applications in neuroscience and may be central to advancing brain decoding research. However, there is currently no available research evaluating the suitability of the NSD for brain decoding tasks. Therefore, there is no research establishing the dataset as a viable tool to assess the role of visual areas in perception. Furthermore, due to the technical challenges associated with ultra-high field fMRI, it is unclear whether the NSD's finer voxel resolution will lead to consistent benefits for natural image reconstruction.

Therefore, the overarching aim of this thesis is to evaluate the viability of the NSD for natural image reconstruction. This will be achieved using three studies, each seeking to address specific objectives as outlined below. In each study, several neural networks will be trained using a conceptual replication of the state-of-the-art architecture proposed by Ren et al. (2021). The networks will be trained on altered variations of the fMRI data provided by the NSD to reconstruct seen images from novel brain data. In line with previous work,

reconstruction quality will be assessed qualitatively and with identification accuracy tasks computed with two leading image similarity metrics (Rakhimberdina et al., 2021).

Study 1

To address the primary aim, in Study 1, the identification accuracy of reconstructions from networks trained on the NSD data will be assessed (for each of the eight participants).

Hypothesis 1. Networks trained on the NSD will produce high-fidelity image reconstructions, performing above chance on identification tasks.

Study 2

The second aim is to investigate the extent to which the NSD's finer voxel resolution results in higher quality reconstructions. To achieve this, several networks will be trained either with data at 1.8 mm or 3 mm voxel resolution.

Hypothesis 2. Relative to coarser resolutions, networks trained with finer voxel resolutions will produce higher quality reconstructions, also reflected in higher identification accuracy.

Study 3

The final aim of this thesis is to investigate the role of low- and high-order areas in visual processing. To address this, the quality of reconstructions generated by networks trained with data either from high or low visual areas will be compared.

Hypothesis 3a. In line with previous research, networks trained on data from low-order visual areas will produce better reconstructions, with improved identification accuracy, than those trained on data from higher visual areas (Gaziv et al., 2022; Ren et al., 2021).

Most importantly, the present thesis seeks to evaluate the extent to which combining high and low visual areas results in an additive benefit to reconstruction quality relative to low visual areas alone, or low visual areas combined with data randomly sampled from non-visual areas.

Hypothesis 3b. Reconstructions from low and high visual areas combined will produce improved reconstructions (and therefore achieve higher identification accuracy) compared to low visual areas and low visual areas combined with non-visual areas.

Methods

Natural Scenes Dataset (NSD)

During the main NSD experiment, eight participants (6 female; 19-32 years) viewed 10,000 natural scene images, three times across 40 scanning sessions. fMRI measurements were collected at 7T with a spatial resolution of 1.8 mm (isotropic). Of the 10,000 images, 9,000 were unique to each individual, and 1,000 images, chosen from the entire image set at random, were presented to all eight participants. More information regarding the NSD can be found at http://naturalscenesdataset.org (Allen et al., 2022).

Data Selection

The NSD provides data in various formats, from raw through various stages of post-processing. Here, the 1.8 mm, denoised preparation of the single-trial beta weights (betas_fithrf_GLMdenoise_RR) were chosen. These beta weights reflect the degree of activation within each voxel as measured by percent BOLD signal change (in response to viewing natural scenes). The denoised data were selected due to having undergone an advanced form of post-processing (i.e., GLMdenoise with ridge regression), previously shown to improve the quality of beta estimates (Allen et al., 2022). Betas were downloaded in subject-native volume space to ensure the anatomical accuracy of the selected ROIs (Hutchison et al., 2014).

ROI Selection

Four collections of manually drawn ROIs were selected, which had been delineated using standard population receptive field (pRF; Benson et al., 2018) and functional localizer (fLoc; Stigliani et al., 2015) experiments. In line with similar work, voxels from areas V1 to

V4 and face- and place-selective regions were selected (Horikawa & Kamitani, 2017a). Additionally, a collection of well-established, body-selective ROIs were included. These have been shown to selectively respond to images of bodies and body parts (Downing et al., 2001; Peelen & Downing, 2005) and hence, should offer valuable encoding information to the network.

fMRI Data Processing

Handling Overlapping Voxels

The four ROI collections were combined per participant to make a combined mask. In cases where voxels had multiple ROI designations, they were assigned to the lowest order visual area (see Table 1 for more details).

Table 1

ROIs Included in Network Training and Evaluation

NSD ROI Collection	Included Areas	Description	ROI Delineation
prf-visualrois	V1v	Ventral subdivision of V1	pRF experiment
	V1d	Dorsal subdivision of V1	
	V2v	Ventral subdivision of V2	
	V2d	Dorsal subdivision of V2	
	V3v	Ventral subdivision of V3	
	V3d	Dorsal subdivision of V3	
	V4	Area V4	
floc-faces	OFA	Occipital Face Area	fLoc experiment
	FFA-1	Fusiform Face Area (subdivision 1)	
	FFA-2	Fusiform Face Area (subdivision 2)	
	mTL-faces	Mid-temporal lobe (face-selective)	
floc-places	OPA	Occipital Place Area	fLoc experiment
	PPA	Parahippocampal Place Area	
	RSC	Retrosplenial Cortex	
floc-bodies	EBA	Extrastriate Body Area	fLoc experiment
	FBA-1	Fusiform Body Area (subdivision 1)	
	FBA-2	Fusiform Body Area (subdivision 2)	
	mTL-bodies	Mid-temporal lobe (body-selective)	

Note. This table outlines all the ROIs used in the full visual cortex collection. Some participants are missing subregions of floc-faces, floc-places, and floc-bodies in either or both hemispheres (see Appendix B for a breakdown of ROI-specific voxel counts). In instances where voxels had labels belonging to two or more different ROI collections, the voxel was rescored to the lowest order area (i.e., the ROI located in the top-most position in the table above). For example, if a voxel had both OFA and EBA labels, this would be rescored to the sub-region belonging to the collection closest to the top. In this example, floc-faces is located higher than floc-bodies; therefore, this voxel would be rescored as OFA. Most overlapping voxels (between 82.1% and 99.4%) were either in V4 (not included in Study 3) or found in areas that would eventually be combined into a collection defined as high visual cortex (HVC, consisting of face-, place-, and body-selective areas).

Beta Processing

Voxel values from the selected ROIs were extracted from all sessions, combined, and then standardized for each participant. This was done in line with best practices for gradient-descent-based machine learning and ensuring the model's timely convergence (Bhandari, 2022).

Train/Test Split

To allow for simple qualitative comparisons between participants and training conditions, the test split was defined as the maximum number of shared images seen at least once by all participants ($N = 872$). For the test split, repetitions were excluded to remove any potential effect of repeated presentations. All remaining images and their repetitions were used for the training set. Due to some participants missing several of the intended 40 scan sessions, there is some variation in the total number of trials in each training set (see Appendix C for details).

Data Resampling

The process outlined above was carried out for the original 1.8 mm preparation of the data and to resampled 3 mm data. In the latter case, each session of betas and the individual ROI collections were downsampled (using nearest neighbour interpolation) to 3 mm voxel resolution before processing.

Overview of Neural Network

The overarching goal of the neural network is to effectively map cognitive data (i.e., fMRI data) to latent visual features from which reconstructions can be decoded – permitting novel, seen images to be predicted from human brain activity. Latent visual features refer to the hidden output features of the network's dual encoders (i.e., Visual and Cognitive encoders). In essence, latent features represent the component visual features of the original image.

Reconstructions are achieved by first training a Visual Encoder, which transforms high-dimensional natural scene images into low-dimensional latent visual features that a Decoder then uses to reconstruct the original image. The trained Visual Encoder then acts as a teacher network, which guides the subsequent training of a Cognitive Encoder. As such, the Cognitive Encoder attempts to map the cognitive data (measured here using fMRI) to the same distribution of latent visual features encoded by the Visual Encoder.

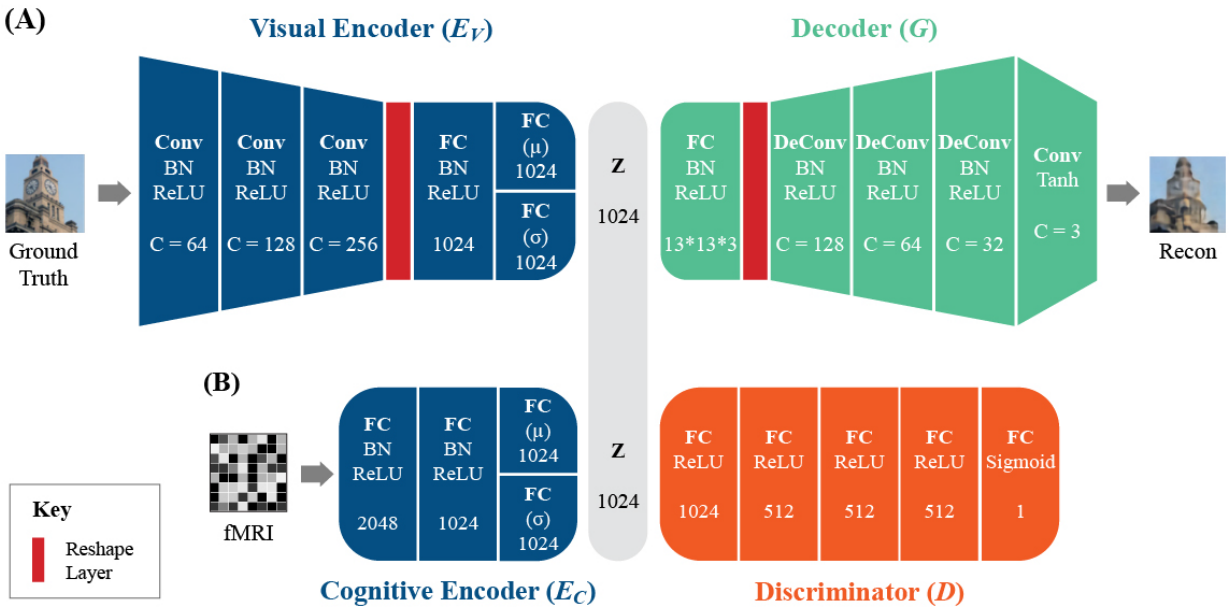
Architecture

The proposed network largely follows the architecture developed by Ren et al. (2021). However, the Variational Autoencoder (VAE) component of Ren's network is replaced with a Wasserstein Autoencoder (WAE). WAEs are known to have similarly stable training processes and share the same fundamental encoder-decoder architecture as VAEs. Still, they can provide better quality reconstructions due to their use of an adversarial Discriminator, which regularizes the distribution of latent features (Tolstikhin et al., 2017).

The sections that follow briefly describe the main components of the proposed neural network architecture (see Figure 4 for technical details).

Figure 4

Overview of the Proposed Architecture



Note. (A) The Visual Encoder consists of three convolutional layers (Conv), three fully connected (FC) layers, and a reshape layer (represented by the red bar). The parallel FC layers output the mean (μ) and log variance (σ) of each batch to the 1024-dimensional latent space Z. The Decoder consists of a single FC layer, reshape layer, three deconvolutional layers (DeConv), and a single convolutional layer (paired with a Tanh activation function) which outputs the final, three-channel (i.e., RGB coloured) image reconstruction. (B) The Cognitive Encoder has four FC layers, two of which act in parallel to output the latent features. The Discriminator consists of five FC layers, with the final layer including a Sigmoid activation function to discriminate between latent distributions. BN, ReLU, C denotes batch normalization, rectified linear unit, and channels, respectively. The numbers beneath FC layers reflect the number of output features.

Discriminator (D)

The Discriminator D plays a vital role in regularizing the encoded latent space (i.e., the output of either encoder). Its function is to discriminate whether a given data point in the

latent space was sampled from a 'real' normally distributed prior P_Z or from the 'fake' encoded space Q_Z (Tolstikhin et al., 2017). In doing so, the Discriminator learns to classify real latent points from fake latent points by maximising the following adversarial loss function:

$$\mathcal{L}_{gan} = \log\big(D(Z_i)\big) + \log\Big(1 - D\big(\tilde{Z}_i\big)\Big), \quad (1)$$

where $\big(D(Z_i)\big)$ is the probability of D correctly classifying a data point sampled from P_Z (Z_i) as real. Conversely, $(1 - D\big(\tilde{Z}_i\big))$ is the probability of D correctly classifying a data point sampled from Q_Z $(\tilde{Z}_i)$ as fake.

Wasserstein Autoencoder

The autoencoder portion of the network refers to the combination of the Decoder G with either Encoder (E_C/E_V). Each Encoder has two main objectives. First, they must generate latent features that provide the Decoder with rich information to generate reconstructions. This is achieved by minimising the following reconstruction error term:

$$\mathcal{L}_{recon} = \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \tilde{Y}_i\big)^2, \quad (2)$$

where $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \tilde{Y}_i)^2$ reflects the mean squared error (MSE) of the reconstruction $\tilde{Y}$ and real image Y. This encourages the encoder-decoder to produce reconstructions more similar to the ground truth image.

The second objective of the encoders is to match the distribution of latent features Q_Z to a prior distribution P_Z. To achieve this, the encoders take advantage of the Discriminator's ability to classify real and fake latent data points. Specifically, the encoders are encouraged to generate latent features which are more likely to be misclassified as having been sampled from the prior distribution. As such, both encoders maximise the following non-saturating loss function:

$$\mathcal{L}_{penalty} = \log\left(D\left(\tilde{Z}_i\right)\right), \quad (3)$$

where $(D\left(\tilde{Z}_i\right))$ is the probability of the Discriminator incorrectly classifying a data point sampled from Q_Z $(\tilde{Z}_i)$ as real. $\mathcal{L}_{penalty}$ works against the $\mathcal{L}_{gan}$ objective, thereby reducing the discrepancy between the encoded and real (or prior) distribution. Notably, the parameters of each encoder-decoder pair are updated jointly.

Training Stages

Training of the network is conducted in three distinct stages, which facilitates the optimization of the unique training objectives at each stage (Ren et al., 2021).

Stage 1

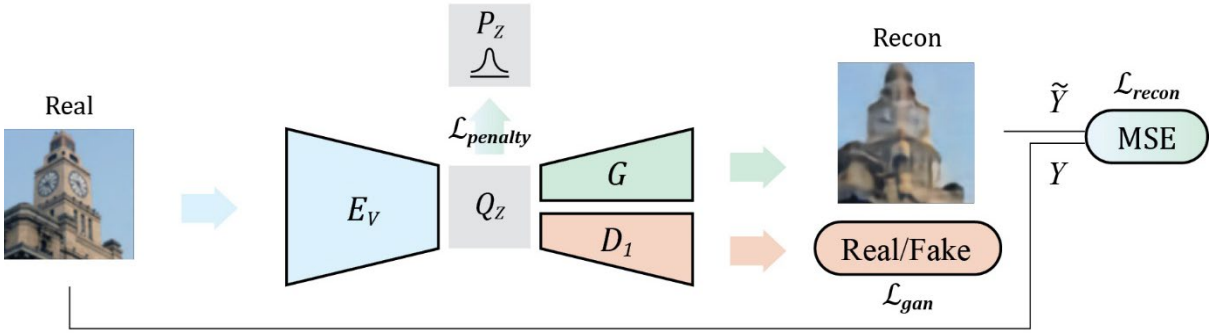
In Stage 1, the Visual Encoder E_V transforms natural scene images into latent visual features that a Decoder G then uses to reconstruct the original image (see Figure 5). In this stage, the parameters of E_V, G, and Discriminator D are updated with the following triple criterion:

$$\mathcal{L} = \mathcal{L}_{penalty} + \mathcal{L}_{recon} + \mathcal{L}_{gan}. \quad (4)$$

Importantly, E_V is trained exclusively on visual stimuli, which provides the encoder with rich knowledge relating to the component visual features of the images (Ren et al., 2021). Given this capability, E_V is used to guide the subsequent training of a Cognitive Encoder E_C. This transfer of knowledge improves the encoding of latent visual features, which results in higher fidelity reconstructions (Ren et al., 2021).

