EXPLORING GAN LATENT SPACES: IMAGE RECONSTRUCTION FROM BRAIN ACTIVITY WITHOUT GROUND TRUTHS

To What Extent Can Image Reconstruction From Brain Activity Be Improved By Using Image Inversion Methods For GAN Latent Spaces W, W+, S, And S+, Particularly When Ground Truth Latents Are Unavailable?

RUOSHI ZHENG

Abstract

This study explores the potential of inverted GAN latent spaces for neural decoding using neural activations in visual cortex. Specifically, we focused on the ability of W, W+, S, and S+ latent spaces to capture image features to reconstruct images from brain activity, particularly when ground truth image features are unavailable. We applied W, W+, S, and S+ inversion methods to four image datasets (two face-centered, two object-centered from which one natural) and used the inverted latents to train linear decoders with corresponding brain activity. The trained decoders were used to make latent predictions and GANs were used to generate images from latents. Performance was measured by both latent similarity and image similarity. Our findings revealed that the inverted latent spaces consistently outperformed the ground truth latent space. W latents captured image features best, while W+ and S+ were most accurate in image reconstruction. We were able to accurately reconstruct natural images without ground truth latents, which shows the potential of inverted latent spaces to decode brain activity from real-life situations. Addressing challenges and opportunities posed by natural image datasets can unlock the full potential of inverted latent spaces for real-world decoding tasks. Overall, this study shows possibilities of neural decoding using inverted latent spaces, paving the way for more sophisticated decoding approaches, and fostering advancements in neurotechnology and real-world applications.

Supervisors
Thirza Dado

Affiliation
This thesis has been submitted to the MSc. Cognitive Neuroscience of The Faculty of Social Sciences, Radboud University

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Thirza Dado, for her support and guidance throughout this research journey. Her accessibility, approachability, and constructive encouragement she provided played a pivotal role in boosting my self-assurance and motivation. Her encouragement and the curiosity she evoked in me were invaluable. I also extend my heartfelt thanks to my co-supervisor, Umut G¨u¸cl¨u, for his ability to transform seemingly hopeless challenges into solvable problems and his inspiring innovative ideas. I deeply appreciate Maarten Snellen for his prompt assistance with the Snellius server, which greatly facilitated my extensive calculations. His support made the technical aspects of my research smoother and more efficient. I am thankful for the members of the Neural Coding Lab for their insightful feedback during my thesis defense. Their contributions not only enriched my research but also provided me with continuous inspiration and motivation for future endeavors. Lastly, I extend my heartfelt thanks to my boyfriend, Sjors Boutkan, for his unwavering love and support throughout this process. His encouragement and understanding were a constant source of strength and positivity. I am truly grateful for the encouragement, support, and inspiration I received from all of these individuals and their roles in making this research journey a rewarding and fulfilling one.

Introduction

Chapter 1

1.1 Background

Cognitive neuroscience plays a critical role in understanding the neural mechanisms that underlie visual perception, providing invaluable insights into how the human brain processes and comprehends visual information [1, 2]. By employing state-of-the-art neural activity measurement techniques researchers can investigate the neural correlates associated with visual perception, visual attention, object recognition, and other crucial aspects of visual processing [1–3]. These techniques allow for detailed examination of brain activity patterns that arise from visual stimuli, facilitating a deeper understanding of how the brain encodes, integrates, and represents visual inputs. By employing computational models, researchers can simulate and decode the complex processes involved in visual perception [2]. This approach not only advances our fundamental knowledge of human cognition but also has implications for neurotechnology and clinical research, fostering advancements in fields such as human-computer interaction and neurorehabilitation. The decoding of brain activity patterns relies on neural activity measurement techniques which enable the decoding of brain activity patterns associated with visual perception [3–5]. By employing functional magnetic resonance imaging (fMRI), researchers can measure changes in blood oxygenation levels, providing spatially precise information about neural activity across the brain [4]. This non-invasive technique allows for the examination of large-scale brain networks involved in visual perception tasks.

In parallel, the use of multi-unit electrode activity (MUA) offers a more direct and high-resolution assessment of neural activity by recording the electrical signals over a small number of neurons [3]. This invasive technique provides exceptional temporal and spatial resolution, enabling researchers to study the detailed dynamics of neural populations during visual perception. By using these neural activity measurement techniques, researchers can study complex patterns of brain activity and gain insights into their responses to specific stimuli.

Understanding the neural basis of visual perception requires the development of techniques for decoding the relationship between neural activity and external stimuli [1, 4]. This field of study, known as neural decoding, plays a crucial role in bridging the gap between neural activity measurement techniques and the perceptual experiences that arise from visual stimuli [2]. Neural decoding aims to unravel the neural code embedded in the recorded neural activity, seeking to decode and interpret the patterns that correspond to the external visual world. By decoding these neural representations, researchers can gain insights into the underlying neural mechanisms that enable visual perception. The potential applications and implications of neural decoding extend beyond fundamental research. Advancements in neurotechnology and brain-computer interfaces can greatly benefit from the insights gained through decoding visual stimuli from brain activity. Besides that, improved decoding methods can lead to enhanced neuroprosthetic devices, enabling, for example, individuals with visual impairments to regain visual perception [6, 7].

1.2 GAN Latent Spaces and Neural Feature

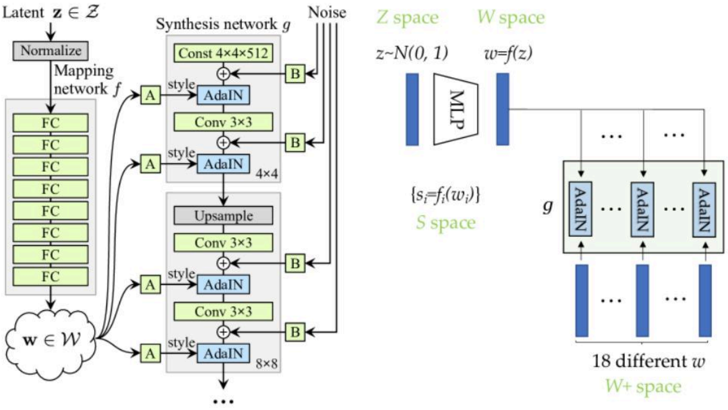
Space Neural decoding offers a computational framework to map brain responses back to sensory stimuli, requiring alignment in shared feature spaces. The process involves predicting the stimulus feature space based on the corresponding neural feature space [8]. This decoding process can be modeled as a composite function of linear and nonlinear transformations.

However, the reconstruction of visual stimuli from brain activity poses challenges due to the complexity and non-linearity of brain activity patterns. Various computational techniques, including deep neural networks and generative models, have been employed to tackle these challenges [9, 10].

Generative adversarial networks (GANs) present a promising approach for reconstructing visual stimuli from brain activity. GANs, comprising a generator and a discriminator, generate realistic images based on learned distributions [11]. In neural decoding, GANs facilitate the generation of visual stimuli from decoded brain activity patterns [12], capturing complex visual features and producing high-quality images.

StyleGAN, renowned for its ability to generate highly realistic images via disentangled latent spaces [13, 14], offers a compelling framework for neural decoding. Multiple GAN inversion methods have been developed, drawing upon the unique latent spaces offered by StyleGANs. Beyond the conventional Z space found in generic GANs, StyleGANs introduce specialized latent spaces such as W, W+, and S spaces.

These spaces can encode various image attributes and allow for manipulation of per-channel mean and variance, giving fine-grained control over image generation [13, 15] (see Figure 1.1).



(a) Style-based generator

(b) Latent spaces of StyleGAN

Figure 1.1: (a) Architecture of the style-based generator. (b) The latent spaces from which the inversion methods are constructed. The synthesis network g and AdaIN in (b) are the same as in (a). Figure edited (removed P and P+ latent spaces) and retrieved from Xia et al. (2022) [16].

We aim to explore the relationship between GAN latent spaces and neural feature spaces. StyleGAN's initial latent space Z and the intermediate latent space W play a crucial role in image synthesis and manipulation. The latent vectors z ∈ Z are sampled from a normal distribution, typically a Gaussian distribution. The vectors w ∈ W are obtained by passing the initial latent vectors z ∈ Z through a fully connected neural network (see Figure 1.1). The W latent space captures disentangled factors of variation, such as pose and lighting, crucial for image synthesis and inversion [13]. Building upon this, the W+ space incorporates different inputs per layer (see Figure 1.1) for fine-grained attribute control, enhancing attribute manipulation [17]. Meanwhile, the S space offers further disentanglement, allowing precise control over feature map variances by modulating convolution kernel weights and additional parameters employed by transformation blocks [15] (see Figure 1.1). In addition, we introduce a novel StyleGAN latent space: the S+ latent space, which combines power of both rich and disentangled style and highly controllable feature representations. More information about S+ can be found in the Method section.

1.3 GAN Inversion Methods

To retrieve the latent representations in the aforementioned latent spaces, we employed image inversion methods provided by the original repositories of StyleGAN. GAN inversion seeks to invert a given image into the latent space of a pretrained GAN model, allowing faithful reconstruction by the generator. It maps a provided real image to the the preferred latent space, resulting in the the acquisition of the latent code. Subsequently, the reconstructed image is generated by the generator of the GAN model (see Figure 1.2). By applying GAN inversion methods, we can deduce image latents in the preferred latent space from observed images. Detailed information about the GAN latent spaces and their inversion procedure can be found in the Method section.

1.4 Reconstruction without Ground Truth Latents

The other main contribution of this research is the exploration of stimulus reconstruction from brain activity without access to ground truth latents. Ground truth latents are represented by StyleGAN's initial latent space Z and are only available for images originally synthesized by the model. Natural images do not have these ground truth images available and need inversion methods to be mapped into a manipulable GAN latent space.