Figure 5

Overview of Stage 1 Network Training



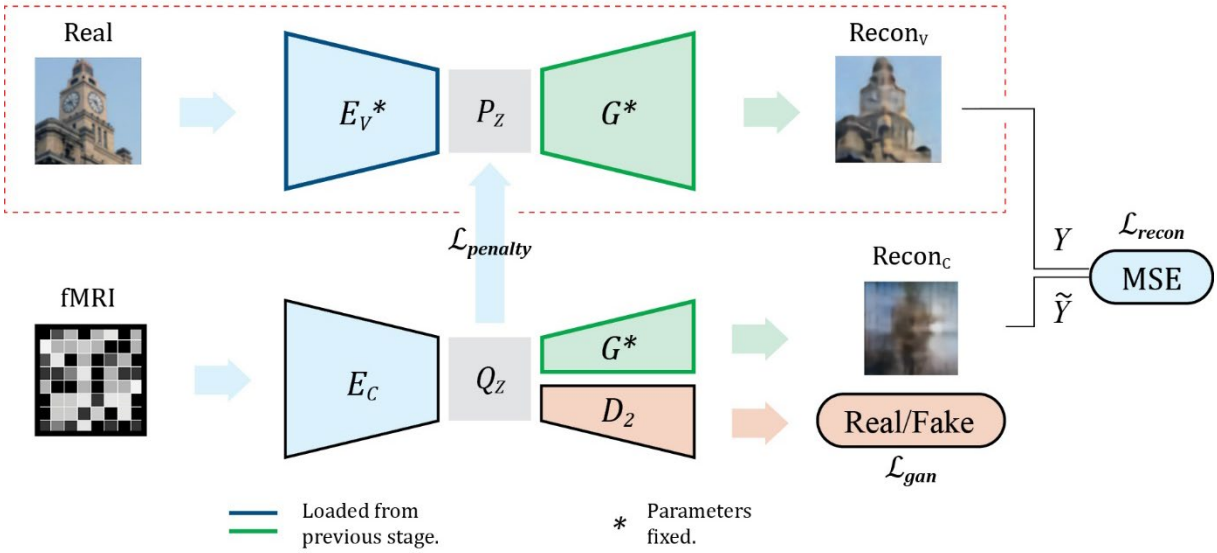
Note. Stage 1 trains a Wasserstein auto-encoder/generative adversarial network (WAE-GAN) to achieve high-fidelity image-to-image reconstructions. Images are fed into a Visual Encoder E_V, which outputs a distribution Q_Z of latent visual features, from which the Decoder G can generate reconstructions. The ground truth and reconstruction are then used to calculate $\mathcal{L}_{recon}$, which is used to update the parameters of E_V and G. Meanwhile, Discriminator D is trained to correctly classify whether latent features were sampled from Q_Z or from the normally distributed prior P_Z (i.e., real to the D). Specifically, the Discriminator assigns real or fake labels to latent visual features sampled from the prior and generated by the encoder. This output is then used to calculate $\mathcal{L}_{gan}$, which reflects the Discriminator's prediction error.

Stage 2

In Stage 2, the trained Visual Encoder E_V acts as a teacher network, which encourages the Cognitive Encoder E_C to mimic its feature output (see Figure 6). The parameters of E_C and the Discriminator D are updated using the triple criterion in Equation 4 above.

Figure 6

Overview of Stage 2 Network Training



Note. Stage 2 trains the Cognitive Encoder E_C, which transforms the cognitive data (fMRI) into latent visual features Q_Z. Here, the same Decoder G, trained during Stage 1, is used with its parameters fixed to generate cognitively derived reconstructions $\tilde{Y}$ (Recon$_C$). In Stage 2, the complete model from Stage 1 works in parallel to generate the 'real' distribution of latent visual features P_Z from the previously trained Visual Encoder E_V. $\mathcal{L}_{penalty}$ forces E_C to output latent features that are similar to those encoded by E_V. A new Discriminator D_2 is trained to correctly classify whether latent features were sampled from Q_Z or from P_Z (i.e., real to the D) by generating labels as described in Figure 5. Finally, the network from Stage 1 is also used to generate visually derived reconstructions (Recon$_V$), which act as the real stimulus Y for $\mathcal{L}_{recon}$, instead of the ground truth image as in Stage 1. Here, $\mathcal{L}_{recon}$ is used to update the parameters of E_C, further encouraging it to output latent visual features that mimic those encoded by E_V (Ren et al., 2021).

Stage 3

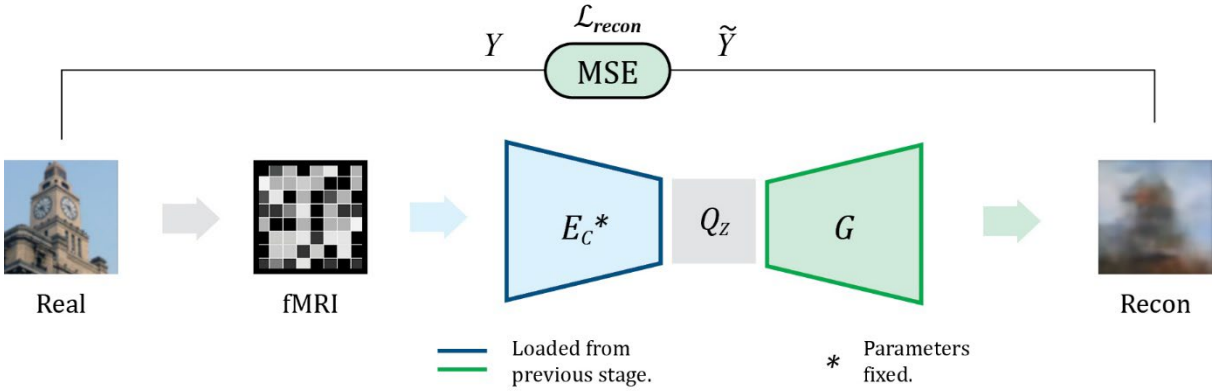
The parameters of Decoder G were fixed in Stage 2 and, thus, have not been trained using cognitive data. As such, Stage 3 optimizes the Decoder's ability to reconstruct seen images based on the features encoded by Cognitive Encoder E_C (see Figure 7). Here, the

parameters of E_C are fixed, which means the Discriminator is no longer required to regularize the latent space. Therefore, only the parameters of G are updated with the following loss function:

$$\mathcal{L} = \mathcal{L}_{recon}. \tag{5}$$

Figure 7

Overview of Stage 3 Network Training



Note. In Stage 3, Cognitive Encoder E_C transforms the cognitive data (fMRI) into latent visual features Q_Z, from which the Decoder G generates reconstructions. Stage 3 aims to optimize the Decoder's ability to reconstruct images from a given set of latent visual features. To achieve this, both E_C and G are loaded from Stage 2, but the parameters of E_C are now fixed. Here, $\mathcal{L}_{recon}$ follows Stage 1 in that it measures the MSE of the cognitively derived reconstruction $\tilde{Y}$ and ground truth image Y.

Implementation Details

Data processing, model testing, and evaluation was conducted on a local machine with an NVIDIA 1080Ti GPU. The NSD data was processed using AFNI (Cox, 1996; Cox & Hyde, 1997), and Python 3.6.13. Final models were implemented using Python 3.8.13 and PyTorch 1.12.0 (Paszke et al., 2017) and were trained on high-performance computing cluster nodes consisting of NVIDIA Tesla V100 GPUs.

Hyperparameters

In all stages, Adam optimizers (i.e., an established algorithm for gradient descent; Alabdullatef, 2020) were used with $\beta = 0.5$. In Stage 1, the initial learning rate for the Encoder and Decoder was 1×10^{-4}, and Discriminator at 5×10^{-5}. In Stages 2 and 3, the learning rate for the Encoder and Decoder was increased to 1×10^{-3} and the Discriminator at 5×10^{-4}. The learning rate was halved every 30 epochs. These parameters were used in line with a previous implementation of a three-stage WAE-GAN for visual reconstructions (Podguzova, 2021).

Stage 1 was trained for 100 epochs with a batch size of 100. However, to control for the variable training set sizes between participants, Stages 2 and 3 were trained for 30,000 and 20,000 iterations (or mini-batches), respectively.

Code Availability

All Python code and scripts for processing data in AFNI are openly available on GitHub (https://github.com/DavidLucha/ImageReconstructionNSD_Thesis).

Procedure

The same pre-trained network from Stage 1 was used for all studies to ensure that the same Visual Encoder guided each Cognitive Encoder. This network was trained on all 72,000 unique images used in the NSD experiment.

In each of the following studies, networks were trained separately on each participant's full set of training data, including all visual cortex voxels (i.e., four combined ROIs) unless otherwise stated.

Study 1

In Study 1, Stages 2 and 3 of the networks were trained on the 1.8 mm preparation of the data. This was to evaluate the viability of the NSD for image reconstruction by testing whether reconstructions for each participant performed above chance.

Study 2

Study 2 investigated the effect of voxel resolution on reconstruction quality and mirrored the process for Study 1. Here, networks were trained on the down-sampled 3 mm variations of the NSD betas to simulate the coarser voxel resolutions of previously used datasets (Horikawa & Kamitani, 2017a).

Study 3

Study 3 examined the contribution of high visual cortex (HVC) activity to reconstruction quality. To achieve this, each participant had four networks trained on different ROIs: V1-V3, HVC, V1-V3 and HVC combined (HVC+), and V1-V3 plus a subset of voxels randomly sampled from non-visual areas (Rand+). Non-visual areas were defined as the whole brain minus all ROIs delineated using the pRF and fLoc experiments. These random voxels were only extracted for the 1.8 mm data and matched the HVC voxel counts for each participant.

Analysis

Evaluation of Reconstruction Quality

Two leading image similarity metrics were used to measure reconstruction quality: Pearson's Correlation Coefficient (PCC) and the Learned Perceptual Image Patch Similarity (LPIPS) Metric.

PCC

PCC is the most widely used metric to evaluate image similarity in the image reconstruction literature (Rakhimberdina et al., 2021). The following equation defines the computation for pixel-wise PCC between pixels in the reconstructed and original image (Rakhimberdina et al., 2021; Shen et al., 2019a):

$$PCC(Y, \tilde{Y}) = \frac{\sum(Y - \mu_Y)(\tilde{Y} - \mu_{\tilde{Y}})}{\sqrt{\sum(Y - \mu_Y)^2 \sum(\tilde{Y} - \mu_{\tilde{Y}})^2}}, \quad (6)$$

where μ_Y and $\mu_{\tilde{Y}}$ reflect the mean values of the ground truth Y and reconstruction $\tilde{Y}$, respectively. PCC is a pixel-wise computation that measures low-level features, where higher PCC values indicate greater average image similarity.

LPIPS

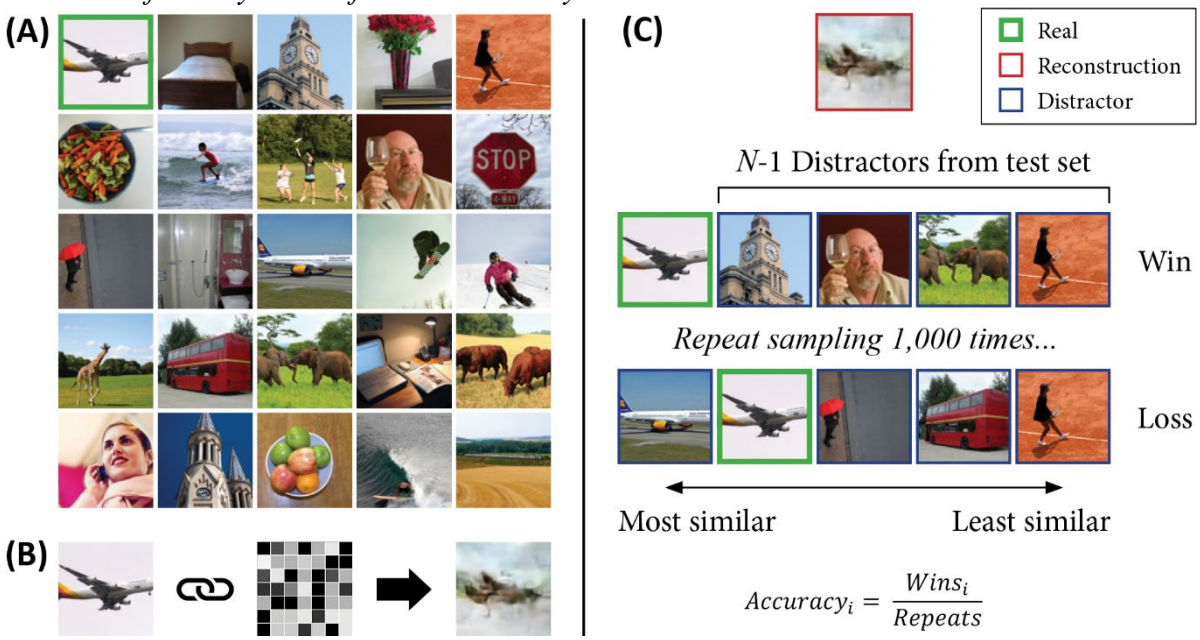
Despite the widespread use of low-level similarity metrics like PCC, these metrics often perform poorly compared to human judgements of image similarity (Zhang et al., 2018). The LPIPS metric has been proposed to address this issue. LPIPS is a distance-based metric that uses deep features from pre-trained neural networks (here, AlexNet version 0.1.4) to compute a measure of high-level perceptual similarity more in line with human judgements, where lower scores reflect greater image similarity (Zhang et al., 2018).

N-Way Identification Accuracy Task

Identification accuracy tasks are widely used as an objective measure of reconstruction quality due to the inherent limitations of comparing the absolute values of similarity metrics between different images (Beliy et al., 2019). In line with previous work, an n-way comparison task was adopted using both PCC (Fang et al., 2020; Ren et al., 2021) and LPIPS metrics (Gaziv et al., 2022). For each metric, three different n-way identification tasks were used with increasing task difficulty (N = 2, 5, 10; see Figure 8 for example).

Figure 8

Overview of N-Way Indentification Accuracy Task



Note. (A) Shows an example of the test set of images. The full test set includes 872 images. (B) For each image in the test set, there is an associated pattern of fMRI activity, which the network uses to generate a reconstruction. Significantly, these images and their associated patterns of activity have not been exposed to the network during training. (C) An example of the 5-way identification accuracy task. For each reconstruction in the test set, the relevant similarity metric is computed between it and its corresponding real image, and to *N*-1 distractor images randomly sampled from the test set. The resulting similarity scores are then used to rank the *N* candidate images from most similar to least similar. Winning trials are defined as instances where the real image is most similar to the reconstruction. Any cases where at least one randomly sampled distractor outscores the real image is a loss. In line with previous work, this random sampling of *N*-1 images is repeated 1,000 times per reconstruction (G. Gaziv, personal communication, September 10, 2022). Accuracy for each reconstructed image is defined as the proportion of winning comparisons divided by the total number of comparisons (i.e., repeats). This task computes an accuracy score for each reconstruction, and the overall identification accuracy for any comparison is the average of all reconstruction accuracy scores in a given test set.

Statistics

Permutation Test

For Study 1, a permutation test was used similar to that reported by Cowen et al. (2014) to test whether networks performed above chance level. Here, the test was adapted to follow the three n-way identification tasks. Specifically, instead of comparing the similarity between the reconstruction and real image with *N*-1 distractor images (as described in Figure 8), the real image was randomly switched with another image in the test. In effect, the real labels corresponding to each reconstruction were randomly permuted, such that any of the 872 test images had an equal chance to be selected as the real to the reconstruction. This was repeated for 100,000 iterations to create a null distribution of mean accuracy scores for each participant, on each n-way task, per metric. The probability of observing each mean accuracy was calculated as the proportion of permutations where mean accuracy was equal to or greater than the observed.

Non-Parametric Tests

In line with current reconstruction research, two-sided Wilcoxon signed-rank tests were used in Studies 2 and 3 to evaluate differences in identification accuracy between conditions (Gaziv et al., 2022; Ren et al., 2021). For comparisons of two or more conditions, a Friedman test (i.e., a non-parametric equivalent of a repeated-measures ANOVA) was used, and significant effects were followed up using Bonferroni-corrected Wilcoxon signed-rank tests.

Bootstrapped Confidence Intervals

Finally, in line with previous work, bootstrapping was used to compute the 95% confidence intervals of mean identification accuracy (Gaziv et al., 2022). Specifically, 10,000 samples of bootstrapped mean accuracies with replacement were generated from any given

set of reconstruction accuracy scores. The resulting distribution of mean accuracy scores was used to compute the confidence intervals.

Contributions

My supervisor, Dr Alex Puckett, suggested an initial avenue for research, and I then proposed specific research questions, which were refined jointly through ongoing discussions. Alex provided a brief conceptual overview of AFNI (i.e., neuroimaging software), and one of his PhD students, Clinton Condon, showed me how to access and navigate the high-performance computer clusters. After this, all the technical knowledge required to carry out the project was acquired independently. This includes learning technical skills such as Python and PyTorch, AFNI, Linux, GitHub, and AWS S3. Any technical issues were solved independently.