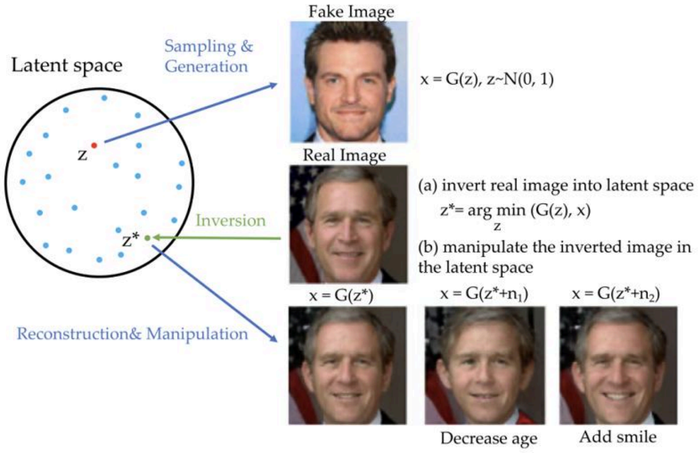


Figure 1.2: Process of GAN inversion, retrieved from Xia et al. (2022) [16]. For the scope of our study, we do not take into account image manipulation.

Once mapped into a GAN latent space, manipulation of certain attributes becomes possible while retaining other attributes for image generation. If the inversion methods successfully capture image features related to neural representations, it signifies the potential for GAN-based decoding in nonsynthesized images.

1.5 Research Objective and Question

In this study, we will explore GAN latent spaces to enhance image reconstruction from brain activity, particularly when ground truth latents are unavailable. We will answer the research question: "To what extent can image reconstruction from brain activity be improved by using image inversion methods for GAN latent spaces W, W+, S, and S+, particularly when ground truth latents are unavailable?" To achieve this, we employ inversion techniques to retrieve the latent representations of images. By leveraging the power of StyleGAN in image generation, combined with inversion techniques, we aim to decode and reconstruct meaningful visual presentations from brain activity. We employ two state-of-the-art generative models, StyleGAN3 [18] and StyleGAN-XL [19], for image generation. StyleGAN3 excels in generating highly realistic and diverse images from a specific category (e.g. faces), capturing fine-grained details, and preserving global coherence. On the other hand, we employ StyleGAN-XL, an extended version of StyleGAN, to handle more complex and diverse datasets consisting of more than one category. Besides that, we employ four inversion methods: W Space Inversion, W+ Space Inversion, S Space Inversion, and S+ Space Inversion. By leveraging these two GAN architectures with the four inversion methods, we aim to achieve accurate image generation and inversion performance across different datasets.

The exploration of the W, W+, S, and S+ latent spaces in image inversion represents a novel approach in the field of neural decoding. By evaluating the effectiveness of these latent spaces in reconstructing images from brain activity without access to ground truth latents, we aim to determine which latent space yields the best results.

This research will contribute to a deeper understanding of the capabilities and limitations of StyleGAN's latent spaces in generating meaningful visual interpretations from brain activity, ultimately advancing the field of neural decoding and facilitating future applications in neurotechnology and brain-computer interfaces.

Method

Chapter 2

2.1 Data

2.1.1 Mapped/Encoded Brain Activity Datasets

2.1.1.1 Mapped Hyper-aligned fMRI

We used a functional magnetic resonance imaging (fMRI) dataset, consisting of brain responses recorded from two healthy participants (S1: 30-year-old male; S2: 32-year-old male) from Dado et al. [20]. This dataset corresponds to the HYPER image dataset described in section 2.1.2. Details about fMRI recording acquisition and preprocessing steps can be found in [20]. The resulting brain maps were represented in terms of z-scores. A voxel mask was created by selecting the most active 4096 voxels from the training set based on amplitude. Hyper-alignment was used to capture the shared neural information across participants. FMRI data was mapped by adding a dense layer at Progessive Growing GAN (PGGAN) to map brain activity to latent vectors that is fit with ordinary least squares. Due to this mapping, the fMRI data was enabled for comparison with GAN latent spaces. The details of the implementation are beyond the scope of this research. For more information, see [20].

2.1.1.2 Encoded Macaque MUA

The multi-unit micro-electrode activity (MUA) dataset, collected from a macaque (male, 7 years old) through chronically implanted electrode arrays, was used to investigate the Brain2GAN Faces, Brain2GAN Objects and THINGS datasets (described in section 2.1.2). The dataset was obtained from Dado et al. [21]. Neural activity was recorded in three visual cortex (VC) regions: V1 (7 arrays), V4 (4 arrays), and the face-selective inferior temporal cortex (IT) (4 arrays), using 15 electrode arrays, yielding a total of 960 channels. The recorded data was preprocessed by normalizing the responses based on a previously established method from Super and Roelfsema [3]. Details about the experiment and data acquisition can be found in [21].

MUA data was fit by an encoding model based on StyleGAN's w latent representation. This was done for each individual microelectrode unit. Due to this mapping, the MUA data was enabled for comparison with GAN latent spaces. The details of the implementation are beyond the scope of this research. For more information, see [21].

2.1.2 Image Datasets

A schematic overview of the image dataset characteristics can be found in Table 2.1.

2.1.2.1 HYPER

The HYPER dataset used in this study was collected and provided by a previous study from Dado et al. [20]. The dataset consists of face images with a resolution of 1024×1024 pixels. These face images were synthesized using a pretrained generator network from a Progressive GAN (PGGAN) model [22]. The generator network generates the images based on randomly sampled 512-dimensional latent vectors from a standard Gaussian distribution. Each synthesized face image is then cropped and resized to a resolution of 224×224 pixels for further processing and analysis. The final dataset consists of a training set of 1050 face images and a test set of 36 synthesized faces. It is important to note that none of the face images used in this research are of real people; they are instead synthesized by a generative model trained on the CelebFaces Attributes Dataset (CelebA), which comprises a large-scale collection of more than 200,000 celebrity images [23].

2.1.2.2 Brain2GAN Faces

The Brain2GAN Faces dataset used in this study was obtained from earlier research from Dado et al. [21]. The dataset consists of photorealistic face images with a resolution of 1024×1024 pixels. These face images were synthesized using a generator network based on the StyleGAN3 architecture, which was pretrained on the high-quality Flickr-Faces-HQ (FFHQ) dataset [13]. The generator network takes as input 512-dimensional z-latent vectors randomly sampled from a standard Gaussian distribution. During synthesis, the z-latents are transformed using learned affine transformations and adaptive instance normalization. The final dataset comprises a training set of 4000 face images and a test set of 100 synthesized faces, each averaged over twenty repetitions. It is important to note that the face images used in this research are not of real individuals but are instead generated by a generative model trained on the FFHQ dataset [13].

2.1.2.3 Brain2GAN Objects

The Brain2GAN Objects dataset used in this study was obtained from the same paper as the Brain2GAN Faces dataset [21]. The dataset comprises RGB images of natural objects with a resolution of 512×512 pixels. These images were synthesized using the StyleGAN-XL architecture, which is designed to handle larger and less-structured datasets. The training set of the dataset consists of 4000 synthesized images, while the test set contains 200 images. The training set images were generated from randomly sampled z-latent vectors, which were mapped to w-latent vectors truncated at 0.7 to ensure image quality and diversity. For the test set, the average w-latent vector of each category was used to generate a single image per category. The images are generated from 200 different classes selected from the Tiny ImageNet dataset, with each class represented by twenty training set stimuli and one test set stimulus. The category labels for the dataset can be found in the supplementary materials of the paper [21].

2.1.2.4 THINGS

The THINGS image dataset used in this study was originally introduced by Hebart et al. [24] as a comprehensive collection of 1,854 diverse object concepts, comprising a total of 26,107 high-quality naturalistic images.

The object concepts were categorized into 27 high-level categories, which effectively captured a substantial portion of the concepts with minimal overlap. Validation of these categories and the corresponding object images was conducted by assessing their relationship to representations in a semantic embedding and a deep convolutional neural network. This analysis provided evidence supporting the meaningful association of the categories with their use in language and affirmed that the object images represented distinct categories while showing variations in fundamental visual properties [24]. In this study, a subset of the THINGS dataset consisting of the first 12 images per concept was employed, resulting in a training set of 22,248 images and a test set comprising 100 images. It is noteworthy that the THINGS dataset consists of natural images, and thus, ground truth latents associated with the image features were not available for analysis.

Dataset	Type	Original Dataset	Activity	Train	Test	Network	Latent Spaces
HYPER	Synthesised Faces	CelebA	fMRI	1050	36	StyleGAN3	W_{gt}, W, $W+$, S, $S+$
Brain2GAN Faces	Synthesised Faces	FFHQ	MUA	4000	100	StyleGAN3	W_{gt}, W, $W+$, S, $S+$
Brain2GAN Objects	Synthesised Objects	Tiny ImageNet	MUA	4000	200	StyleGAN-XL	W_{gt}, W, $W+$, S, $S+$
THINGS	Natural Objects	THINGS	MUA	22248	100	StyleGAN-XL	W, $W+$, S, $S+$

Table 2.1: Characteristics of image datasets used in this study. The THINGS dataset is the largest dataset and the only dataset consisting of natural images instead of synthesised images. Because of the latter, the ground truth W latents were not available for the THINGS dataset.

2.2 GAN Models

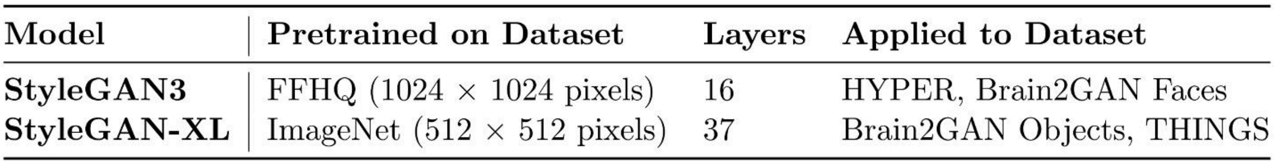
In this section, we present an overview of the two pretrained GAN models employed in our study: StyleGAN3 and StyleGAN-XL. The pretrained StyleGAN3 model was used for the HYPER and Brain2GAN Faces datasets. The pretrained StyleGAN-XL model was used for the Brain2GAN Objects and THINGS datasets. A schematic overview of the characteristics of the used GAN models can be found in Table 2.2.