Much of the underlying framework of my final network architecture was adapted from existing implementations found on GitHub (and these specific contributions are referenced throughout my GitHub repository). One of the main external contributions was using a WAE with Ren's three-stage training process, which was first conceptualized by GitHub user 'MariaPdg'. With their permission, I adapted the underlying structure of their network. However, a considerable amount of effort remained involved in adapting the code for my purposes. Specifically, this was in formatting the NSD data, defining entirely new loss functions, parameter tuning, networking optimization, and proposing a novel synthesis of existing network architectures by adding visual feature guidance (see Ren et al., 2021) and changing the objective of the autoencoder regularizer in Stage 2 (see Appendix D).

Finally, the data analysis plan was developed under my supervisor's guidance and with some minor input from Ashley York and Brendan Zietsch. However, the analyses were implemented independently. I designed all figures except where stated otherwise, and interpretation of results was developed through discussions with my supervisor.

Results

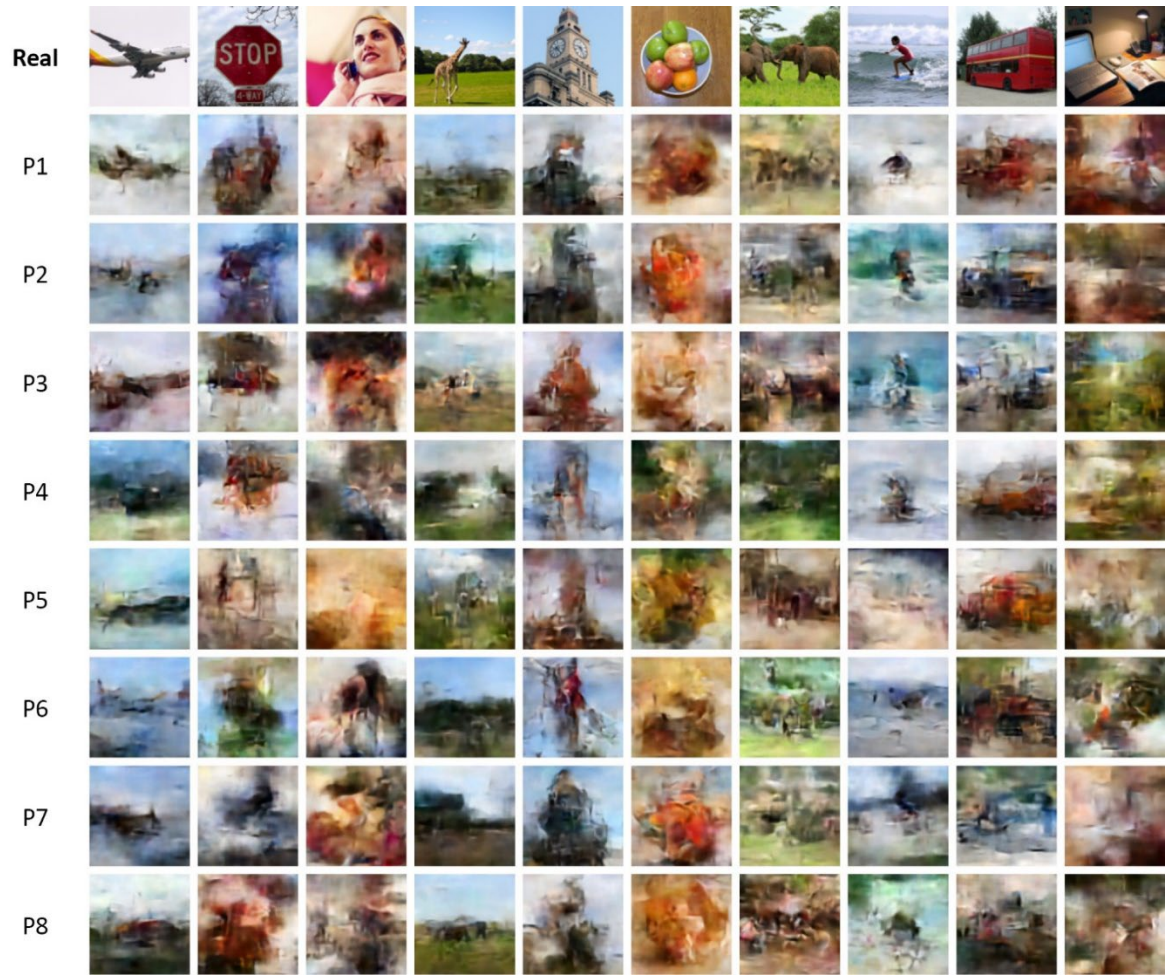
Please note: all subsequent results and qualitative examples are based on reconstructions generated from test data (i.e., data not seen by the network during training).

Study 1 – Viability of the NSD for Natural Image Reconstruction

To investigate the viability and consistency of reconstructions generated using data from the NSD, the reconstruction quality of networks trained on each participant's full set of training data was evaluated. Figure 9 shows a comparison of reconstructions between participants. Variability between participant's reconstructions is particularly notable in reconstructions of the stop sign, where reconstructions vary in their portrayal of the colour of the original stimulus. Despite this variability, most reconstructions successfully capture important visual features, including colour and structure, particularly in images that appear to have high contrast. For example, reconstructions of the plane, giraffe and building tend to capture the edge details and colour at the intersection of foreground and background.

Figure 9

Qualitative Comparison of Reconstructions Across Participants



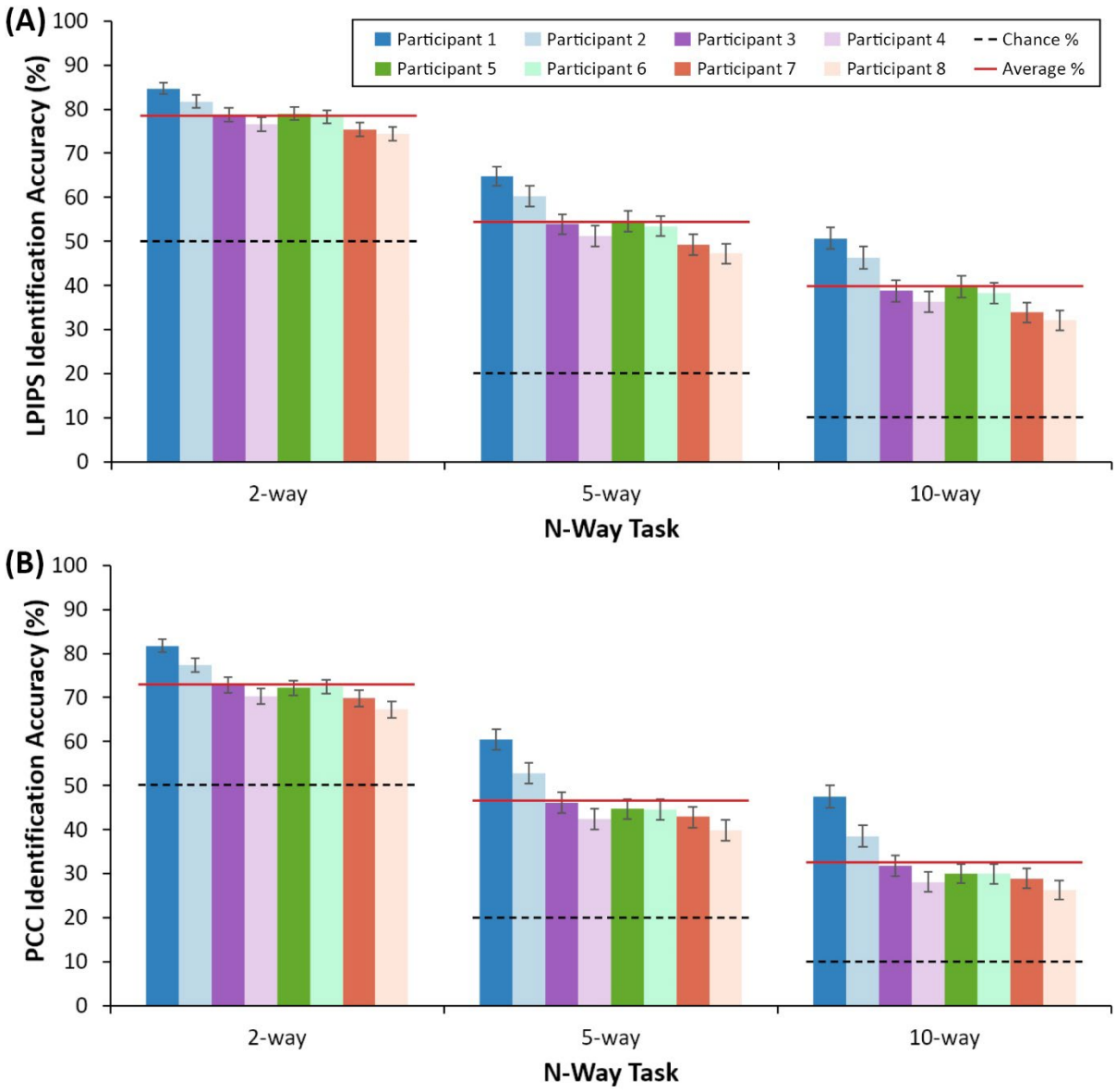
Note. Variability of reconstruction quality across participants (P1-P8). The top row shows ground truth ('real') image. Subsequent rows show reconstructions generated by each participant's trained network. More reconstruction examples can be found in Appendix G or downloaded from the Open Science Framework[1].

Mean accuracy across all participants for the 2-, 5- and 10-way tasks using LPIPS was 78.6%, 54.3%, and 39.5%, respectively. Similarly, the 2-, 5- and 10-way mean accuracy scores calculated using PCC were 73.0%, 46.7%, and 32.6%, respectively. Importantly, the permutation tests indicated that for all eight participants, across each n-way task and both metrics, mean accuracy was significantly above chance, $ps < .001$ (see Figure 10).

[1] https://osf.io/xq3cu/?view_only=2e61bd7f209e45f0ae8dfc49289f80b9

Figure 10

Variability in Identification Accuracy Across Participants



Note. Reconstruction accuracy scores across eight participants using the n-way comparison task on LPIPS (Panel A) and PCC (Panel B). Error bars reflect bootstrapped 95% confidence intervals. The red line shows participants' average accuracy score (%), and the black line shows theoretical chance level accuracy (%). However, performance against chance was tested statistically using the permutation test described in the Methods section.

Interestingly, Participant 1 consistently outperformed all other participants, scoring significantly higher mean accuracies on the 5-way LPIPS task, 95% CI [62.55%, 66.86%],

and PCC metrics, 95% CI [58.18%, 62.85%]. This trend was consistent across all n-way tasks and for both metrics, except for the 10-way LPIPS task, where the difference between Participant 1 (95% CI [48.19%, 53.12 %]) and Participant 2 (95% CI [43.81%, 48.81%]) was non-significant (see Appendix E for the full list of accuracy results). This was the only finding that differed between n-way tasks. As such, for brevity, all subsequent results will be reported using only the 5-way identification accuracy task. However, results for the 2- and 10-way tasks can be found in Appendix F.

Main Analysis

Consistent with the previous literature, comparisons were made with reconstruction accuracies pooled across all participants (N = 6,976; Gaziv et al., 2022). Bootstrapped 95% confidence intervals are also provided, which minimize assumptions about distributions and provide a more intuitive estimate of mean differences (DiCiccio & Efron, 1996; Jung et al., 2019). In this framework, mean accuracy scores can be considered significantly different if the bootstrapped confidence intervals of two conditions are non-overlapping (Knezevic, 2008).

Assumption Checks

Non-parametric tests are well established in the literature as the most appropriate test for comparisons between neural networks (Demšar, 2006). Additionally, a series of Lilliefors-corrected Kolmogorov-Smirnov tests were conducted to demonstrate that the distributions of accuracy scores were non-normal and, therefore, suitable for non-parametric analyses. All tests indicated that the distributions of raw reconstruction accuracy scores significantly deviated from normality, $ps < .001$. The Wilcoxon signed-rank test also requires that the distribution of difference scores between the dependent groups be symmetrical (King & Eckersley, 2019). As such, the skewness of each distribution of difference scores used in

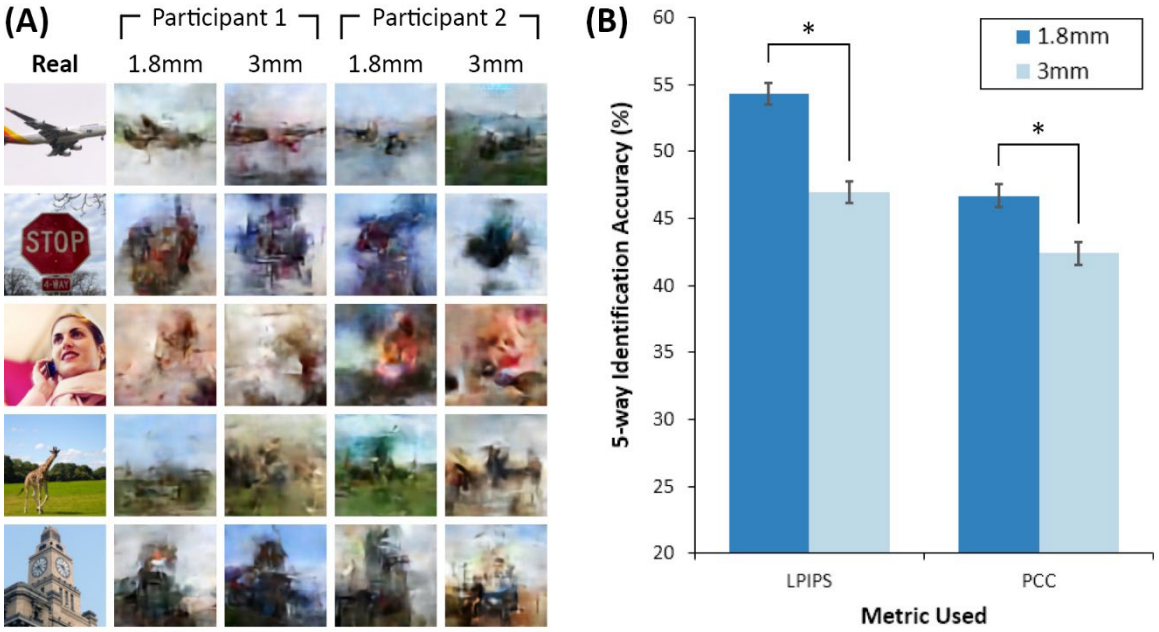
Studies 2 and 3 was evaluated. All values for skewness were excellent (George & Mallery, 2019), falling between the range of +/- 1 (skewness < |0.60|).

Study 2 – Effect of Voxel Resolution

Differences in reconstructions between networks trained on 1.8 mm and 3 mm voxels were evaluated to examine the effect of the NSD's finer voxel resolution on reconstruction quality. These differences can be observed in the qualitative comparisons of reconstructions in Figure 11A. Specifically, reconstructions on 1.8 mm voxels appear more consistent in their generation of edge features like horizons and the outlines of objects and in their reconstruction of colour, such as the red in the stop sign. The degradation of these features on reconstructions from 3 mm is particularly noticeable in reconstructions from Participant 2.

Figure 11

The Effect of Voxel Resolution on Reconstruction Quality



Note. (A) Qualitative comparison of reconstructions from Participant 1 and Participant 2. The left column shows ground truth 'real' image. For Participants 1 and 2, reconstructions from test fMRI data at each voxel resolution are displayed. More reconstruction examples can be found in Appendix H. (B) Quantitative comparison of the effect of voxel resolution on identification accuracy in the 5-way task for each metric. The accuracy scores are pooled across all participants' test reconstructions (N = 6,976). Error bars show 95% CI by bootstrapping (N = 10,000). The lower bound for the y-axis is set at chance level accuracy. * $p < .001$, and significant difference defined by non-overlapping 95% confidence intervals.