2.2.1 StyleGAN3 FFHQ

For our experiments on face-centered datasets (HYPER, Brain2GAN Faces), we leveraged the StyleGAN3 network which was pretrained on the 1024×1024 FFHQ dataset [13]. This pretrained model can be obtained from the StyleGAN3 NVIDIA GPU Cloud model catalog. We use a pretrained version that incorporates rotation and translation equivariance. The FFHQ dataset consists of 70,000 high-quality PNG images with a resolution of 1024×1024 pixels [13]. It offers a wide array of variations in terms of age, ethnicity, and image background, making it well-suited for capturing diverse image synthesis scenarios. Moreover, the dataset encompasses various accessories such as eyeglasses, sunglasses, hats, and more, further enhancing its coverage of real-world complexities.The images within the dataset are sourced from the popular photo-sharing platform Flickr and have been automatically aligned and cropped using the dlib library. To ensure compliance with licensing requirements, only images under permissive licenses were included in the dataset. Additionally, a series of automatic filters were applied to remove undesirable images, and any remaining ambiguous cases were carefully pruned with the assistance of Amazon Mechanical Turk.

2.2.2 StyleGAN-XL ImageNet

We used the pretrained network of StyleGAN-XL on ImageNet [25] for the object-centered datasets (Brain2GAN Objects, THINGS). This ImageNet dataset used for this pretrained network consisted of images with a resolution of 512×512 pixels. The pretrained model can be downloaded via this link. The ImageNet dataset contains a collection of diverse images, covering a wide range of object categories and variations. It consists of over a million high-quality images distributed across a total of 1000 distinct object classes [25]. The dataset has played an important role in advancing computer vision research and has been widely adopted as a standard benchmark for imagerelated tasks [26]. The pretrained version of StyleGAN-XL on the ImageNet dataset has enabled the model to generate detailed and realistic images.



Model	Pretrained on Dataset	Layers	Applied to Dataset
StyleGAN3	FFHQ (1024 × 1024 pixels)	16	HYPER, Brain2GAN Faces
StyleGAN-XL	ImageNet (512 × 512 pixels)	37	Brain2GAN Objects, THINGS

Table 2.2: Characteristics of GAN models used in this study.

2.3 GAN Inversion and Latent Spaces

In this section, we delve into the details of four distinct GAN inversion techniques: W Space Inversion, W+ Space Inversion, S Space Inversion, and S+ Space Inversion. By inverting images to these spaces, we aim to investigate the effectiveness of these inversion methods in combination with brain activity. Specifically, we seek to train a linear model using the inverted images paired with corresponding brain responses, allowing us to predict latent spaces and generate images from brain activity. A schematic overview of the latent spaces can be found in Table 2.3. The timing of our inversion step and a schematic overview of the experimental setup can be found in Figure 2.1.

2.3.1 Inversion Algorithm Overview

The GAN inversion process aims to reconstruct images from brain activity by iteratively adjusting latent codes until the synthesized image from the generator closely matches the input image. The algorithm involves the use of a pretrained StyleGAN generator network and VGG16 as feature detector network for extracting features from both the target and synthesized images. The pretrained VGG16 network is accessed from the NVIDIA API model library. Direct optimization is employed to refine both the latent vector w and the noise vector n in reconstructing the input image x, guided by the Learned Perceptual Image Patch Similarity (LPIPS) perceptual loss function [27]. Optimizing the noise vector n with a noise regularization term enhances the inversion process by preventing the noise vector from containing crucial information. Consequently, once wp is determined, the influence of n values on the final visual appearance decreases [13]. The overall optimization function is formulated as follows:

$$(w_p, n) = \arg\min_{w,n} \mathcal{L}_{LPIPS}(x, G(w, n; \theta)) + \lambda_n \mathcal{L}_n(n)$$

where G(w, n, θ) represents the generated image using a generator G with weights θ. Notably, StyleGAN's mapping network (converting from ζ to γ) is not utilized. Here, LLP IP S denotes the perceptual loss, Ln represents a noise regularization term, and λn is a hyperparameter. An Adam optimizer is set up to optimize the latent vector and the noise inputs. Once the optimization process is complete, it returns the optimized inverted GAN image latent corresponding to the given target image. Besides optimization-based inversion, as applied in this study, other methods for GAN inversion include learning-based inversion and a hybrid approach. Learning-based inversion involves training an encoder network to map images to latent space, optimizing the reconstruction process to closely resemble the original images. Optimization-based inversion directly minimizes pixel-wise reconstruction loss through back-propagation, refining the latent code to reconstruct the input image faithfully (= similarity between the real image and the generated one [16]). The hybrid approach combines both strategies, utilizing an encoder to generate an initial latent code, which is then optimized further. While our primary focus was on optimization-based inversion, we also explored one hybrid method, known as pivotal tuning inversion (PTI) [28]. PTI is a generator-tuning technique which uses an initial latent code serves as a pivot, which is slightly adjusted to the pretrained generator to ensure accurate reconstruction of the input image. PTI is able to align an image from an external domain with an internal latent code to achieve faithful reconstruction. While our primary focus was on optimization-based inversion, we also explored the hybrid method PTI. For further details and experiments regarding PTI, see Appendix 6.1.

2.3.2 W Space Inversion

W Space Inversion begins by sampling latent vectors z ∈ Z where after StyleGAN transforms the native z latent vectors into mapped style vectors w using a nonlinear mapping network f, which is represented by an 8-layer multilayer perceptron (MLP) (see Figure 1.1a). This intermediary latent space is referred to as the W space. Through the mapping network and affine transformations, the W space in StyleGAN exhibits a higher degree of feature disentanglement compared to the Z space. The algorithm iteratively optimizes the latent vector and noise inputs to minimize the distance between the synthesized and target images. The applied code can be found in the w projector.py file of the PTI Github page.

Once the optimization process is complete, it returns the optimized inverted 512-dimensional w-latent corresponding to the given target image. This wlatent is copied across the mapping network's number of input layers (16 for StyleGAN3, 37 for StyleGAN-XL) to match the structure of the generator's input.

2.3.3 W+ Space Inversion

W+ Space Inversion extends W Space Inversion by projecting images into an extended latent space W+. This method concatenates identical w vectors from each layer in the StyleGAN architecture, creating a w+ nd-array. Optimization is performed on this array, and noise regularization is applied to the noise buffers in each layer of the StyleGAN network. Whereas W Space Inversion copies the optimized 512-dimensional latent vector across the GAN input layers, W+ Space Inversion directly inputs the optimized 512 × nr GAN input layers-dimensional w+ latent vector to each of the generator's layers via AdaIN (see Figure 1.1b). The applied code can be found in the w plus projector.py file of the PTI Github page.

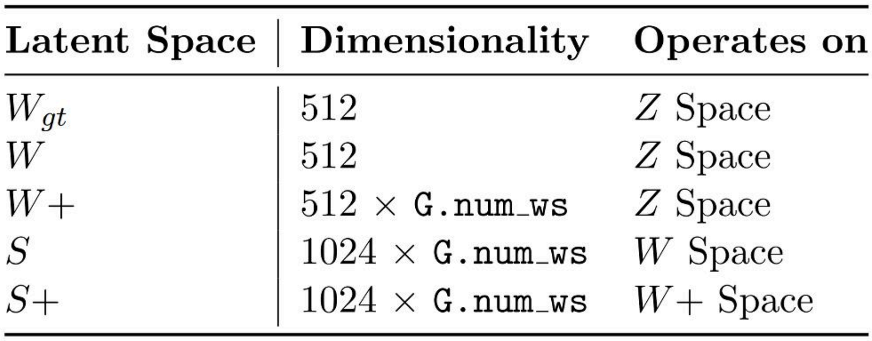
The extended latent space W+ introduces increased feature entanglement due to the input of the optimized 512 × nr GAN input layers-dimensional latent vector instead of copies of the same optimized 512-dimensional latent vector as in W Space Inversion. Besides using all inputs simultaneously, we explored feature entanglement in W+ space. We studied both distinct and cumulative vectors in W+ space for image generation with StyleGAN3 and StyleGAN-XL. More details about our implementation of feature entanglement and its implications for image generation and manipulation can be found in Appendix 6.2.

2.3.4 S Space Inversion

S Space Inversion leverages the disentangled style representations, such as in Appendix 6.2, but now provided by the S latent space of StyleGAN. Latent vectors in W Space go through a affine transformation process to be converted to corresponding 1024-dimensional style vectors in S Space. This enables precise control over the style attributes of the generated images. Dimensions of each block are modified when passing the style vectors to ensure compatibility with StyleGAN's 512-dimensional feature space requirements. Although S Space Inversion utilizes the 512-dimensional feature space during the image synthesis process, the full 1024-dimensional feature space is used when training the linear decoder (described in section 2.4). This enables the decoder to learn and map the neural representations to the rich and disentangled style attributes captured by the S latent space.

2.3.5 S+ Space Inversion

S+ Space Inversion is an exploratory technique that operates on the w+ latents, similar to W+ Space Inversion. It converts latent vectors in W+ Space to corresponding style vectors in S+ Space, allowing for enhanced image manipulation. Similarly to S Space Inversion, dimensionality is modified when passing the s+ vectors through the StyleGAN synthesis process, reducing the feature space from 1024 to 512 dimensions. The linear decoder (described in section 2.4) is again trained on the full 1024-dimensional feature space. As a novel concept, S+ Space Inversion requires further evaluation and experimentation to understand its capabilities and limitations fully.



Latent Space	Dimensionality	Operates on
W_{gt}	512	Z Space
W	512	Z Space
$W+$	$512 \times \texttt{G.num\_ws}$	Z Space
S	$1024 \times \texttt{G.num\_ws}$	W Space
$S+$	$1024 \times \texttt{G.num\_ws}$	$W+$ Space

Table 2.3: Characteristics of latent spaces used in this study. The dimensionality is shown for each latent space, where 512 or 1024 is the dimensionality per latent vector and G.num ws is the number of StyleGAN synthesis layers. For StyleGAN3 this is 16 and for StyleGAN-XL 37.

Since StyleGAN's architecture expects a 512 × G.num ws dimensionality input, the 512-dimensional latent vectors of Wgt and W are copied G.num ws times before feeding it to the generator. The 1024-dimensional latent vectors of S and S+ are reduced to 512 dimensions by applying affine transformations.