For objective assessment, identification accuracy scores between voxel resolution conditions were compared using a series of two-tailed Wilcoxon signed-rank tests (see Figure 11B). A Bonferroni correction was applied to account for three n-way tests across both metrics (N = 6), resulting in an alpha level of .008. Results indicated that models trained on 1.8 mm voxels achieved greater reconstruction accuracy scores measured using the LPIPS metric (M = 54.3%, SD = 34.8%, 95% CI [53.51%, 55.14%]) than did those trained on 3 mm

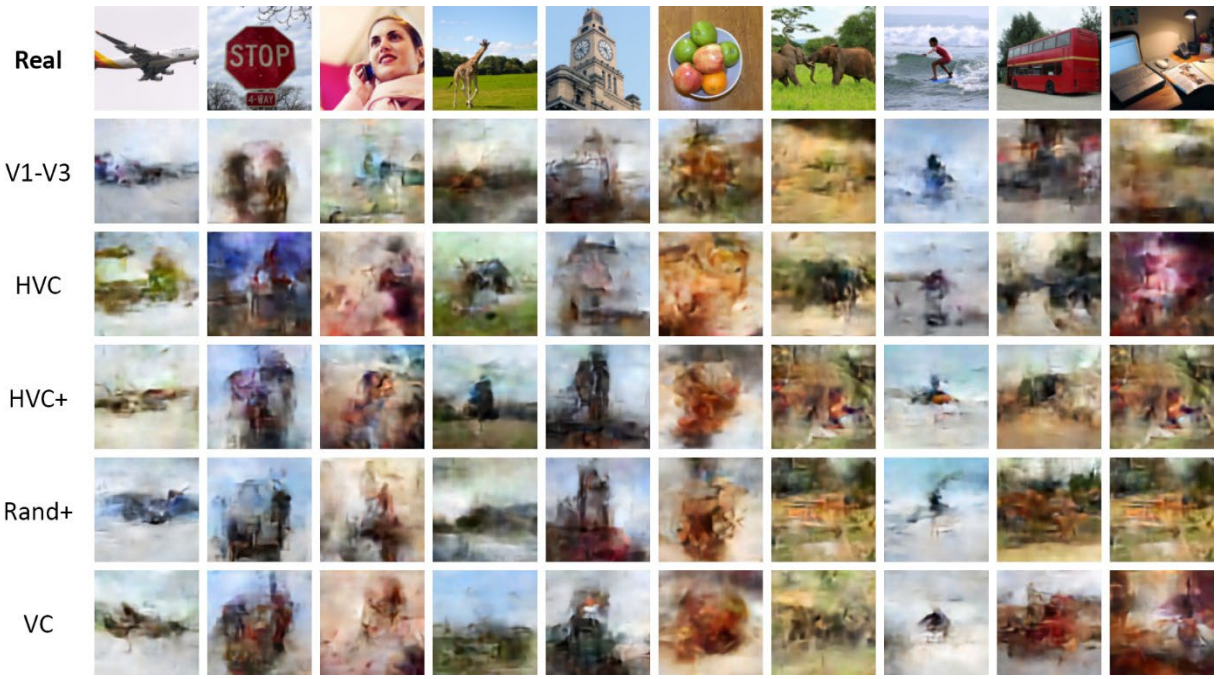
voxels (M = 47.0%, 95% CI [46.14%, 47.77%]), W = 8,035,776.00, r_{pb} = .31, p < .001. Similarly, when calculating reconstruction accuracy scores using the PCC metric, networks trained on 1.8 mm voxels (M = 46.7%, 95% CI [46.14%, 47.77%]) scored significantly higher in the 5-way identification task than those trained on 3 mm voxels (M = 42.4%, 95% CI [41.58%, 43.23%]), W = 9,757,611.00, r_{pb} = .17, p < .001.

Study 3 – Effect of ROIs on Reconstruction Quality

Finally, to investigate the influence of different ROIs on reconstruction quality (see Figure 12 for examples), identification accuracy results from networks trained on voxels either from V1-V3 or HVC were compared using a series of two-sided Wilcoxon signed-rank tests. Again, a Bonferroni correction was applied to account for six comparisons, resulting in a critical significance level of .008. Results revealed that models trained on V1-V3 voxels (M = 44.6%, 95% CI [43.76%, 45.40%]) achieved significantly greater reconstruction accuracy than did those trained only on HVC voxels (M = 40.9%, 95% CI [40.12%, 41.70%]) when measured using LPIPS metric, W = 9,976,656.00, r_{pb} = .14, p < .001. Similarly, when using the PCC metric for the 5-way task, networks trained on V1-V3 voxels (M = 43.4%, 95% CI [42.52%, 44.20%]) achieved significantly greater identification accuracy than networks trained on HVC voxels (M = 33.7%, 95% CI [32.99%, 34.53%]), W = 8,550,719.00, r_{pb} = .28, p < .001 (see Figure 13A).

Figure 12

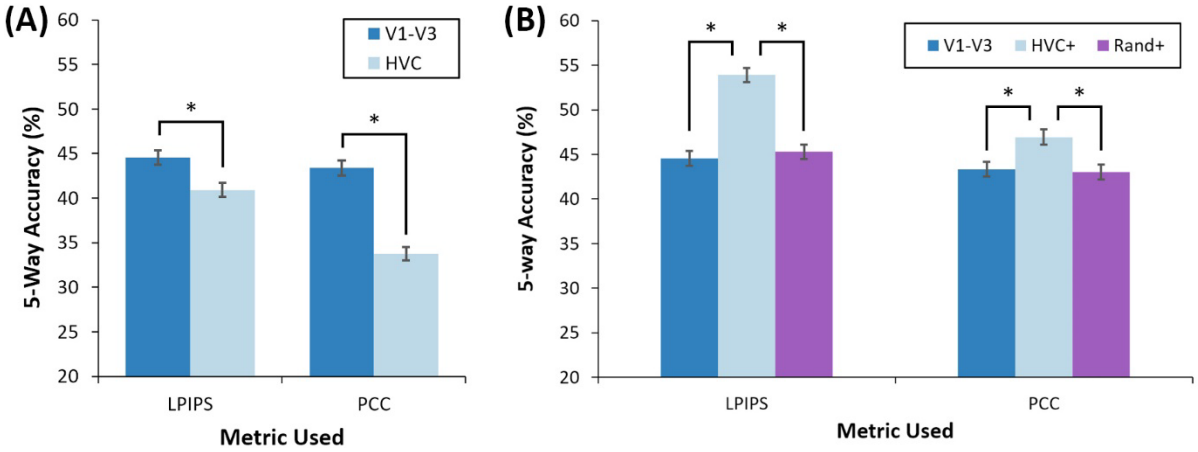
Qualitative Comparison of Reconstructions from Different ROIs



Note. Qualitative comparison of reconstructions from Participant 1 networks trained on different ROIs. More ROI-specific reconstructions can be found in Appendix I. The top row shows ground truth 'real' image, and each following row shows the reconstructions from test fMRI data from networks trained solely on the ROI defined. The final row shows reconstructions from networks trained on all visual cortex voxels (i.e., the same reconstructions from Study 1). There is notable degradation of reconstruction quality when generated from HVC alone, relative to V1-V3 or V1-V3 with HVC (HVC+). For example, the edges of the plane and building are blurrier for HVC reconstructions than for V1-V3. Rand+ does not appear to provide visual detail over V1-V3 alone, whereas HVC+ appears to add some semantic information – for example, in the giraffe reconstruction where it generates a tree-like object in the foreground.

Figure 13

The Effect of Different ROIs on Reconstruction Quality



Note. (A) Quantitative comparison of identification accuracy from networks trained only on V1-V3 or HVC voxels. (B) Quantitative comparison of identification accuracy from networks trained on either V1-V3 alone, V1-V3 with HVC (HVC+), and V1-V3 plus a random subset of non-visual cortex voxels, matching the voxel counts of HVC (Rand+). For both panels, accuracy is calculated for both metrics in the 5-way identification task. Again, results for 2- and 10-way tasks can be found in Appendix F. Accuracy scores are pooled across all participants' test reconstructions (N = 6,976). Error bars show 95% CI by bootstrapping. * p < .001, and significant difference defined by non-overlapping 95% confidence intervals.

Next, a series of Friedman tests were used to evaluate the additive effect of HVC in conjunction with V1-V3 (HVC+), compared to V1-V3 alone or with a random subset of voxels (Rand+; see Figure 13B). Using the LPIPS metric, the Friedman test revealed a significant difference between identification accuracy from networks trained on either V1-V3, HVC+ and Rand+, $\chi^2(2) = 713.99$, $p < .001$. Similarly, with the PCC metric, a Friedman test revealed a significant difference between accuracy scores from the same three ROIs, $\chi^2(2) = 116.29$, $p < .001$.

These significant results were followed up with Wilcoxon signed-rank tests with a Bonferroni-corrected alpha level of .003 to account for 18 comparisons (three comparisons

for each n-way task on two metrics). First, for identification tasks conducted using the LPIPS metric, the results of a Wilcoxon signed-rank tests revealed that accuracy from networks trained on HVC+ voxels (M = 53.9%, 95% CI [53.08%, 54.71%]) were significantly greater than both V1-V3 alone (W = 7,390,471.00, r_{pb} = .36, p < .001), and Rand+ (M = 45.3%, 95% CI [44.48%, 46.09%]), W = 7,489,396.00, r_{pb} = .35, p < .001. However, there was no significant difference in identification accuracy between networks trained on V1-V3 alone, and Rand+, W = 11,111,607.50, r_{pb} = .07, p = .025. Likewise for the PCC metric, the results showed that identification accuracy of reconstructions generated from HVC+ voxels (M = 46.9%, 95% CI [46.07%, 47.78%]) were significantly higher than those generated from V1-V3 (W = 10,287,704.00, r_{pb} = .12, p < .001), and Rand+ (M = 43.0%, 95% CI [42.21%, 43.87%]), W = 10,011,753.00, r_{pb} = .14, p < .001. Again, there was no significant difference in accuracy between networks trained on V1-V3 alone, and Rand+, W = 11,457,687.00, r_{pb} = -.01, p = .299.

Discussion

The current thesis aimed to establish the viability of the NSD for natural image reconstruction using machine learning. Additionally, it aimed to determine the extent to which the NSD's finer voxel resolution might allow for higher quality reconstructions and to evaluate the role of low- and high-order visual areas in perception. To achieve this, 48 neural networks were trained to reconstruct seen images from carefully manipulated variations of the NSD. Reconstruction quality was measured using three n-way identification accuracy tasks computed with two leading image similarity metrics for low-level visual features (PCC) and high-level perceptual similarity (LPIPS).

Viability of NSD

The current findings are consistent with the first hypothesis, such that networks trained on the NSD data generated high fidelity reconstructions that consistently performed

above chance across multiple task difficulty levels (i.e., multiple n-way tasks). Therefore, to my knowledge, this study is the first to show that the NSD can be used for brain decoding tasks such as natural image reconstruction. This is a critical contribution to extending the potential application of the dataset beyond encoding models, which predict neural responses to visual stimuli (Gu et al., 2022; Khosla & Wehbe, 2022). Furthermore, by using the most comprehensive deep learning fMRI dataset currently available, the present thesis marks an essential step in overcoming the technical limitations of small-scale datasets. Namely, that machine learning approaches for natural image reconstruction have long required larger datasets to reconstruct more complex visual stimuli (Rakhimberdina et al., 2021).

The results are also consistent with previous research demonstrating the suitability of machine learning approaches for decoding visual content from fMRI data (Rakhimberdina et al., 2021). Specifically, the results suggest that the proposed network effectively encodes visual features from novel brain data to reconstruct recognizable features of the original image. This finding indicates that the NSD data accurately maps natural scene images to stable patterns of brain activity, and, as a result, the present reconstructions achieve similar levels of identification accuracy as those reported in previous works (Beliy et al., 2019; Gaziv et al., 2022; Ren et al., 2021). This is regardless of the fact that the size of the current test set likely increases the difficulty of the identification tasks due to increased similarity between candidate images (see Appendix J for examples).

Despite the overall success of the current approach, there is notable variation in the quality of reconstructions between participants, though this is not uncommon in the previous literature (Gaziv et al., 2022; Ren et al., 2021). Investigating the causal factors in these individual differences lies beyond the scope of this thesis. However, previous research indicates that differences in the training set size of each participant (Shen et al., 2019a), variation in the voxel counts (Gaziv et al., 2022), and variability in the overall quality of data

might contribute to discrepancies in decoding accuracy (Allen et al., 2022). Nevertheless, qualitative comparisons of the present reconstructions show a large degree of overlap, suggesting that the networks learn similar heuristics for decoding the cognitive data despite being trained on unique sets of images and their evoked responses.

Finally, most existing literature assumes that the theoretical chance level for each n-way ($N = 2, 5, 10$) task is simply $1/N$. However, one of the main strengths of the present work is in utilizing a novel permutation test to establish an empirical null distribution for the n-way task. In doing so, the number of assumptions made about the null distribution is minimized (Lee & Kuhl, 2016), which is particularly important given that the distributions of accuracy scores were highly non-normal. Overall, this consideration offers a robust and powerful test that provides an exact measure of the likelihood of the observed accuracies under the null hypothesis.

Voxel Resolution

Consistent with predictions, finer voxel resolutions led to higher quality reconstructions, reflected in higher identification accuracy. These findings are in line with previous studies, which show that finer voxel resolutions can improve the decoding of fMRI data (Gardumi et al., 2016). However, research also indicates that the effect of voxel resolution may be task-dependent (Sengupta et al., 2017). In this context, the current findings suggest that natural image reconstruction may be a decoding task that benefits from imaging with increased spatial resolution. This may be related to the fact that the information required to reconstruct complex natural scenes is encoded at multiple scales (Olman & Yacoub, 2011), which may be more accurately captured at a resolution of 1.8 mm.

Partial volume effects may also drive the effect of voxel resolution on reconstruction quality. As discussed, smaller voxels decrease the likelihood that voxel activity captures multiple tissue types or points from both sides of opposing sulcal banks (Gardumi et al.,

2016). Furthermore, at 1.8 mm, voxels more closely match the size of ocular dominance columns (~1 mm; Cheng, 2007). Therefore, sampling from visual areas where cortical columns are ubiquitous (e.g., low-order visual areas), with less partial volume effects, may provide the neural networks with rich information reflecting the complex functional organization of the visual cortex (Zeki et al., 1991).

The Effect of Low- and High-Order Visual Areas

A perturbation paradigm was used to manipulate the visual areas included in network training to investigate the effect of low- and high-order visual areas on reconstruction quality. As hypothesized and in line with previous work, the current results show that low-order visual areas produce improved reconstructions compared to higher visual areas (Gaziv et al., 2022; Han et al., 2019; Ren et al., 2021). Crucially, a strength of the current study is in considering the nuanced role that these higher visual areas play. Specifically, in line with the functional specialization of the visual cortex, it was hypothesized that high-order visual areas would be supplementary to low visual areas – providing semantic and contextual information above and beyond low-level visual features (Courtney & Ungerleider, 1997). Consistent with this view, the results reveal that reconstructions from low- and high-order visual areas combined produced higher quality reconstructions than lower visual areas alone as well as lower visual areas combined with non-visual areas. These findings suggest that the benefit of high-order visual areas is particularly notable in conjunction with lower areas and not in isolation. These results align with previous research that shows that high-order visual areas provide semantic information that can improve the quality of natural image reconstructions (Gaziv et al., 2022).

Going beyond this previous work, the present study rules out the possibility that this additive effect of high-visual areas might be explained simply by providing more voxels to the network, thereby increasing model complexity. Yet despite doubling or sometimes even

tripling the voxel count of V1-V3, data from non-visual areas do not provide any further benefit to reconstruction quality when added to low visual areas. This finding is critical because it reveals that high visual areas offer a distinct benefit to reconstruction quality that is not found when the same amount of data is randomly sampled from non-visual regions of the brain. This further establishes that high-order visual areas provide rich semantic and contextual information for visual perception. Furthermore, this is in line with research showing that natural scenes are not simply processed as a culmination of low-level visual features but are characterized by additional contextual information provided by semantic-rich high-order visual areas (Doerig et al., 2022).

Enhanced Effect of High-Order Visual Areas Measured Using LPIPS

Beyond the proposed hypotheses, one unexpected finding further highlights the specialized roles of low- and high-order visual areas. Specifically, the results revealed that the additive effect of high-order visual areas was considerably stronger for the LPIPS metric than for PCC. One plausible explanation for these results relates to what these metrics attempt to measure. On the one hand, PCC is calculated pixel-wise and only captures fine spatial details (i.e., low-level features; Shen et al., 2019a). By comparison, LPIPS is thought to capture high-level features, including high-order structure, object categories, and semantic information (Mozafari et al., 2020; Zhang et al., 2018). Therefore, if high-order visual areas are critical for integrating semantic information into reconstructions, then it is likely that the LPIPS metric captures these feature-specific benefits much more so than the PCC metric.