2.4 Linear Decoder

We used linear mapping to evaluate whether the GAN latent- and neural representation effectively encode the same stimulus properties [10]. More complex nonlinear transformations would not be appropriate to use for this task since nonlinearities will fundamentally alter the underlying representations [21]. To train the linear decoder, we employed scikit-learn's linear model module to conduct linear regression with the mapped/encoded brain activity (see Section 2.1.1) and GAN image latent as input [29]. We used four types of mapped/encoded brain activity: mapped fMRI from HYPER, encoded MUA from Brain2GAN Faces, encoded MUA from Brain2GAN Objects, and encoded MUA from THINGS. We used five different inputs for GAN image latents: ground truth GAN image latents Wgt (except for the THINGS dataset) and inverted GAN image latents from spaces W, W+, S, and S+. The training step of the linear decoder is indicated with pink arrows in Figure 2.1. For the fMRI data, we mapped brain activity to latents representing the overall visual cortex (VC), as visual areas were not individually measured in the study we retrieved the fMRI data from [20]. The latents captured the global VC responses associated with the presented images. On the other hand, for the study from which we retrieved the MUA data from [21], we were able to map brain activity to latents representing both global and individual VC areas: full, V1, V4, and IT. Each visual area was represented by its corresponding dimensions in the latents. This allowed us to investigate the relationship between the visual areas and the corresponding image features more precisely. To construct the input for the linear decoder, we leveraged two types of GAN image latents. For datasets where ground truth features were available (HYPER, Brain2GAN Faces, Brain2GAN Objects), we used the corresponding ground truth image latents as one set of inputs. These ground truth image features provided a direct representation of the underlying image properties.

In addition to the ground truth image latents, we applied our image inversion methods to the original images from these datasets (including THINGS). By inverting the images using the image inversion methods, we obtained the inverted image latents. These inverted image latents aimed to capture the latent space dynamics and explore the potential benefits of using the inverted image features for predicting brain activity.

Using the training data consisting of mapped brain activity (fMRI or MUA) to latents paired with the corresponding image latents, we fit a linear regression model to predict brain activations in the test sets. The prediction step is indicated with the orange test arrow in Figure 2.1. Each latent pair served as a training set, enabling the linear decoder to learn the mapping between image features and brain activity. This approach allowed us to evaluate the effectiveness of both the ground truth image features and the inverted image features in predicting brain activity.

By comparing the performance of the linear decoder using different sets of image latents (ground truth and inverted) and brain activity (fMRI and MUA), we aimed to assess the potential of image inversion methods and the role of visual cortex regions in predicting brain activity.

This analysis not only provided insights into the relationship between neural representations and image features but also enabled us to explore the predictive power of inverted image features and the contributions of different visual areas, providing a comprehensive understanding of the mapping between brain activity and image properties.

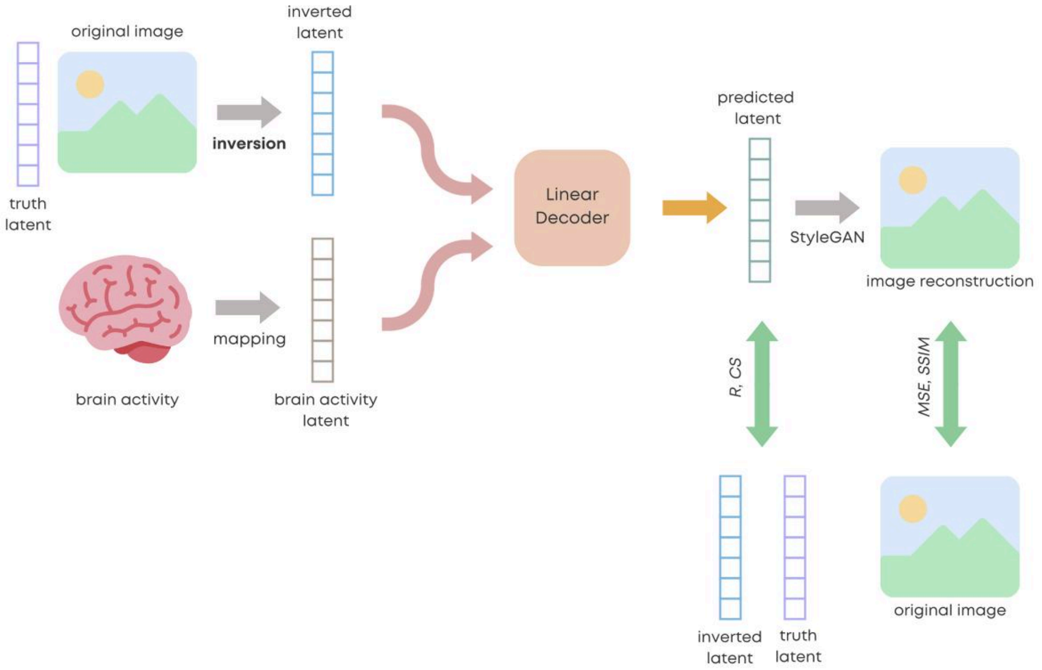


Figure 2.1: Overview of experimental setup. The inversion step is presented in bold. For the THINGS dataset, the truth latents (Wgt) were unavailable. Therefore, we performed the R and CS evaluation metrics solely with the inverted latents (W, W+, S, S+) for THINGS. For the other datasets, we performed the R and CS evaluation metrics with both the inverted latents and the truth latents. Grey arrow = mapping from image to latent or other way around, pink arrows = training input for linear decoder, orange arrow = testing the trained linear decoder model with brain activity as input, green arrows = evaluation for latent similarity and image similarity.

2.5 Evaluation

In order to assess the quality of our predictions, we compared the latent vectors and the reconstructed images with the ground truth images. These evaluation methods provide insights into the efficacy of the linear decoder and the relationship between the latent spaces and the resulting image reconstructions. The evaluation steps are indicated with the green arrows in Figure 2.1.

2.5.1 Latent Similarity

We obtained predicted latents through the trained linear decoder which learned the mapping between the image latents and the corresponding brain activity during the training phase. The predicted latents represented the latent space dynamics associated with the test images. These predicted latents were also in the form of image features, allowing us to use them as inputs for StyleGAN to reconstruct images.

Quantitative evaluation metrics were employed to provide objective measures of the performance of our approach. We used Pearson Correlation (R) and Cosine Similarity (CS) to compare the predicted brain test latents with the inverted test image latents (with the exception of the ground truth). These metrics provide a measure of the correlation and similarity between the two sets of latents. A higher R or CS score indicates a stronger linear or cosine relationship between the predicted latents and the image properties, suggesting a better alignment between the predicted and inverted/ground truth test latents.

2.5.2 Image Reconstruction Accuracy

To further evaluate the quality of our predictions, we used the generated images obtained from the predicted latents. By comparing the reconstructed images with the ground truth images from the test set, we could quantitatively and qualitatively assess the accuracy and fidelity of the predicted latents. For the evaluation of the reconstructed images, we used Mean Squared Error (MSE) and Structural Similarity Index (SSIM) as quantitative evaluation metrics. MSE quantifies the average squared difference between the pixel values of the reconstructed images from the predicted latents and the ground truth test images. A lower MSE value indicates a smaller average difference between the reconstructed and ground truth images, indicating higher reconstruction accuracy. On the other hand, SSIM measures the structural similarity between the two sets of images, taking into account luminance, contrast, and structural information. Details of these terms can be found in [30]. SSIM values range from 0 to 1, with 1 indicating a perfect match between the reconstructed and ground truth images.

2.5.3 Curse of Dimensionality

During evaluation, we encountered a challenge with the R and CS scores between the predicted latents and test latents in the W+, S, and S+ latent spaces. It appeared that their R and CS scores were unrealistically high compared to their MSE and SSIM values. This discrepancy was attributed to the high dimensionality of the feature space, which can lead to the ”Curse of Dimensionality”, where spurious correlations or patterns may be found by chance. In high-dimensional spaces, there is an increased risk of overfitting, where the model memorizes the training data without capturing the underlying patterns, resulting in inflated evaluation metrics but poor generalization to new data.

To address these issues, we employed Principal Component Analysis (PCA) to reduce the dimensionality of the W+, S, and S+ latent spaces to 512- dimensional vectors. We used the scikit-learn library’s implementation of PCA [29] in our analysis. We specified the number of components to be retained as 512 using the n components parameter, fit the PCA model on the test latents to learn the principal components, and transformed the predicted latents using the fitted PCA model. To ensure the dimensionality reduction was successful, we verified the shapes of the transformed arrays (n test images × 512).

By applying PCA in this manner, we effectively reduced the dimensionality of the latent spaces, retaining the most informative components while discarding less important ones. This dimensionality reduction enabled a more robust evaluation of the model’s performance by enhancing our ability to assess the correlation and similarity between predicted and test latents.

Results

Chapter 3

3.1 Evaluation Metrics

The evaluation of the linear decoder and the inversion techniques involved the calculation of various metrics, including Pearson Correlation Coefficient (R), Pearson P value (P), Cosine Similarity (CS), Mean Squared Error (MSE), and Structural Similarity Index (SSIM). The evaluation was conducted for all image datasets, latent spaces, and VC region combinations.

For each predicted latent from brain activity and its corresponding (inverted) test latent, R, P, and CS were computed. R measures the linear relationship between two vectors. It quantifies the strength and direction of the association between the predicted latent and the test latent. The Pvalue determines whether the observed correlation is statistically significant or occurred by chance. CS measures the similarity between two vectors based on the cosine of the angle between them. It indicates how close the predicted latent is to the test latent in terms of direction.

Additionally, for each reconstructed image and its corresponding test image, the MSE and SSIM were calculated. The MSE measures the average squared differences between the pixel values of the reconstructed image and the test image. It quantifies the overall dissimilarity between the two images. The SSIM compares the structural information between the reconstructed image and the test image. It evaluates the similarity in terms of luminance, contrast, and structure.

To ensure a comprehensive evaluation, these metrics were computed for all predicted latents and reconstructed images, considering all test latents and -images in the respective datasets. For each dataset, latent space type, and VC region combination, the means and standard errors of these metrics were calculated.