Limitations

The present work offers several significant contributions to the literature; however, there are some fundamental limitations to consider. First, due to time constraints and the project's technical complexity, implementing a human-based evaluation of image similarity was not viable. Much of the recent literature uses a subjective n-way identification accuracy

task that involves human observers selecting the candidate image most similar to the reconstruction (Rakhimberdina et al., 2021). The lack of such an assessment might restrict the validity of the current measures of reconstruction quality. To address this, the present thesis implemented a high-level perceptual similarity metric (LPIPS) that is known to perform in line with human judgements (Zhang et al., 2018). Furthermore, previous research has established that subjective and objective (i.e., metric-based) identification accuracy tasks are highly correlated (Gaziv et al., 2022; Ren et al., 2021). Notwithstanding these considerations, improving identification accuracy measured using similarity metrics may not guarantee that two sets of reconstructions are qualitatively different. It is therefore recommended that future research implements both objective and subjective assessments of image similarity to reliably capture the nature of human perception. This would allow research to better quantify whether differences in reconstruction quality measured using similarity metrics were of practical significance (i.e., apparent to human observers).

Second, the current findings suggest that finer voxel resolution improves reconstruction accuracy and that reduced partial volume effects may explain this. However, voxel-based imaging may not represent cortical activity most appropriately. More specifically, by using volumetric data (i.e., fMRI data collected in voxel-space), the proposed network treats every voxel as independent and is therefore ignorant of the spatial relationships within the cortex, which contain rich stimulus-related information (Bullmore & Sporns, 2009; Ortega et al., 2018). In contrast, surface-based approaches use folded meshes or graphs, which preserve the folded structure of the brain, thereby maintaining its spatial relationships (Ribeiro et al., 2022; Sarasua et al., 2021). As such, there is an emerging field known as geometric deep learning, which seeks to adapt deep learning models which can exploit the spatial relationships that are preserved in surface-based data (Bronstein et al., 2017). These approaches have proven effective in predicting the retinotopic organization of

the visual cortex (Ribeiro et al., 2021) and decoding brain states using fMRI (Zhang et al., 2021). These findings suggest that natural image reconstruction may also benefit from handling its input data in a manner which maintains the spatial relationships of the visual cortex. Therefore, future research should investigate whether geometric approaches can be successfully applied to natural image reconstruction. This research may even look to compare the relative quality of volumetric and surface-based reconstructions to determine the extent to which these spatial relationships play a role in visual perception.

Conclusion

This thesis used deep learning to reconstruct images from novel brain activity measured using fMRI. The current results present three key findings. First, the NSD is suitable for decoding tasks such as natural image reconstructions, which suggests that the data accurately maps natural scene images to their stable patterns of brain activity. Second, the improved voxel resolution of the NSD provides significant benefits to reconstruction quality. Finally, high-order visual areas can improve reconstruction quality, particularly when working in conjunction with lower visual areas. This reinforces the role of high-order visual areas as providing contextual and semantic information for visual perception. Overall, this thesis demonstrates how deep learning can facilitate neuroscience inquiry and highlights that the NSD provides high-fidelity, large-scale data, which may prove vital in bridging the gap between these fields.

References

Alabdullatef, L. (2020). *Complete guide to Adam optimization*. https://towardsdatascience.com/complete-guide-to-adam-optimization-1e5f29532c3d

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. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, *25*(1), 116-126. https://doi.org/10.1038/s41593-021-00962-x

Beliy, R., Gaziv, G., Hoogi, A., Strappini, F., Golan, T., & Irani, M. (2019). From voxels to pixels and back: Self-supervision in natural-image reconstruction from fMRI. *Advances in Neural Information Processing Systems*, *32*. https://proceedings.neurips.cc/paper/2019/file/7d2be41b1bde6ff8fe45150c37488ebb-Paper.pdf

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. (2018). The Human Connectome Project 7 Tesla retinotopy dataset: Description and population receptive field analysis. *Journal of Vision*, *18*(13), 23-23. https://doi.org/10.1167/18.13.23

Bhandari, A. (2022). *Feature scaling for machine learning: Understanding the difference between normalization vs. standardization*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., & Vandergheynst, P. (2017). Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, *34*(4), 18-42. https://doi.org/10.1109/MSP.2017.2693418

Brouwer, G. J., & Heeger, D. J. (2009). Decoding and reconstructing color from responses in human visual cortex. *Journal of Neuroscience*, *29*(44), 13992-14003. https://doi.org/10.1523/JNEUROSCI.3577-09.2009

Bullmore, E., & Sporns, O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, *10*(3), 186-198. https://doi.org/10.1038/nrn2575

Cai, Y., Hofstetter, S., van der Zwaag, W., Zuiderbaan, W., & Dumoulin, S. O. (2021). Individualized cognitive neuroscience needs 7T: Comparing numerosity maps at 3T and 7T MRI. *NeuroImage*, *237*, 118184. https://doi.org/10.1016/j.neuroimage.2021.118184

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

Chen, M., Han, J., Hu, X., Jiang, X., Guo, L., & Liu, T. (2014). Survey of encoding and decoding of visual stimulus via FMRI: An image analysis perspective. *Brain Imaging and Behavior*, *8*(1), 7-23. https://doi.org/10.1007/s11682-013-9238-z

Cheng, K. (2007). Revealing columnar architectures using fMRI: Challenges and possibilities. *VISION*, *19*(1), 29-35. https://doi.org/10.24636/vision.19.1_29

Courtney, S. M., & Ungerleider, L. G. (1997). What fMRI has taught us about human vision. *Current Opinion in Neurobiology*, *7*(4), 554-561. https://doi.org/10.1016/S0959-4388(97)80036-0

Cox, D. D., & Dean, T. (2014). Neural networks and neuroscience-inspired computer vision. *Current Biology*, *24*(18), R921-R929. https://doi.org/10.1016/j.cub.2014.08.026

Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, *29*(3), 162-173. https://doi.org/10.1006/cbmr.1996.0014

Cox, R. W., & Hyde, J. S. (1997). Software tools for analysis and visualization of fMRI data. *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In Vivo*, *10*(4-5), 171-178. https://doi.org/10.1002/(SICI)1099-1492(199706/08)10:4/5<171::AID-NBM453>3.0.CO;2-L

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, *7*, 1-30. https://www.jmlr.org/papers/volume7/demsar06a/demsar06a.pdf

DiCiccio, T. J., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, *11*(3), 189-228. https://doi.org/10.1214/ss/1032280214

Doerig, A., Kietzmann, T. C., Allen, E., Wu, Y., Naselaris, T., Kay, K., & Charest, I. (2022). Semantic scene descriptions as an objective of human vision. *arXiv e-prints arXiv:2209.11737*. https://doi.org/https://doi.org/10.48550/arXiv.2209.11737

Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science*, *293*(5539), 2470-2473. https://doi.org/10.1126/science.1063414

Du, B., Cheng, X., Duan, Y., & Ning, H. (2022). fMRI brain decoding and its applications in brain-computer interface: A survey. *Brain Sciences*, *12*(2). https://doi.org/10.3390/brainsci12020228

Du, C., Du, C., Huang, L., & He, H. (2018). Reconstructing perceived images from human brain activities with Bayesian deep multiview learning. *IEEE Transactions on Neural*

Networks and Learning Systems, *30*(8), 2310-2323. https://doi.org/10.1109/TNNLS.2018.2882456

Du, C., Li, J., Huang, L., & He, H. (2019). Brain encoding and decoding in fMRI with bidirectional deep generative models. *Engineering*, *5*(5), 948-953. https://doi.org/10.1016/j.eng.2019.03.010

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

Fang, T., Qi, Y., & Pan, G. (2020). Reconstructing perceptive images from brain activity by shape-semantic GAN. *Advances in Neural Information Processing Systems*, *33*, 13038-13048. https://proceedings.neurips.cc/paper/2020/file/9813b270ed0288e7c0388f0fd4ec68f5-Paper.pdf

Gardumi, A., Ivanov, D., Hausfeld, L., Valente, G., Formisano, E., & Uludağ, K. (2016). The effect of spatial resolution on decoding accuracy in fMRI multivariate pattern analysis. *NeuroImage*, *132*, 32-42. https://doi.org/10.1016/j.neuroimage.2016.02.033

Gauthier, I., Tarr, M. J., Moylan, J., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). The fusiform "face area" is part of a network that processes faces at the individual level. *Journal of Cognitive Neuroscience*, *12*(3), 495-504. https://doi.org/10.1162/089892900562165

Gaziv, G., Beliy, R., Granot, N., Hoogi, A., Strappini, F., Golan, T., & Irani, M. (2022). Self-supervised natural image reconstruction and large-scale semantic classification from brain activity. *NeuroImage*, *254*, 119121. https://doi.org/10.1016/j.neuroimage.2022.119121

George, D., & Mallery, P. (2019). Descriptive Statistics. In *IBM SPSS statistics 25 step by step: A simple guide and reference* (pp. 112-119). Routledge.

Goense, J., Bohraus, Y., & Logothetis, N. K. (2016). fMRI at high spatial resolution: Implications for BOLD-models. *Frontiers in Computational Neuroscience*, *10*, 66. https://doi.org/10.3389/fncom.2016.00066

Grill-Spector, K., & Malach, R. (2004). The human visual cortex. *Annual Review of Neuroscience*, *27*, 649-677. https://doi.org/10.1146/annurev.neuro.27.070203.144220

Gu, Z., Jamison, K., Sabuncu, M., & Kuceyeski, A. (2022). Personalized visual encoding model construction with small data. *arXiv preprint arXiv:2202.02245*. https://doi.org/10.48550/arXiv.2202.02245

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

Haynes, J.-D., & Rees, G. (2005). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature Neuroscience*, *8*(5), 686-691. https://doi.org/10.1038/nn1445

Heeger, D. J., & Ress, D. (2002). What does fMRI tell us about neuronal activity? *Nature Reviews Neuroscience*, *3*(2), 142-151. https://doi.org/10.1038/nrn730

Horikawa, T., & Kamitani, Y. (2017a). Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, *8*(1), 1-15. https://doi.org/10.1038/ncomms15037

Horikawa, T., & Kamitani, Y. (2017b). Hierarchical neural representation of dreamed objects revealed by brain decoding with deep neural network features. *Frontiers in Computational Neuroscience*, *11*, 4. https://doi.org/10.3389/fncom.2017.00004

Hutchison, J. L., Hubbard, N. A., Brigante, R. M., Turner, M., Sandoval, T. I., Hillis, G. A. J., Weaver, T., & Rypma, B. (2014). The efficiency of fMRI region of interest

analysis methods for detecting group differences. *Journal of Neuroscience Methods*, *226*, 57-65. https://doi.org/10.1016/j.jneumeth.2014.01.012

Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, *76*(6), 1210-1224. https://doi.org/10.1016/j.neuron.2012.10.014

Jung, K., Lee, J., Gupta, V., & Cho, G. (2019). Comparison of bootstrap confidence interval methods for GSCA using a Monte Carlo simulation. *Frontiers in Psychology*, *10*, 2215. https://doi.org/10.3389/fpsyg.2019.02215

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, *8*(5), 679-685. https://doi.org/10.1038/nn1444

Kamitani, Y., & Tong, F. (2006). Decoding seen and attended motion directions from activity in the human visual cortex. *Current Biology*, *16*(11), 1096-1102. https://doi.org/10.1016/j.cub.2006.04.003

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. https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, *452*(7185), 352-355. https://doi.org/10.1038/nature06713

Khosla, M., & Wehbe, L. (2022). High-level visual areas act like domain-general filters with strong selectivity and functional specialization. *bioRxiv*. https://doi.org/10.1101/2022.03.16.484578

King, A. P., & Eckersley, R. (2019). Inferential statistics III: Nonparametric hypothesis testing. In A. P. King & R. Eckersley (Eds.), *Statistics for biomedical engineers and*

scientists (pp. 119-145). Academic Press. https://doi.org/10.1016/B978-0-08-102939-8.00015-3

Knezevic, A. (2008). Overlapping confidence intervals and statistical significance. *StatNews: Cornell University Statistical Consulting Unit*, *73*. https://web.archive.org/web/20190430205057id_/http://www.cscu.cornell.edu/news/statnews/stnews73.pdf

Kriegeskorte, N., & Bandettini, P. (2007). Analyzing for information, not activation, to exploit high-resolution fMRI. *NeuroImage*, *38*(4), 649-662. https://doi.org/10.1016/j.neuroimage.2007.02.022

Lee, H., & Kuhl, B. A. (2016). Reconstructing perceived and retrieved faces from activity patterns in lateral parietal cortex. *Journal of Neuroscience*, *36*(22), 6069-6082. https://doi.org/10.1523/JNEUROSCI.4286-15.2016

Lemm, S., Blankertz, B., Dickhaus, T., & Müller, K.-R. (2011). Introduction to machine learning for brain imaging. *NeuroImage*, *56*(2), 387-399. https://doi.org/10.1016/j.neuroimage.2010.11.004

Mazaika, P. (2009). Percent signal change for fMRI calculations. https://cibsr.stanford.edu/content/dam/sm/cibsr/documents/tools/methods/artrepair-software/FMRIPercentSignalChange.pdf

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. https://doi.org/10.1016/j.neuron.2008.11.004

Monti, M. M. (2011). Statistical analysis of fMRI time-series: A critical review of the GLM approach. *Frontiers in Human Neuroscience*, *5*, 28. https://doi.org/10.3389/fnhum.2011.00028

Mozafari, M., Reddy, L., & VanRullen, R. (2020). Reconstructing natural scenes from fMRI patterns using BigBiGAN. *2020 International Joint Conference on Neural Networks* 1-8. https://doi.org/10.1109/IJCNN48605.2020.9206960

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. https://doi.org/10.1073/pnas.87.24.9868

Olman, C. A., & Yacoub, E. (2011). High-field fMRI for human applications: An overview of spatial resolution and signal specificity. *The Open Neuroimaging Journal*, *5*, 74. https://doi.org/10.2174/1874440001105010074

Ortega, A., Frossard, P., Kovačević, J., Moura, J. M., & Vandergheynst, P. (2018). Graph signal processing: Overview, challenges, and applications. *Proceedings of the IEEE*, *106*(5), 808-828. https://doi.org/10.1109/JPROC.2018.2820126

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. (2017). Automatic differentiation in pytorch. https://openreview.net/pdf?id=BJJsrmfCZ

Peelen, M. V., & Downing, P. E. (2005). Selectivity for the human body in the fusiform gyrus. *Journal of Neurophysiology*, *93*(1), 603-608. https://doi.org/10.1152/jn.00513.2004

Podguzova, M. (2021). *Image reconstruction from human brain activity* [GitHub repository]. https://github.com/MariaPdg/thesis-fmri-reconstruction

Poldrack, R. A., & Farah, M. J. (2015). Progress and challenges in probing the human brain. *Nature*, *526*(7573), 371-379. https://doi.org/10.1038/nature15692

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

Ras, G., Xie, N., van Gerven, M., & Doran, D. (2022). Explainable deep learning: A field guide for the uninitiated. *Journal of Artificial Intelligence Research*, *73*, 329-397. https://doi.org/10.1613/jair.1.13200

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. https://doi.org/10.1016/j.neuroimage.2020.117602

Ribeiro, F. L., Bollmann, S., Cunnington, R., & Puckett, A. M. (2022). An explainability framework for cortical surface-based deep learning. *arXiv preprint arXiv:2203.08312*. https://doi.org/10.48550/arXiv.2203.08312

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

Sarasua, I., Lee, J., & Wachinger, C. (2021). Geometric deep learning on anatomical meshes for the prediction of Alzheimer's disease. *2021 IEEE 18th International Symposium on Biomedical Imaging*, 1356-1359. https://doi.org/10.1109/ISBI48211.2021.9433948

Sengupta, A., Yakupov, R., Speck, O., Pollmann, S., & Hanke, M. (2017). The effect of acquisition resolution on orientation decoding from V1 BOLD fMRI at 7 T. *NeuroImage*, *148*, 64-76. https://doi.org/10.1016/j.neuroimage.2016.12.040

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*, 21. https://doi.org/10.3389/fncom.2019.00021

Shen, G., Horikawa, T., Majima, K., & Kamitani, Y. (2019b). Deep image reconstruction from human brain activity. *PLoS Computational Biology*, *15*(1), e1006633. https://doi.org/10.1371/journal.pcbi.1006633

Singh, J., & Banerjee, R. (2019). A study on single and multi-layer perceptron neural network. *2019 3rd International Conference on Computing Methodologies and Communication*, 35-40. https://doi.org/10.1109/ICCMC.2019.8819775

Stigliani, A., Weiner, K. S., & Grill-Spector, K. (2015). Temporal processing capacity in high-level visual cortex is domain specific. *Journal of Neuroscience*, *35*(36), 12412-12424. https://doi.org/10.1523/JNEUROSCI.4822-14.2015

Thirion, B., Duchesnay, E., Hubbard, E., Dubois, J., Poline, J. B., Lebihan, D., & Dehaene, S. (2006). Inverse retinotopy: Inferring the visual content of images from brain activation patterns. *NeuroImage*, *33*(4), 1104-1116. https://doi.org/10.1016/j.neuroimage.2006.06.062

Tolstikhin, I., Bousquet, O., Gelly, S., & Schoelkopf, B. (2017). Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*. https://doi.org/10.48550/arXiv.1711.01558

Torrisi, S., Chen, G., Glen, D., Bandettini, P. A., Baker, C. I., Reynolds, R., Liu, J. Y.-T., Leshin, J., Balderston, N., & Grillon, C. (2018). Statistical power comparisons at 3T and 7T with a GO/NOGO task. *NeuroImage*, *175*, 100-110. https://doi.org/j.neuroimage.2018.03.071

Viessmann, O., & Polimeni, J. R. (2021). High-resolution fMRI at 7 tesla: Challenges, promises and recent developments for individual-focused fMRI studies. *Current Opinion in Behavioral Sciences*, *40*, 96-104. https://doi.org/10.1016/j.cobeha.2021.01.011

Wandell, B. A., & Winawer, J. (2011). Imaging retinotopic maps in the human brain. *Vision Research*, *51*(7), 718-737. https://doi.org/10.1016/j.visres.2010.08.004

Wen, H., Shi, J., Zhang, Y., Lu, K.-H., Cao, J., & Liu, Z. (2018). Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral Cortex*, *28*(12), 4136-4160. https://doi.org/10.1093/cercor/bhx268

Yacoub, E., Harel, N., & Uğurbil, K. (2008). High-field fMRI unveils orientation columns in humans. *Proceedings of the National Academy of Sciences*, *105*(30), 10607-10612. https://doi.org/10.1073/pnas.0804110105

Zeki, S., Watson, J., Lueck, C., Friston, K. J., Kennard, C., & Frackowiak, R. (1991). A direct demonstration of functional specialization in human visual cortex. *Journal of Neuroscience*, *11*(3), 641-649. https://doi.org/10.1523/JNEUROSCI.11-03-00641.1991

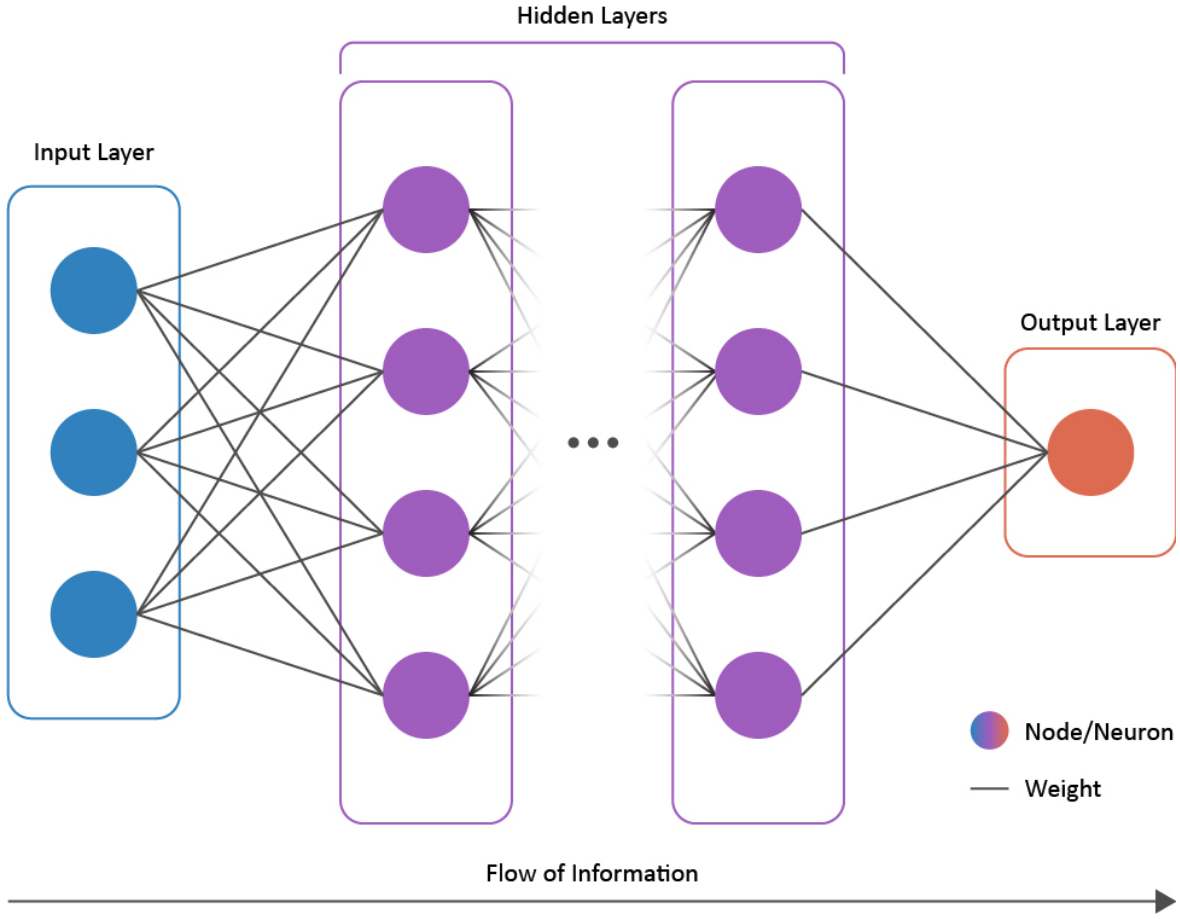
Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 586-595. https://openaccess.thecvf.com/content_cvpr_2018/papers/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.pdf

Zhang, Y., Tetrel, L., Thirion, B., & Bellec, P. (2021). Functional annotation of human cognitive states using deep graph convolution. *NeuroImage*, *231*, 117847. https://doi.org/10.1016/j.neuroimage.2021.117847

Zou, J., Han, Y., & So, S.-S. (2009). Overview of artificial neural networks. In D. J. Livingstone (Ed.), *Artificial neural networks: Methods and applications* (pp. 14-22). Humana Press. https://doi.org/10.1007/978-1-60327-101-1_2

Appendix A

Deep Neural Networks Explained



Note. Deep neural networks are comprised of an input and output layer, plus a variable number of hidden layers. Each of these layers is made of multiple nodes, and each connection between two nodes represents a weight that defines an input's relative importance. Nodes take a weighted sum of their inputs plus a bias (i.e., a constant term), the result of which is passed through a non-linear function known as an activation function (Singh & Banerjee, 2019). This process occurs for each neuron within each hidden layer until arriving at an output. Next, a loss function is calculated, which reflects the error between the observed and desired output. Most neural networks learn via a backpropagation algorithm, which iteratively updates the weights and biases of the model, intending to reduce the model's error (see Zou et al., 2009 for review).

Appendix B

Voxel Counts for All Included ROIs from Participants in NSD (1.8 mm)

Participant	V1v	V1d	V2v	V2d	V3v	V3d	V4	OFA	FFA-1	FFA-2	mTL-faces	aTL-faces	OPA	PPA	RSC	EBA	FBA-1	FBA-2	mTL-bodies	Total voxels
1	594	756	834	599	646	541	687	177	477	310	0	159	1607	1033	566	2798	197	134	0	12115
2	544	558	615	460	566	531	483	322	289	529	0	275	1366	994	813	3195	0	985	0	12525
3	724	530	653	488	422	506	426	545	697	396	0	159	1330	1251	838	3237	606	87	0	12895
4	485	392	479	384	463	345	475	430	447	463	163	0	1235	960	813	2863	68	343	127	10935
5	458	655	520	561	475	450	542	527	452	455	103	148	1330	1221	771	4042	204	395	296	13605
6	399	728	586	594	533	668	477	487	341	485	0	80	1224	1229	845	4005	686	140	142	13649
7	524	618	558	428	371	355	397	261	346	138	80	231	1052	912	694	2761	0	355	0	10081
8	522	552	567	466	479	410	495	260	519	648	273	203	1354	946	799	2890	270	210	121	11984

Note. Shows the voxel counts for each participant in the 1.8 mm preparation of the NSD data. All ROIs were delineated using pRF and fLoc experiments. Further details about these experiments and descriptions of the areas can be found at https://cvnlab.slite.page/p/X_7BBMgghj/ROIs

Appendix C

Available Sessions and Training Set Size for Each NSD Participant

Participant	Sessions Available	Total Train Set Size
Participant 1	37	24980
Participant 2	37	24980
Participant 3	29	19637
Participant 4	27	18265
Participant 5	37	24980
Participant 6	29	19637
Participant 7	37	24980
Participant 8	27	18265

Note. Some of the NSD participants missed several of the intended 40 scan sessions. Furthermore, at the time of writing, each participant's final three sessions of the main NSD experiment were withheld from public access. This table shows the number of sessions available for this thesis and, therefore, the total number of image/fMRI pairs in the training set.

Appendix D

Adapting the WAE/GAN Implementation

Given the word limit of the thesis and in an attempt to keep the project relevant to psychology and neuroscience, I was unable to go into depth concerning the technical contributions of the project. One such contribution is proposing a novel neural network architecture by adapting visual feature guidance as proposed by Ren et al. (2021). Recall the current network architecture was primarily adapted from code available on GitHub (Podguzova, 2021). This code largely follows the d-VAE/GAN architecture put forward by Ren et al. (2021), with some changes to adapt the network to PyTorch and to replace the VAE component of the model with a WAE. However, I sought to more accurately adapt visual feature guidance by changing the loss function of the network.

The code initially had implemented $\mathcal{L}_{recon}$ in Stage 2 as the mean-squared error (MSE) of the cognitive reconstruction (i.e., the reconstruction from the Cognitive Encoder) and the ground truth image. However, in Stage 2, $\mathcal{L}_{recon}$ was changed to reflect the original paper better. Specifically, Stage 2 now takes the MSE of cognitive reconstruction and the visual reconstruction (i.e., the reconstruction from the Visual Encoder trained in Stage 1). This is thought to allow for Cognitive Encoder to mimic the Visual Encoder better, which should have an enhanced ability to encode latent visual features due to being trained directly on the visual stimuli (Ren et al., 2021).

Furthermore, the code from Podguzova (2021) implements the Stage 2 WAE loss in a manner that is inconsistent with the logic of the original d-VAE/GAN network proposed by Ren et al. (2021). This is nuanced, because the goal of the discriminator in the context of the VAE is quite different from the WAE implemented here. However, to my knowledge, the discriminator in a WAE/GAN is set up to regularize the encoded latent features to match some ideal prior distribution (Tolstikhin et al., 2017). In Stage 1, this is implemented

correctly, such that the encoded distribution (i.e., the output of the Visual Encoder) is regularized to mimic a normally-distributed prior, which the discriminator sees as the 'real'. This makes sense, because the non-saturating loss function can force the 'fake' distribution (i.e., the encoded features) look more like the 'real' distribution as described in the main text.

However, in Stage 2, under the framework of visual feature guidance, the goal would be to make the encoded features of the Cognitive Encoder mimic those of the Visual Encoder trained in Stage 1. However, the code implements this in the opposite manner. Specifically, the output of the Cognitive Encoder was established as the 'real' to the discriminator instead of the output of the Visual Encoder. As such, to be more in line with Ren et al. (2021), the loss function was adjusted to reflect the visual feature guidance logic. More specifically, the real to the discriminator in Stage 2 was changed such that the 'real' of the discriminator was the output of the Visual Encoder, and the 'fake' was the output of the Cognitive Encoder. I believe that this more accurately reflects the goal of the network. However, future research is needed to investigate whether this translates to the use of WAEs for image reconstruction.

Technical Contributions of the Thesis

Overall, the current thesis provides further evidence for the suitability of dual encoder approaches. In line with Ren et al. (2021), the proposed architecture emphasises knowledge transfer between visual and cognitive encoders, which allows the latter to encode a richer set of latent visual features. Furthermore, by replacing the variational autoencoder component of Ren's d-VAE/GAN (Ren et al., 2021), the thesis shows dual encoders, visual feature guidance, knowledge distillation, and the use of multiple training stages can be effectively extended to new architectures. This may be critical for creating new architectures which can leverage these principles to improve the quality of natural image reconstructions.

Appendix E

Accuracy Results from N-Way Tasks Using LPIPS and PCC

	2-Way (%)		5-Way (%)		10-Way (%)	
Condition	LPIPS *M* [95% CI]	PCC *M* [95% CI]	LPIPS *M* [95% CI]	PCC *M* [95% CI]	LPIPS *M* [95% CI]	PCC *M* [95% CI]
Subject 1	84.1 [83.4, 86.0]	81.7 [80.2, 83.2]	64.7 [62.5, 66.9]	60.5 [58.2, 62.8]	50.6 [48.2, 53.1]	47.4 [44.9, 50.0]
Subject 2	81.7 [80.2, 83.2]	77.3 [75.7, 78.9]	60.2 [57.9, 62.5]	52.8 [50.5, 55.2]	46.3 [43.8, 48.8]	38.5 [36.1, 41.0]
Subject 3	78.7 [77.1, 80.2]	72.8 [71.0, 74.5]	54.0 [51.7, 56.2]	46.1 [43.7, 48.4]	38.8 [36.4, 41.1]	31.7 [29.4, 34.1]
Subject 4	76.6 [75.0, 78.2]	70.3 [68.5, 72.1]	51.2 [48.9, 53.5]	42.4 [40.1, 44.7]	36.4 [34.0, 38.7]	28.1 [25.9, 30.3]
Subject 5	79.0 [77.5, 80.5]	72.2 [70.5, 73.9]	54.6 [52.3, 56.9]	44.7 [42.3, 46.9]	39.7 [37.3, 42.1]	30.0 [27.8, 32.2]
Subject 6	78.3 [76.7, 79.8]	72.5 [70.8, 74.1]	53.4 [51.2, 55.7]	44.5 [42.1, 46.9]	38.3 [35.9, 40.6]	29.9 [27.7, 32.2]
Subject 7	75.4 [73.8, 77.0]	69.9 [68.0, 71.7]	49.2 [46.9, 51.5]	42.9 [40.5, 45.2]	33.8 [31.6, 36.1]	28.9 [26.7, 31.2]
Subject 8	74.4 [72.8, 76.0]	67.3 [65.4, 69.2]	47.3 [45.0, 49.5]	39.8 [37.5, 42.2]	32.1 [29.9, 34.4]	26.3 [24.2, 28.5]
1.8mm [A]	78.6 [78.0, 79.1]	73.0 [72.4, 73.6]	54.3 [53.5, 55.1]	46.7 [45.9, 47.5]	39.5 [38.7, 40.4]	32.6 [31.8, 33.4]
3mm [A]	73.9 [73.3, 74.5]	69.7 [69.1, 70.4]	47.0 [46.1, 47.8]	42.4 [41.6, 43.2]	32.1 [31.3, 32.9]	28.7 [27.8, 29.5]
V1-V3 [A]	72.0 [71.4, 72.6]	70.4 [69.7, 71.0]	44.6 [43.8, 45.4]	43.4 [42.5, 44.2]	30.1 [29.3, 30.9]	29.6 [28.9, 30.5]
HVC [A]	69.5 [68.8, 70.1]	63.3 [62.6, 63.9]	40.9 [40.1, 41.7]	33.7 [33.0, 34.5]	26.3 [25.6, 27.1]	20.6 [19.9, 21.2]
HVC+ [A]	78.4 [77.8, 78.9]	73.2 [72.5, 73.8]	53.9 [53.1, 54.7]	46.9 [46.1, 47.8]	38.9 [38.0, 39.7]	32.9 [32.1, 33.8]
Rand+ [A]	72.4 [71.8, 73.0]	70.1 [69.5, 70.7]	45.3 [44.5, 46.1]	43.0 [42.2, 43.9]	30.7 [29.9, 31.5]	29.3 [28.5, 30.1]

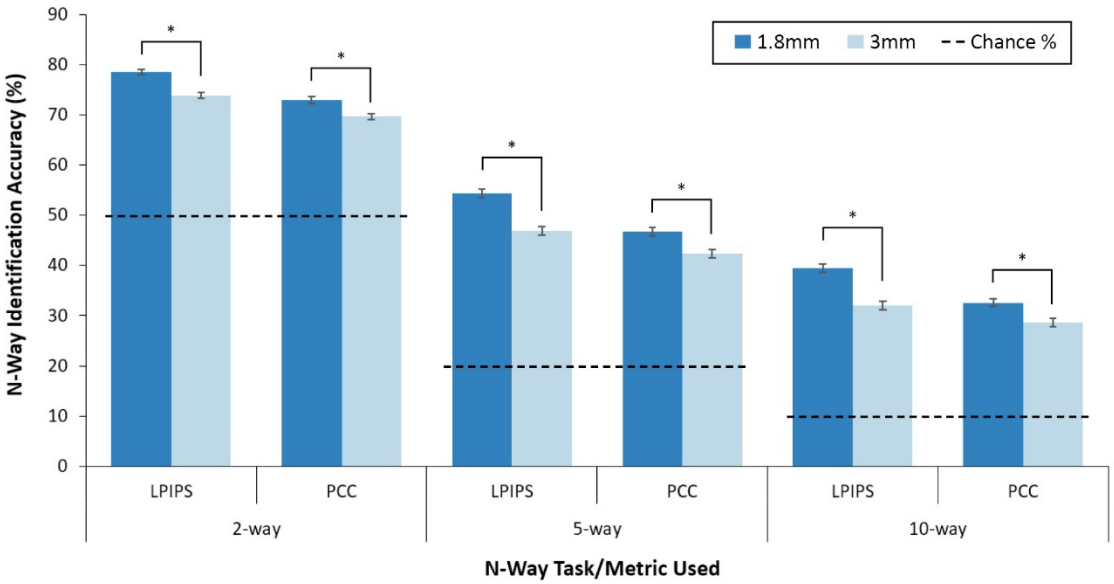
Note. Mean and 95% CI (calculated using bootstrapping; $N = 10{,}000$) of identification accuracy scores on all three n-way tasks and both metrics. Standard deviation is not reported here because they misrepresent the variability of accuracy scores[2]. That is, given the nature of the accuracy task, most reconstructions either win most trials or lose most of their respective trials. Therefore, the variability in accuracy scores in a given test set is quite large. However, the bootstrapped means suggest that mean accuracy scores over the test set are robust to repeat sampling.