The mean values (μ) provide an indication of the average performance across all image pairs, while the standard errors (σx^{-}) offer insights into the variability and precision of the results. By considering these aggregated metrics, we can assess the overall performance of the linear decoder and the GAN-based techniques across different datasets, latent spaces, and VC region combinations. In the subsequent sections, we will present and discuss the results obtained from these evaluation metrics.

3.2 Image Dataset Results

The full quantitative evaluation results for the image datasets in different latent spaces and VC regions are presented in Appendix 6.3, Table 6.1. In the subsequent sections, we will discuss the results per dataset.

3.2.1 HYPER

See Figure 3.1 for the plotted quantitative results for the HYPER dataset. The latent space Wgt yielded low R (mean = 0.0123 ± 0.0074) compared to other latent spaces.

The CS (mean = 0.0124 ± 0.0074) and SSIM (mean = 0.0665 ± 0.0269) were also relatively low for Wgt. The latent space W showed a higher R (mean = 0.5349 ± 0.0141), indicating better performance in capturing image features in latent space from brain activity. The latent space W+ showed intermediate performance with respect to Pearson correlation (R = 0.1552 ± 0.0133), while having a high SSIM (mean = 0.5224 ± 0.0186) and low MSE (mean = 0.1070 ± 0.0104). The same MSE and SSIM values applied to the S+ latent space. The latent space S yielded a moderate R (mean = 0.4386), indicating satisfactory preservation of image content.

The evaluation results for the HYPER dataset demonstrate the varying performance of different latent spaces in preserving image content. Latent space W outperformed others with higher R and CS values, indicating better preservation of original image features. On the other hand, latent spaces W+ and S+ outperformed others with lower MSE and higher SSIM values. Latent space Wgt performed poorly in comparison with the other latent spaces. Image reconstructions from a HYPER stimulus for the different latent spaces can be found in Figure 3.2.

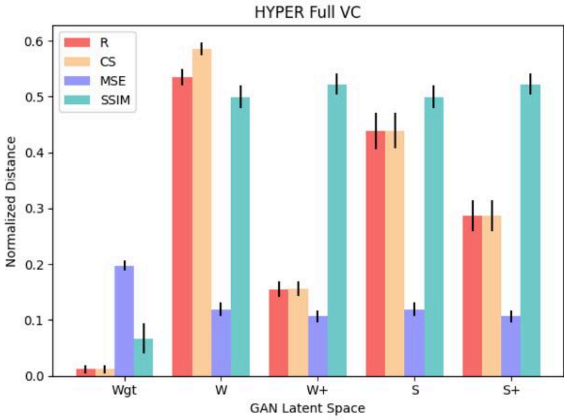


Figure 3.1: Quantitative evaluation results (mean and standard error) for the HYPER dataset in different latent spaces for full VC. R = Pearson correlation, CS = cosine similarity, MSE = mean squared error, SSIM = structural similarity index measurement.

3.2.2 Brain2GAN Faces

See Figure 3.3 for the plotted quantitative results for the Brain2GAN Faces dataset. The latent spaces W, W+, S, and S+ showed higher R, CS, SSIM, and lower MSE values than Wgt space, indicating better preservation of original image features in these latent spaces. The latent space W showed the highest R (mean = 0.7371 ± 0.0052) and CS (mean = 0.7754 ± 0.0043). The best performance for MSE (mean = 0.0371 ± 0.0016) and SSIM (mean = 0.6453 ± 0.0116) were for the W+ and S+ latent spaces, with both having the same values for these measurements. This indicated better similarity in image reconstruction to the ground truth images.

When exploring specific brain regions, the performance of the latent spaces seems to degrade slightly. Across all latent spaces, the full VC region consistently yielded the best results in terms of the evaluation metrics. Latent space W showed the highest correlation in full VC, suggesting that it can more accurately capture the underlying structures when taking into account the full VC. IT also showed relatively good predictive capabilities. In V1 and V4, the results are consistently lower than in the full VC region and IT. Image reconstructions from a Brain2GAN Faces stimulus for the different latent spaces can be found in Figure 3.4.

3.2.3 Brain2GAN Objects

See Figure 3.5 for the plotted quantitative results for the Brain2GAN Objects dataset. Latent space W consistently outperforms Wgt, W+, S, and S+ across all brain regions in terms of the evaluation metrics. It achieves the highest R (mean = 0.7328 ± 0.0048) and CS (mean = 0.8229 ± 0.0030), indicating a high similarity between the predicted and test latents. This suggests that W can capture the underlying features in the dataset more accurately. However, W+ and S+ show the lowest MSE values for the V4 region. Highest SSIM values were retrieved for the full VC regions of these latent spaces.

Across all latent spaces, the full VC region consistently returned better results in terms of the evaluation metrics, except for MSE, for which V4 showed the best performance. Latent space W consistently outperformed other spaces in the latent similarity measures, indicating its effectiveness in capturing features of natural object. On the other hand, latent spaces W+ and S+ outperformed other spaces in the image similarity measures, indicating its effectiveness in reconstructing images from features of natural object. Image reconstructions from a Brain2GAN Objects stimulus for the different latent spaces can be found in Figure 3.6.

3.2.4 THINGS

There was no ground truth latent space data available for the THINGS dataset. Therefore, we could use the THINGS dataset to test our method on non-GAN generated images without ground truth data. See Figure 3.7 for the plotted quantitative results for the Brain2GAN Objects dataset. Also here, latent space W consistently outperformed W+, S, and S+ in terms of latent similarity. W achieves the highest R value (mean = 0.6838 ± 0.0102), indicating a strong relationship between the predicted and test latents compared to other latent spaces. The lowest MSE and highest SSIM values can, again, be accredited to the W+ and S+ latent spaces, suggesting lower overall image reconstruction errors. Image reconstructions from a THINGS stimulus for the different latent spaces can be found in Figure 3.8.

3.3 Comparative Analysis

Across all datasets, the inverted latent spaces (W, W+, S, and S+) consistently outperform the ground truth latent space (Wgt) in terms of various evaluation metrics. The W latent space stands out as a consistently strong performer, demonstrating higher correlations between predicted and test latents across different datasets. On the other hand, the W+ and S+ latent spaces excel in image similarity measures, indicating their accuracy in reconstructing images from the latent features.

When evaluating the performance of latent spaces across different brain regions, we find that the full VC region consistently yields better results in terms of the evaluation metrics. In the Brain2GAN Faces dataset, latent space W performs best in the full VC region, followed by IT. For face images, the performance degrades slightly in V1 and V4 regions compared to the full VC and IT regions. In the Brain2GAN Objects dataset, the full VC region returns better results across all latent spaces, except when measuring MSE, with which the V4 region shows the best performance. For images of natural objects, the performance degrades slightly in V1 and IT regions compared to the full VC and V4 regions.

In summary, our comparative analysis reveals that the choice of latent space and brain region influences performance of the linear decoder. The predicted latents, using the linear decoder fit on inverted w latents, consistently demonstrated higher correlation with its corresponding test set (inverted test w latents) and better preservation of underlying image features across different datasets. On the other hand, the W+ and S+ latent spaces excel in image reconstruction, with the lowest MSE values and the highest SSIM values. Additionally, the full VC region consistently yields the best results in terms of evaluation metrics, indicating its relevance for the decoding task. A set of W, W+, and V4 image reconstructions can be found in Figure 3.9 for object-centered datasets. A set of W, W+, and IT image reconstructions can be found in Figure 3.10 for face-centered datasets.

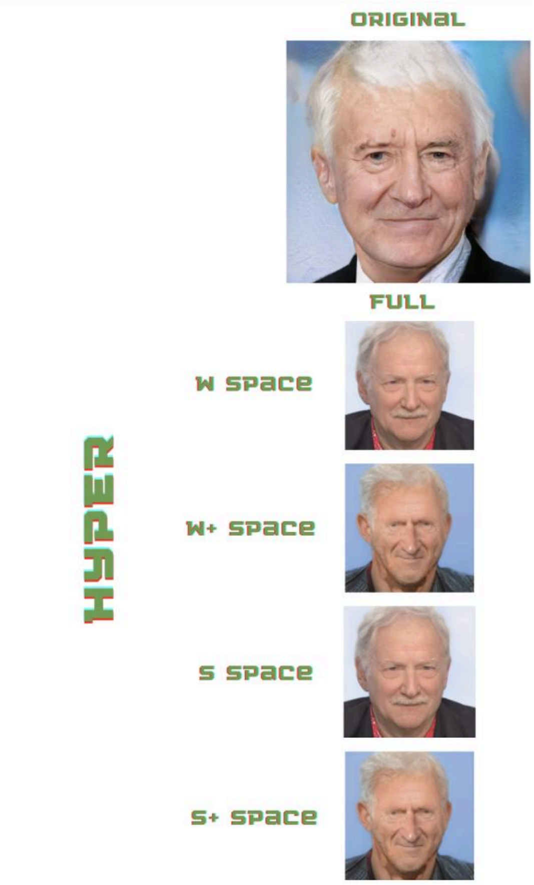


Figure 3.2: Image reconstructions for different latent spaces for a stimulus from the HYPER dataset.

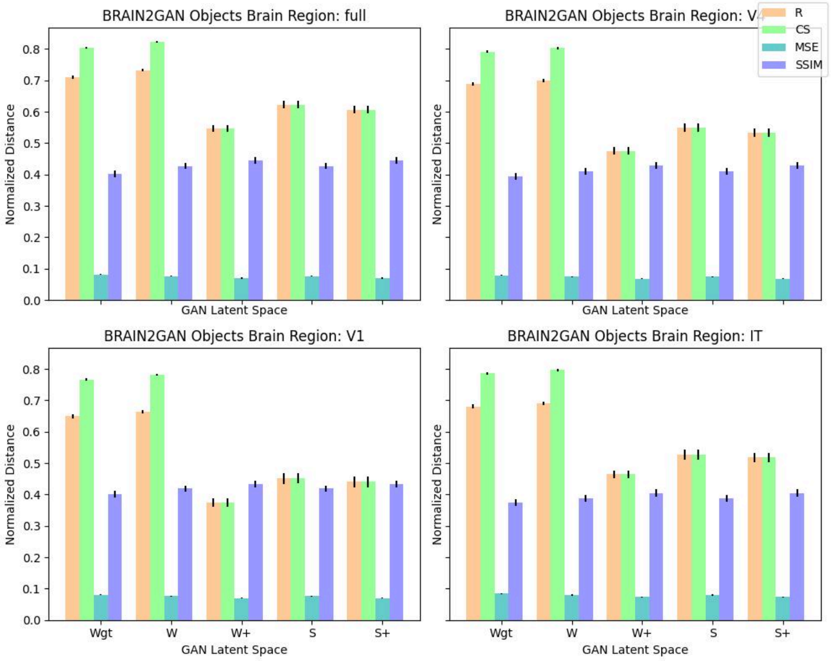


Figure 3.3: Quantitative evaluation results (mean and standard error) for the Brain2GAN Faces dataset in different latent spaces and VC regions. R = Pearson correlation, CS = cosine similarity, MSE = mean squared error, SSIM = structural similarity index measurement.

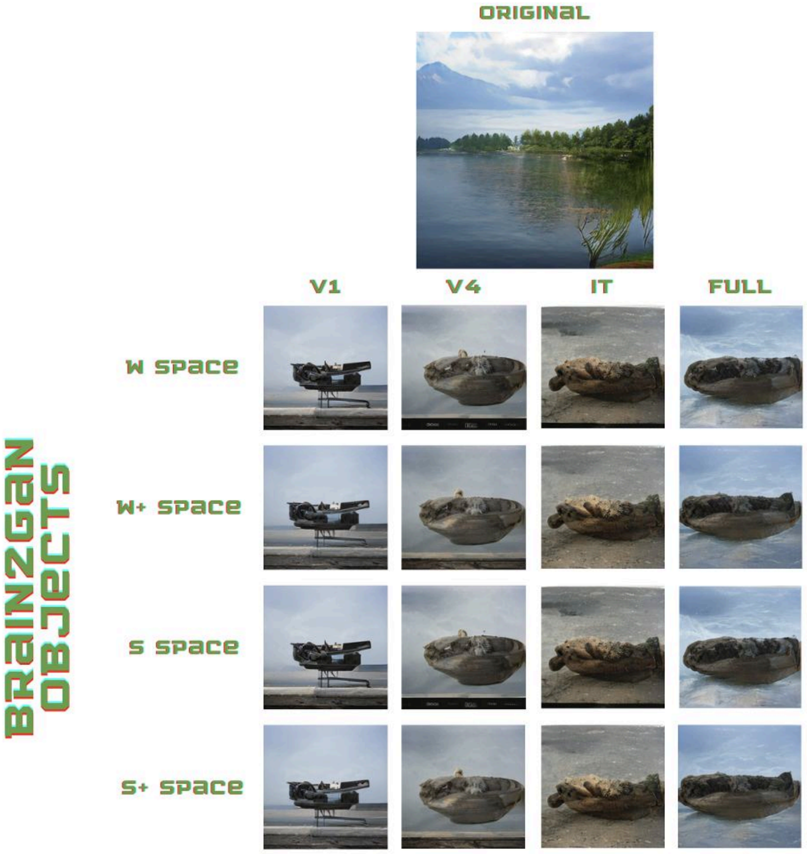


Figure 3.6: Image reconstructions for different latent spaces and brain regions for a stimulus from the Brain2GAN Objects dataset.

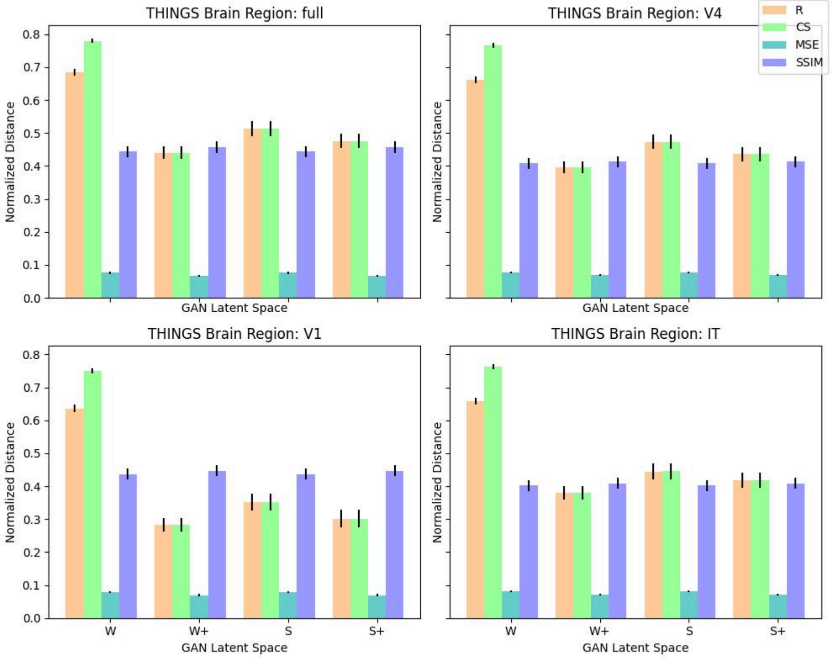


Figure 3.7: Quantitative evaluation results (mean and standard error) for the THINGS dataset in different latent spaces and VC regions. R = Pearson correlation, CS = cosine similarity, MSE = mean squared error, SSIM = structural similarity index measurement.

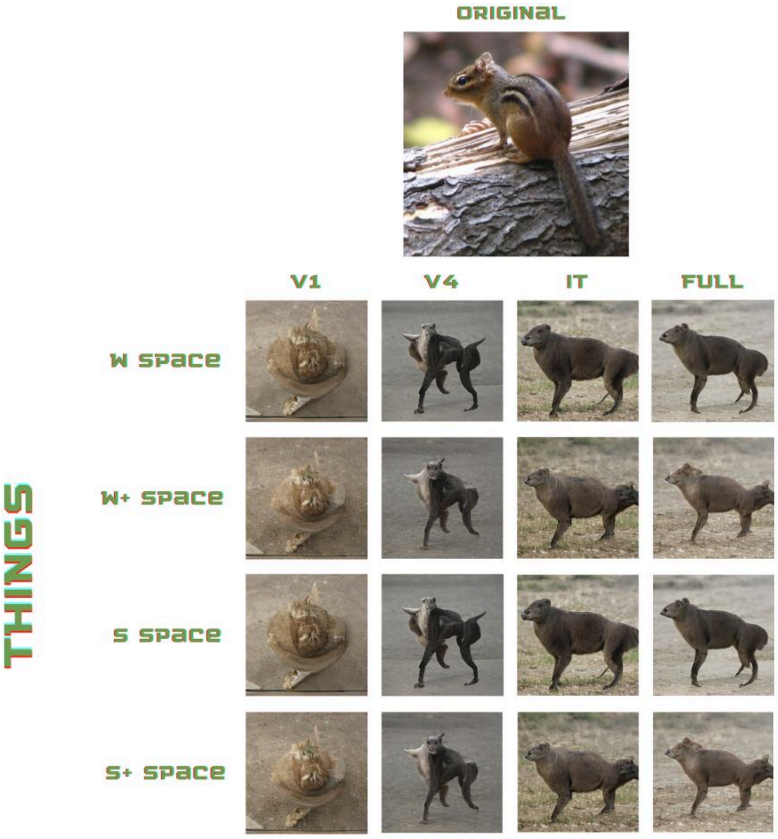


Figure 3.8: Image reconstructions for different latent spaces and brain regions for a stimulus from the THINGS dataset.

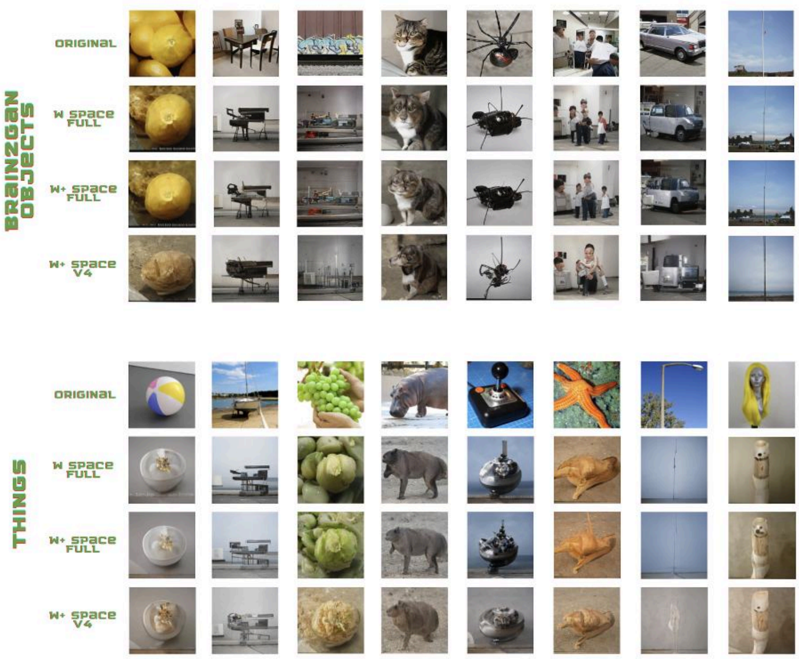


Figure 3.9: Image reconstructions for W and W+ latent spaces and V4 brain region for object-centered stimuli.

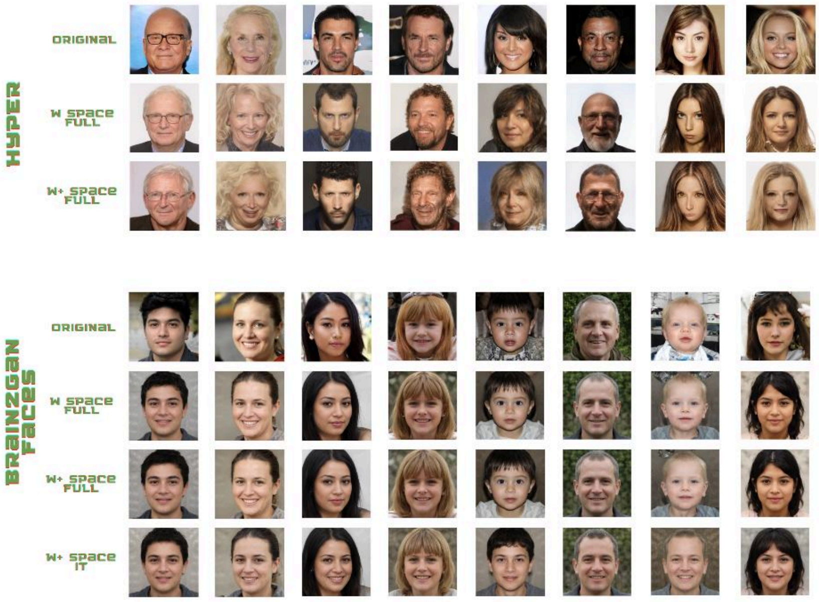


Figure 3.10: Image reconstructions for W and W+ latent spaces and IT brain region for face-centered stimuli.