[A] Accuracy is calculated with reconstructions pooled across all eight participants ($N = 6{,}976$)

[2] Full results including means, medians, standard deviations, results of statistical tests, arrays of all raw accuracy scores, and effect sizes can be found at https://osf.io/xq3cu/?view_only=2e61bd7f209e45f0ae8dfc49289f80b9

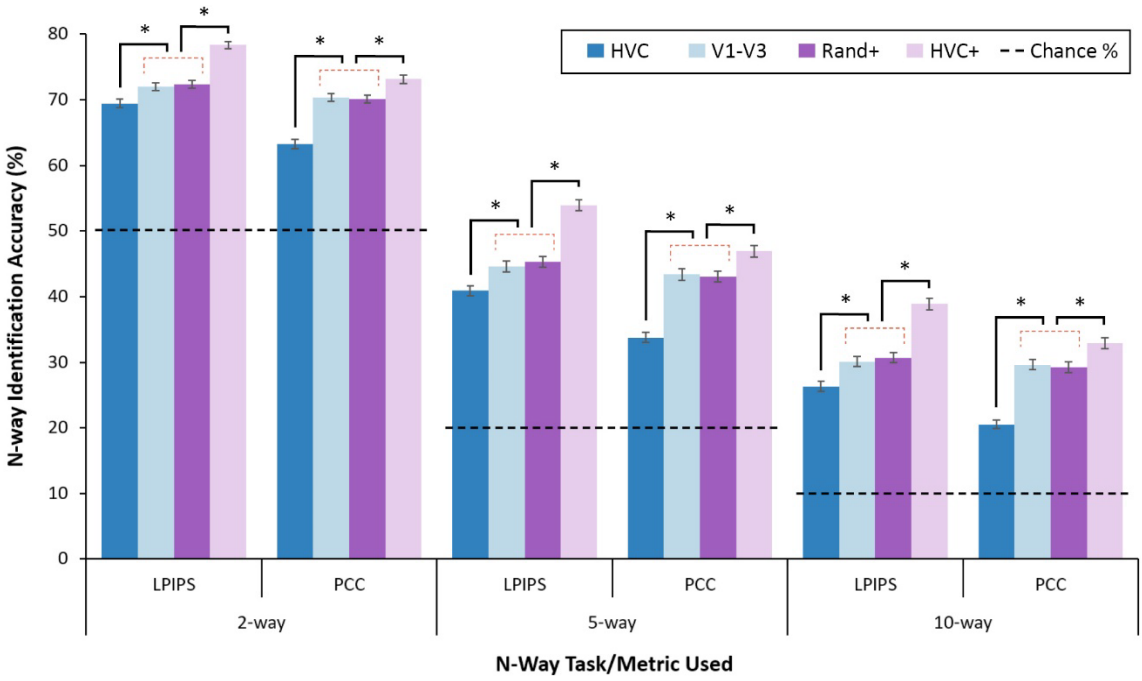
Appendix F

Full Results for Study 2 – Voxel Resolution



Note. The effect of voxel resolution is consistent across all task difficulties. The dotted line shows theoretical chance accuracy, and error bars show bootstrapped 95% CI.

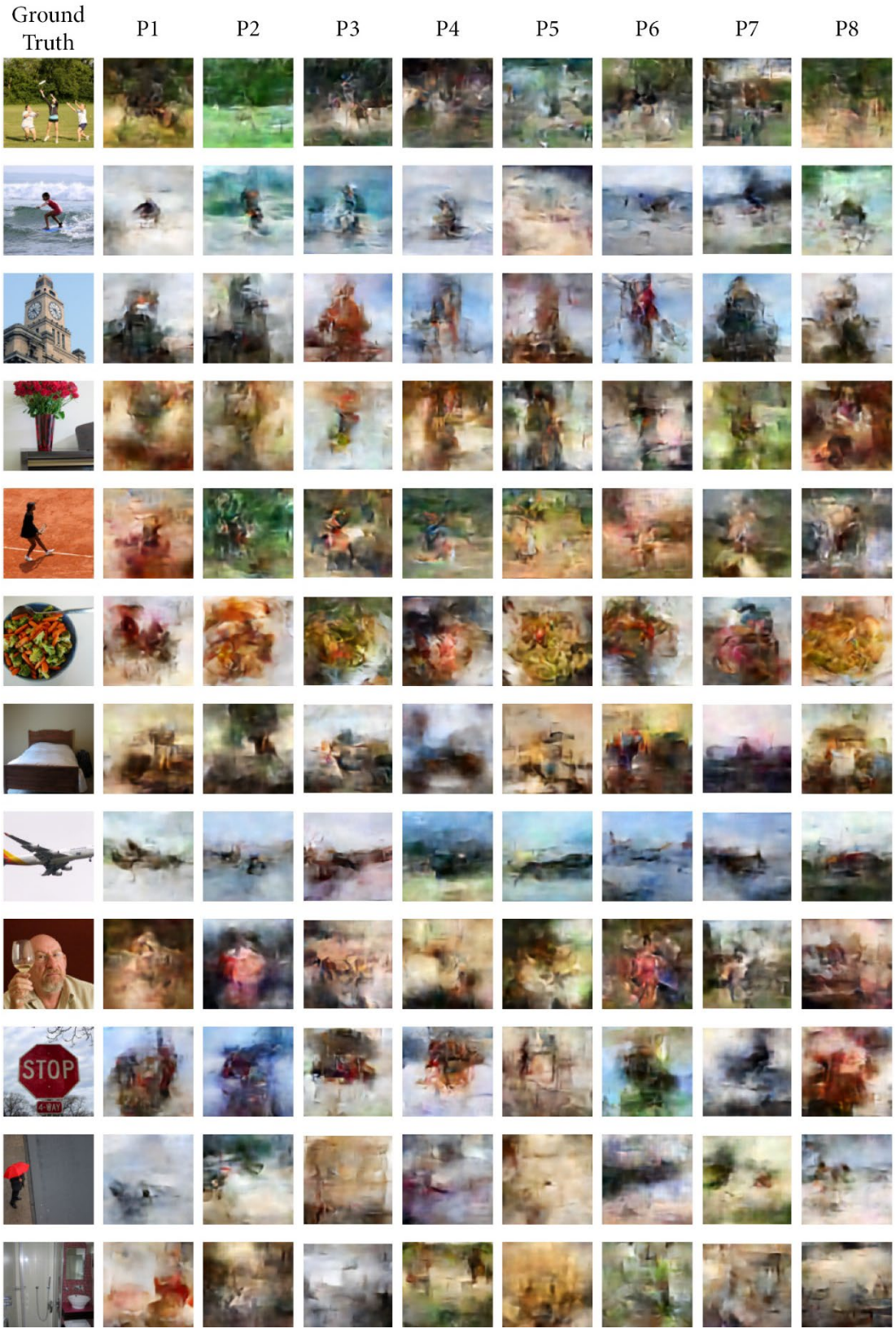
Full Result for Study 3 – ROI Specific Reconstructions



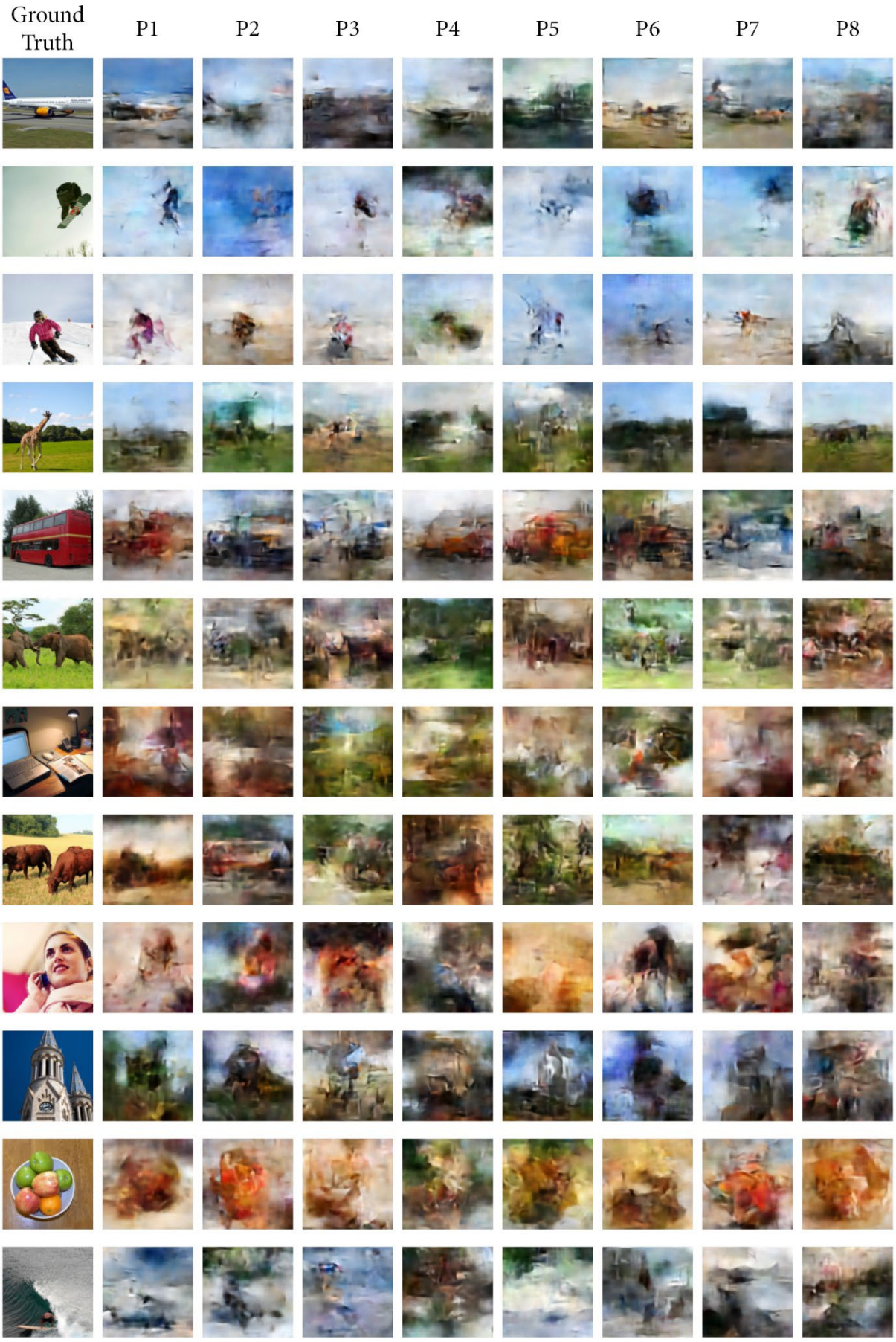
Note. All effects were consistent across n-way tasks. Firstly, that V1-V3 performs better than HVC alone. Second, no significant difference between V1-V3 and Rand+. Thirdly, HVC+ outperforms both V1-V3 and Rand+. Finally, this third effect is amplified on LPIPS relative to PCC. The red dotted line shows that the significance indicator relates to V1-V3 and Rand+. The black dotted line shows theoretical chance accuracy. Error bars show bootstrapped 95% CI.

Appendix G

More Examples of Reconstructions – All Participants (1.8 mm data)[3]

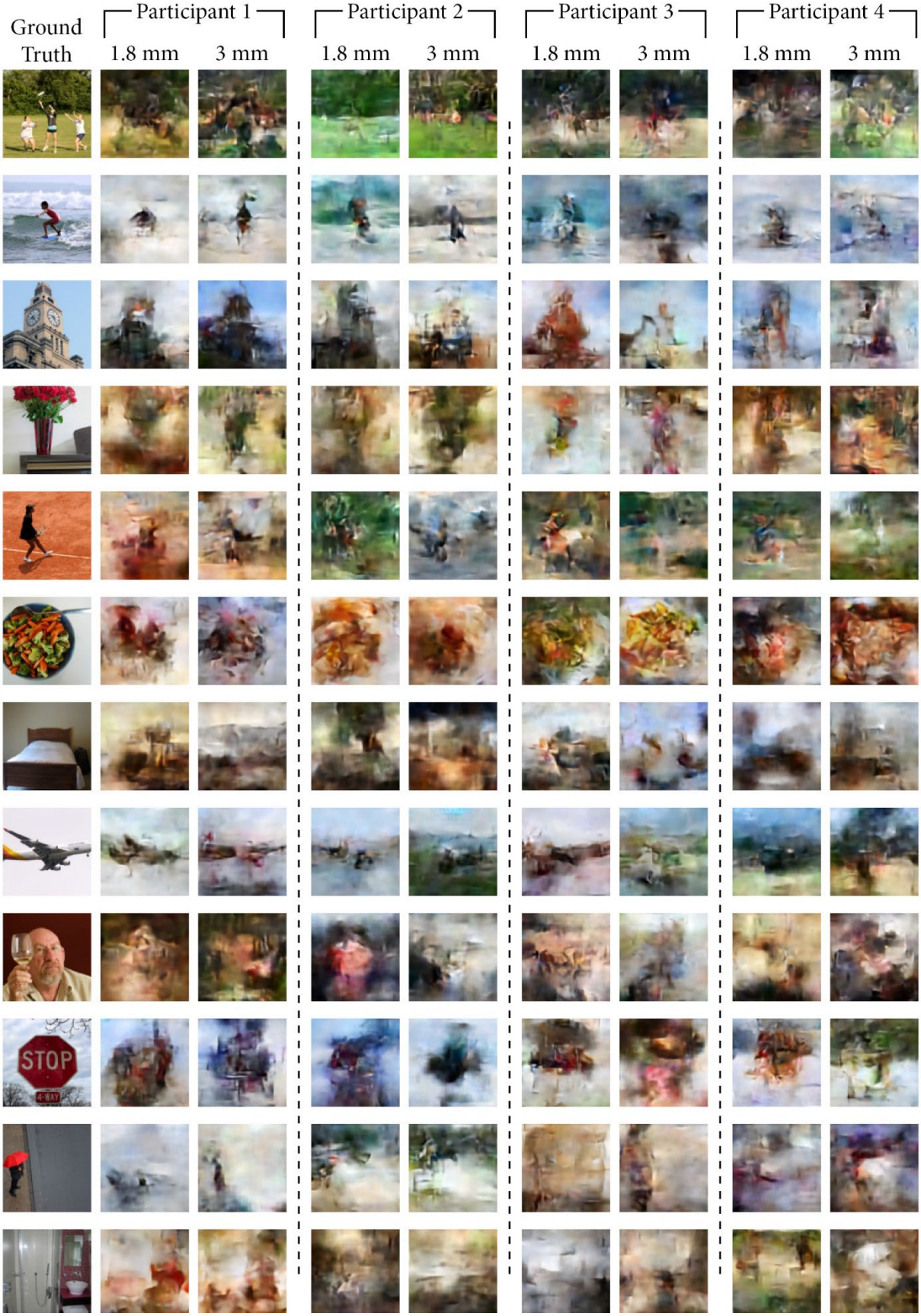


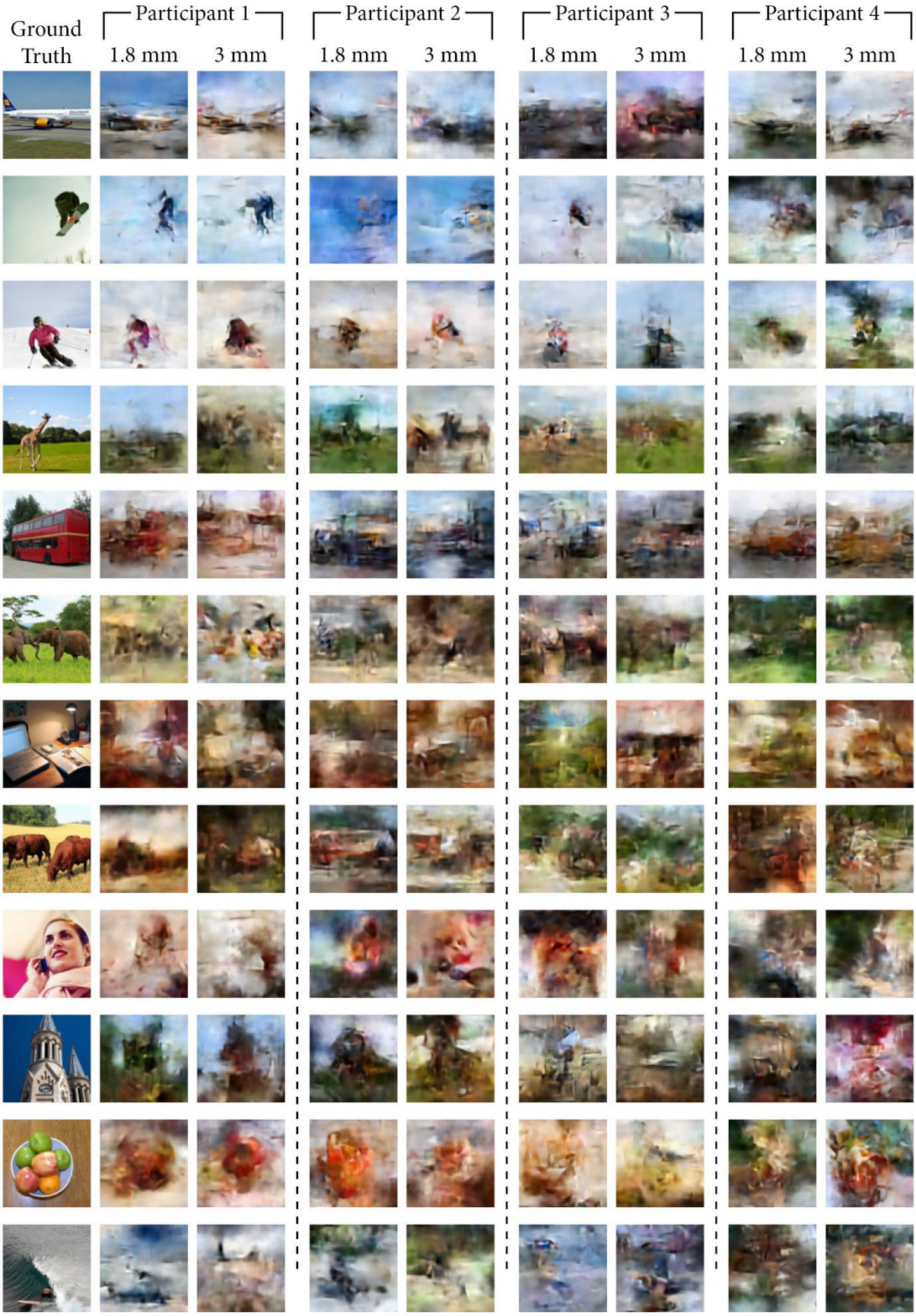
[3] All reconstructions from each 48 networks can be downloaded at https://osf.io/xq3cu/?view_only=2e61bd7f209e45f0ae8dfc49289f80b9


Ground Truth
P1
P2
P3
P4
P5
P6
P7
P8

Appendix H

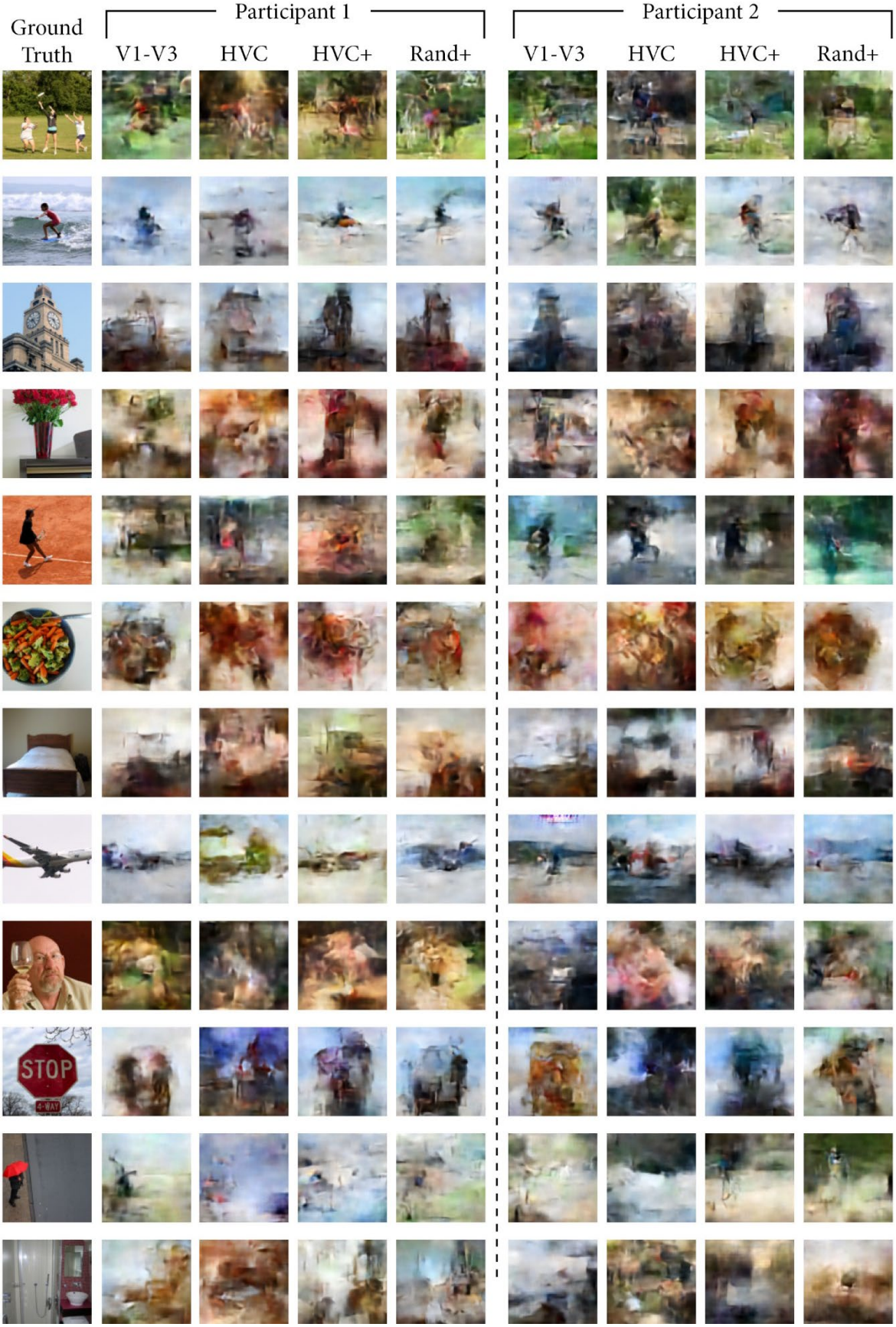
More Examples of Reconstructions – Voxel Resolution Comparison (Participants 1-4)

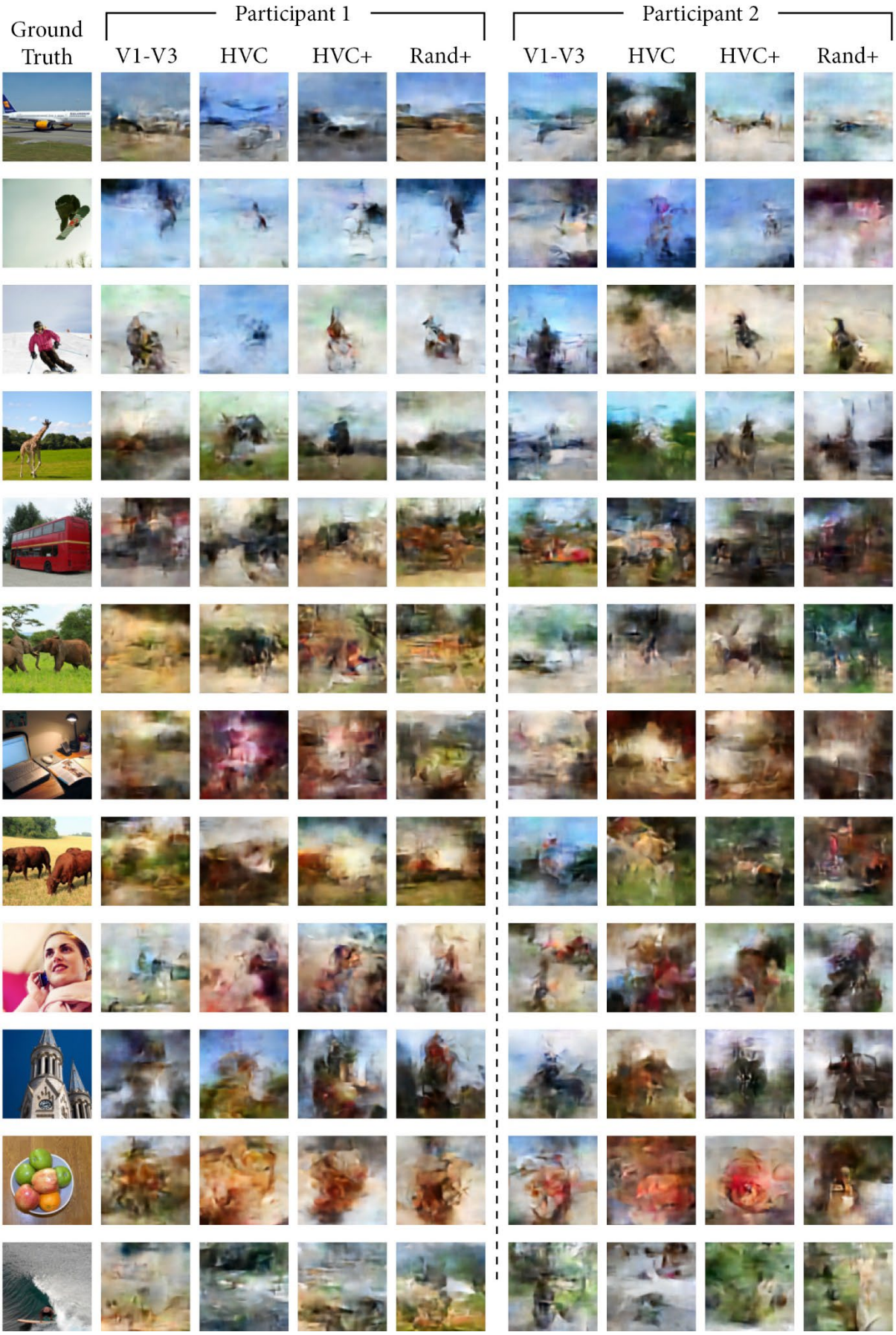



Ground Truth
Participant 1
1.8 mm
3 mm
Participant 2
1.8 mm
3 mm
Participant 3
1.8 mm
3 mm
Participant 4
1.8 mm
3 mm

Appendix I

More Examples of Reconstructions – Comparison of ROIS (Participants 1 & 2)



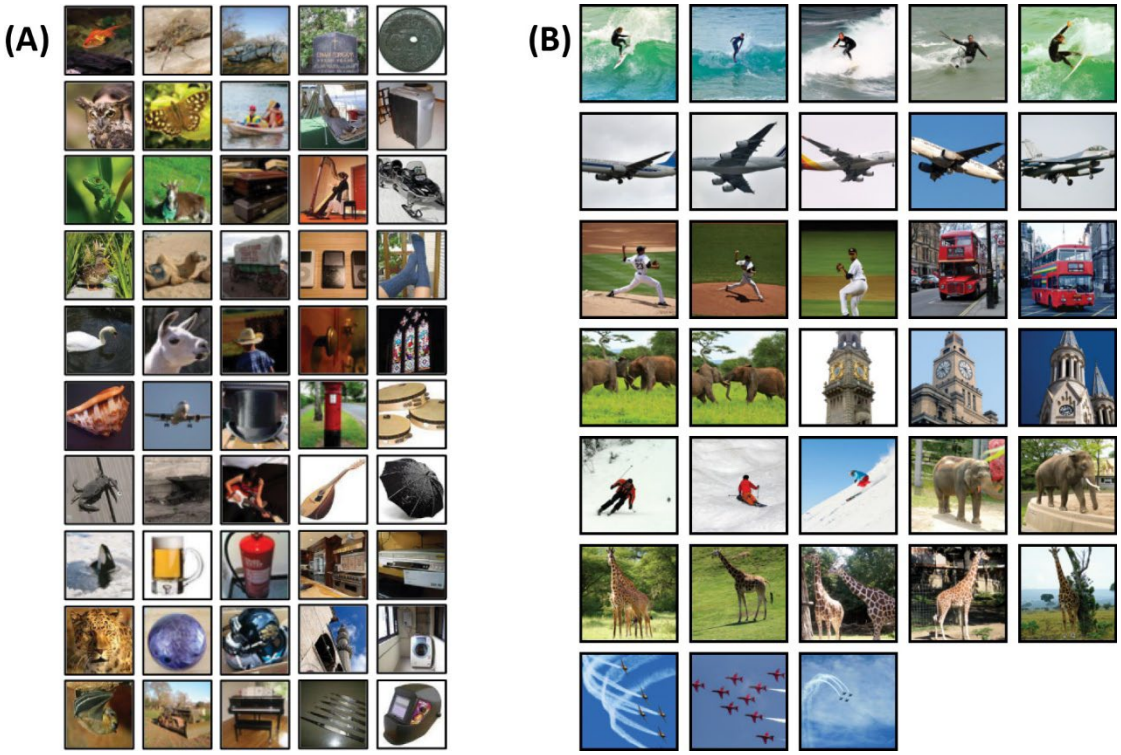

Ground Truth
Participant 1
V1-V3
HVC
HVC+
Rand+
Participant 2
V1-V3
HVC
HVC+
Rand+

Appendix J

Increased Difficulty of N-Way Task Due to Large Test Set

The current test set (N = 872) is over 17 times larger than that used with the Generic Object Decoding Dataset (N = 50; Horikawa & Kamitani, 2017a). The sheer scale of the current test set might thus increase the difficulty of each n-way identification task due to an increased likelihood of similarity between candidate and ground truth images. To demonstrate, Panel A below shows the entire test set from the Generic Object Decoding dataset. This is the most widely used dataset in previous natural image reconstruction studies (Rakhimberdina et al., 2021). In some cases, there are structural similarities between images (e.g., the black umbrella and welding helmet). However, for the most part, it appears that most of the stimuli are visually distinct with little semantic overlap.

Demonstrating the Similarities Found in Two Test Sets



By contrast, Panel B shows a small collection of images from the current test set, which share some similarities. At the very least, this indicates that there are several images

with similar semantic categories (e.g., many images of giraffes). In the most extreme cases (e.g., the two pictures with the interlocked elephants), images almost appear identical, sharing many similarities regarding colour, structure, and semantics. Therefore, given that each reconstruction was assessed with 1,000 random resamples of each n-way task, it is very likely that many of those trials included multiple candidate images that were similar structurally and/or semantically. In other words, even with close-to-perfect reconstructions of the original image, the similarities found in the test set would have made the n-way task far more difficult than those implemented in previous studies.