Discussion

4.1 Overview of Findings

In this study about neural decoding using inverted latent spaces, we analyzed four diverse datasets: HYPER, Brain2GAN Faces, Brain2GAN Objects, and THINGS. Our findings demonstrate the varying performance of these spaces in preserving image content and capturing brain activity.

For the HYPER dataset, latent space W outperformed others, showing higher R values, indicating better feature preservation. Latent spaces W+ and S+ excelled in image similarity measures with lower MSE and higher SSIM scores. In the Brain2GAN Faces dataset, latent space W exhibited the highest R and CS values, while W+ and S+ excelled in MSE and SSIM measures. For the Brain2GAN Objects dataset, W consistently outperformed other spaces, achieving higher R and CS values. W+ and S+ showed the lowest MSE in the V4 region and the highest SSIM for full VC regions. In the absence of ground truth data, we evaluated the THINGS dataset, with W again showing the highest R values, and W+ and S+ excelling in image similarity metrics.

Across all datasets, inverted latent spaces consistently outperformed the ground truth latent space Wgt in various evaluation metrics. Latent space W showed the highest R and CS values, indicating a high similarity between predicted and test latents, and better preservation of original image features. Furthermore, latent spaces W+ and S+ demonstrated superior performance in terms of image similarity measures, showcasing their effectiveness in reconstructing images from features.

In most cases, the full VC region yielded the best results, indicating its importance in neural decoding tasks. For face images, the performance degrades slightly in V1 and V4 regions compared to the full VC and IT regions. For natural object images, the performance degrades slightly in V1 and IT regions compared to the full VC and V4 regions.
These findings underscore the importance of selecting suitable latent spaces for neural decoding objectives. In the following sections, we will discuss our observations in more detail.

4.2 Inverted Latent Spaces Outperform Ground Truth Latent Space

The inverted latent spaces (W, W+, S, S+) consistently outperformed the ground truth latent space (Wgt). This raises questions about the relationship between neural data, image features, and the inversion process.

These findings could be attributed to the relationship between neural data and image features. Neural systems are highly complex and do not necessarily encode external stimuli in a straightforward manner. Rather, they process and represent information in ways optimized for specific cognitive processes [31, 32]. The inverted latent spaces may prioritize certain image features that are crucial for these cognitive processes, while ground truth latent spaces are based on the features initially used by the generator for image synthesis.

Furthermore, ground truth latent spaces are typically derived from the generator's synthesis process, which can introduce biases [33]. In contrast, the inversion process aims to reverse-engineer the latent space by working towards the highest image similarity, potentially avoiding some of these biases and aligning more closely with how the brain processes visual information.

These findings suggest that the inversion process could reveal latent representations that bridge the gap between neural data and image features more effectively. These insights may not only advance the field of neural decoding but also influence image processing, computer vision, and cognitive science.

4.3 Evaluation of Latent Similarity

The consistent superior performance of W latents in latent similarity evaluation raises questions about their underlying characteristics and their potential alignment with neural representations of image features. As we did not employ PCA for W and Wgt latents, it is possible that the original 512-dimensional W space captures key features and relationships more accurately, avoiding potential information loss that may occur with dimensionality reduction techniques. By retaining the full dimensionality of W latents, they might preserve a richer set of image features, contributing to their higher latent similarity scores.

Another aspect to consider is the difference in dimensionality between the W and W+, S, and S+ latent spaces. By reducing the dimensionality to 512 dimensions using PCA for the latter spaces, we aimed to mitigate the challenges posed by the ”Curse of Dimensionality”. However, this reduction could potentially result in the loss of some fine-grained features captured in the original higher-dimensional space, leading to relatively lower latent similarity for W+, S, and S+ compared to the W latent space.

Future research could explore alternative dimensionality reduction techniques that might better preserve critical image features. Techniques beyond PCA might offer alternative strategies to mitigate the challenges of overfitting in high-dimensional feature spaces. For example, t-Distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) are dimensionality reduction techniques known for their ability to preserve local structures and non-linear relationships in data [34]. By applying these methods, we may retain more relevant information while reducing dimensionality.

4.4 Evaluation of Image Similarity

One of our key observations is that, while W latents outperformed in latent similarity, W+, S, and S+ latents exhibited superior performance in image similarity metrics in terms of low MSE and high SSIM. This suggests that these latent spaces might excel in reconstructing images from features, even though they showed relatively lower latent similarity scores.

The evaluation metrics used to assess the performance of latent spaces play a crucial role in understanding the strengths and limitations of each model. As our results demonstrate, W latents consistently achieve the highest R and CS scores. However, it is essential to have a balance between latent similarity and image similarity.

While W latents excel in capturing underlying features and correlations, the W+ and S+ spaces demonstrate remarkable image similarity in terms of low MSE and high SSIM. The latter two metrics focus on the quality of image reconstruction which is important for applications using image generation and synthesis.

In the context of neural decoding, where the objective is to transform brain activity into stimuli, image similarity becomes a critical aspect of the evaluation process. While measurements such as MSE and SSIM focus on pixel-wise similarity, the importance lies in ensuring that the reconstructed images resemble to the original stimuli in terms of underlying features and visual characteristics.

To address this concern, we recognize that images can be similar in the features they display, even if their pixel-wise values differ. Alternative methods for image similarity assessment can provide a more comprehensive evaluation of the reconstructed images. One such approach involves using feature detectors like VGG16 to extract features from both the reconstructed and test images [21]. By calculating the difference in these extracted features, we can obtain a measure of feature-level similarity, providing insights into how well the latent space preserves visual features. By comparing features at different layers of the VGG16 network can reveal hierarchical similarities between images, capturing both low-level and high-level visual representations.

4.5 The Role of Visual Cortex Regions

While the full VC consistently demonstrated the highest decoding accuracy across all datasets, specific VC subregions showed varying influences on decoding performance. For the Brain2GAN Faces dataset, latent space W performed best in the full VC region, followed closely by IT. Decoding accuracy in V1 and V4 regions showed slight degradation compared to the full VC and IT regions. Interestingly, for the Brain2GAN Objects and THINGS datasets, a similar pattern emerged, with the V4 region closely following the full VC in terms of decoding accuracy across latent spaces. This suggests a consistent trend where specific VC subregions play crucial roles in decoding specific types of visual information. The IT region seems more important for face-related stimuli, while the V4 region appears to be more relevant for object-related stimuli.

Potential explanations for this pattern could be linked to the hierarchical nature of visual processing in the brain. Studies have shown that the ventral stream, which includes the V4 and IT regions, are involved in high-level visual processing. Studies have shown that IT neurons exhibit high selectivity for face stimuli, responding strongly to faces compared to other objects or stimuli [35–37]. On the other hand, the explicit representation of curvature in V4 is an effective method for encoding boundary elements of natural objects, making it more relevant for object-related stimuli [38–40]. The differing functional characteristics of these VC subregions could explain their varying degrees of importance in decoding different types of visual stimuli.

Future research could explore the dynamic interactions between VC subregions and latent spaces to facilitate the development of more targeted decoding approaches. We could, for example, delve deeper into the specific interactions of IT and V4 regions with latent spaces for face-

and object- centered datasets, probing the neural representations of various visual stimuli. This could potentially lead to more tailored and effective decoding strategies for various brain regions.

4.6 Naturalness and Size of Dataset

Although THINGS was the larger dataset among the object datasets, it showed marginal improvements in image similarity compared to Brain2GAN Objects, and its latent similarity results were worse in comparison to Brain2GAN Objects. However, the THINGS dataset was the only dataset consisting of natural images. It should be noted that GAN-generated images have certain advantages over natural images. GANs can learn complex underlying data distributions and can generate high-quality, realistic images that preserve essential features while decreasing unwanted biases [41]. As a result, the inversion process, which involves mapping brain activity to the GAN latent space, is generally more straightforward and efficient with GAN-generated images.

In contrast, the inversion of natural images introduces several challenges. Natural images exhibit a wide range of variations, including lighting conditions, backgrounds, and object poses, which can complicate the inversion process. Additionally, the unique features and characteristics specific to GAN-generated images, learned during GAN training, may not adequately represent certain aspects of natural images during inversion. Therefore, the higher image similarity observed for the THINGS dataset is a significant finding in the context of neural decoding. Reconstructing images accurately from brain activity, even in the absence of ground truths and with natural images, demonstrates the potential for real-world applications where ground truth data might not be available. In practical scenarios, humans perceive and interact with natural objects without access to the exact ground truth, relying on their neural representations to make inferences about the surrounding environment.

While GAN-generated datasets like Brain2GAN Objects and Brain2GAN Faces offer certain advantages in terms of image quality and diversity, the success of THINGS in neural decoding from brain activity opens new avenues for decoding tasks in situations where GAN-generated images might not be readily available or feasible. This finding highlights the robustness of the proposed decoding framework and its potential to provide meaningful insights into neural representations of natural objects, even in real-world, uncontrolled conditions.

However, when working with natural images, such as THINGS, there could be a need for further refinement of the decoding pipeline to account for realworld variations. Future research could explore techniques to address these challenges and optimize the use of natural image datasets for specific neural decoding objectives. To address these challenges, it is essential to recognize and account for the differences between GAN-generated images and natural images in the context of neural decoding. Future research should focus on developing robust techniques that effectively handle the complexities and variations present in natural images, thus enabling successful decoding from brain activity under real-world, uncontrolled conditions.

4.7 Link to Research Question and Implications

Our study tried to answer the research question: “To what extent can image reconstruction from brain activity be improved by using image inversion methods for GAN latent spaces W, W+, S, and S+, particularly when ground truth latents are unavailable?” Our findings provide insights into the relationship between latent spaces and neural decoding, as well as the broader implications of our work.

The core findings of our study, demonstrating the superior performance of inverted latent spaces compared to ground truth latent space (Wgt), directly address the research question’s essence. This observation shows that, even in the absence of ground truth data, inverted latent spaces can be used to enhance image reconstruction from brain activity. It challenges the idea that ground truth latents are the optimal choice for neural decoding.

Our findings align with previous studies that have shown the complexities of neural representations and their potential divergence from stimulus-specific ground truth features [6, 7, 20, 21, 42, 43]. Our results extend these insights by demonstrating the benefits of employing inverted latent spaces. Future studies may benefit from incorporating inverted latent spaces into their methods, particularly when ground truth latents are unavailable.

Our findings deepen our understanding of how neural data interacts with image features. They underscore the dynamic nature of neural encoding and the capability of neural representations to emphasize specific cognitive processes. This knowledge could shape future research in the fields of neuroscience and cognitive science, enhancing our comprehension of information processing in the brain.

In the realm of artificial intelligence and computer vision, our research offers a new perspective on latent space utilization. The demonstrated inversion methods can yield representations that align more closely with brain activity. This could potentially advance applications such as image recognition, synthesis, and interpretation.

4.8 Diversity and Quality in Dataset

Considering the impact of using diverse and large datasets for inversion and for training the linear decoder, we have observed differences in the performance of latent spaces. In the case of the HYPER dataset, which included 1050 low-quality images from the CelebA dataset for inversion and linear decoder training, we encountered challenges with distorted reconstructions and noticeable bias toward individuals of White ethnicity (see Figure 4.1). These limitations were worsened by the lower quality of the accompanying fMRI data, which had lower spatial and temporal resolution than the MUA data.

However, when employing the Brain2GAN Faces dataset, which featured 4000 images sourced from the more extensive, high-quality, and diverse FFHQ dataset for training, considerable improvements became evident. The reconstructions in this context were not only more realistic but also showed better preservation of characteristics in individuals from various ethnic backgrounds. This enhanced performance can be attributed not only to the superior image quality but also to the higher quality of the accompanying brain activity data. The Brain2GAN Faces dataset benefited from high-quality responses, which offered improved spatial and temporal resolution compared to the fMRI data, resulting in a more robust and faithful mapping between neural representations and image latents.

4.9 Ethical Considerations

The macaque monkey data that was used in this study followed guidelines outlined in the NIH Guide for Care and Use of Laboratory Animals. Experimental protocols were approved by the local institutional animal care and use committee of the Royal Netherlands Academy of Arts and Sciences to ensure compliance with ethical standards for animal research. It is essential to acknowledge that, while this research was conducted in accordance with established ethical guidelines, the welfare of the animal participants remains a serious concern.

The human data that was used in this study was in full compliance with ethical guidelines for human subject research. Ethical protocols were followed to protect the rights, privacy, and well-being of human participants involved in data collection and analysis. There was an informed consent procedure in advance, providing participants with comprehensive information about the study's objectives, procedures, and potential risks. As researchers, we hold a moral responsibility to safeguard the interests and welfare of both human and animal subjects.

This study underscores the importance of maintaining a conscientious approach to ethical considerations in future research, especially when it comes

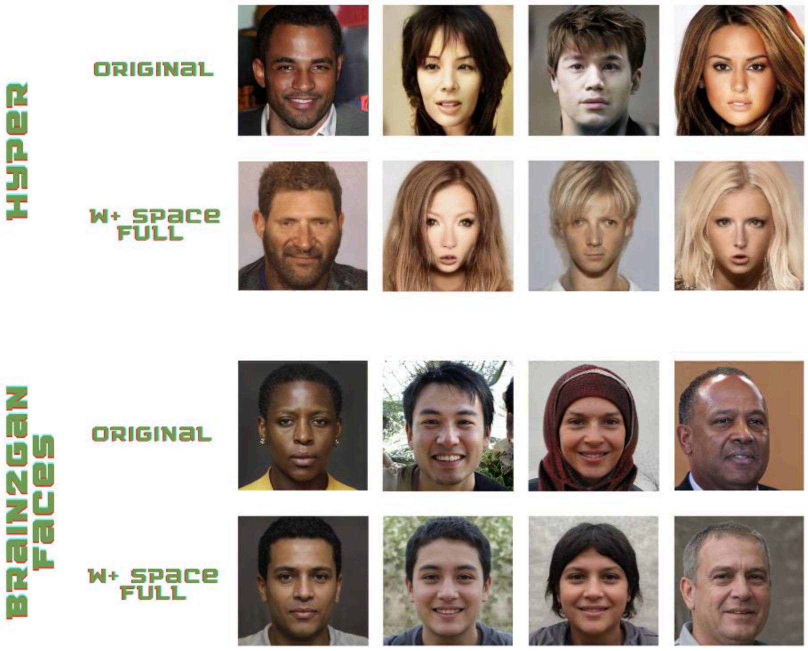


Figure 4.1: HYPER image reconstructions for W+ latent space in face-centered stimuli of individuals from various ethnic backgrounds

to non-human participants who lack the capacity for autonomous decisionmaking. The sacrifice of the macaque monkey in this experiment highlights the ethical complexities and directs us to reevaluate and prioritize the wellbeing of all participants in scientific research.

Ethical implications of brain reading technologies and consumer neurotechnology [44] were considered. The study acknowledges concerns about mental privacy and potential misuse of neurotechnologies. While the current neural decoding approaches employed in this study do not allow for involuntary access to individuals' thoughts, ongoing discussions on the responsible use of neural decoding techniques and safeguarding mental privacy are essential.

Conclusion

In this study, we explored latent spaces to enhance image reconstruction from brain activity, particularly when ground truth latents are unavailable. We investigated the potential of inverted latent spaces for decoding brain activity and reconstructing visual stimuli. Our findings show the power of different latent spaces and answered our research question: "To what extent can image reconstruction from brain activity be improved by using image inversion methods for GAN latent spaces W, W+, S, and S+, particularly when ground truth latents are unavailable?"

The core findings of our study show that inverted latent spaces consistently outperformed the ground truth latent space. On top of that, we were able to accurately reconstruct natural images without ground truth latents, which shows the potential of inverted latent spaces to decode brain activity from real-life situations. We showed that inverted latent spaces offer promising results for enhancing image reconstruction from brain activity. It challenges the conventional wisdom that ground truth latents are the optimal choice for neural decoding.

Through a comprehensive evaluation using multiple metrics, we assessed the performance of the latent spaces Wgt, W, W+, S, and S+. Notably, the W latents were most capable in capturing relevant image features from brain activity compared to other spaces. W+ and S+ showed promising results in image similarity. The trade-offs between latent similarity measures and image similarity metrics underscore the importance of careful evaluation metric selection.

Our analysis of the impact of brain region selection on decoding performance revealed distinctive patterns in face-centered and object-centered datasets. In face datasets, the W latent space demonstrated optimal performance in full visual cortex and inferior temporal regions, while in object datasets, W outperformed others in both full visual cortex and V4 regions. These findings emphasize that different brain regions can give specific information and should be taken into account when designing decoding models.

Addressing the challenges and opportunities presented by natural image datasets like THINGS can enrich our understanding of neural decoding with real-world stimuli. The ability to accurately reconstruct from natural images, even without ground truth data, shows the potential of inverted latent spaces to decode brain activity from real-life situations. By considering diverse, large, and high-quality natural datasets, future research can explore the potential of inverted latent spaces and their suitability for real-world decoding tasks.

In conclusion, this study demonstrates the potential of inverted latent spaces for neural decoding. Our findings show that, even in the absence of ground truth data, inverted latent spaces offer promising results for enhancing image reconstruction from brain activity. They highlight the effectiveness of W latents in capturing relevant image features from brain activity and of W+ and S+ latents in reconstrucing images from brain activity. By addressing challenges and opportunities posed by natural image datasets, future research can unlock the full potential of inverted latent spaces for real-world decoding tasks. Overall, this study shows possibilities of neural decoding using inverted latent spaces, paving the way for more sophisticated decoding approaches, and fostering advancements in neurotechnology and real-world applications.

References

1. Grill-Spector, K. & Malach, R. The human visual cortex. Annu. Rev. Neurosci. 27, 649–677 (2004).

2. Wandell, B. A. & Winawer, J. Imaging retinotopic maps in the human brain. Vision research 51, 718–737 (2011).

3. Super, H. & Roelfsema, P. R. Chronic multiunit recordings in behaving animals: advantages and limitations. Progress in brain research 147, 263–282 (2005).

4. Logothetis, N. K. What we can do and what we cannot do with fMRI. Nature 453, 869–878 (2008).

5. Christopher deCharms, R. Applications of real-time fMRI. Nature Reviews Neuroscience 9, 720–729 (2008).

6. Haxby, J. V., Connolly, A. C. & Guntupalli, J. S. Decoding neural representational spaces using multivariate pattern analysis. Annual review of neuroscience 37, 435–456 (2014).

7. Warren, D. J. et al. Recording and decoding for neural prostheses. Proceedings of the IEEE 104, 374–391 (2016).

8. Van Gerven, M. A., Seeliger, K., G¨u¸cl¨u, U. & G¨u¸cl¨ut¨urk, Y. Current advances in neural decoding. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, 379–394 (2019).

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

10. G¨u¸cl¨u, U. & van Gerven, M. A. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. Journal of Neuroscience 35, 10005–10014 (2015).

11. Goodfellow, I. et al. Generative adversarial nets. Advances in neural information processing systems 27 (2014).

12. Yamins, D. L. & DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. Nature neuroscience 19, 356–365 (2016).

13. Karras, T., Laine, S. & Aila, T. A style-based generator architecture for generative adversarial networks in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2019), 4401– 4410.

14. Karras, T. et al. Analyzing and improving the image quality of stylegan in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2020), 8110–8119.

15. Wu, Z., Lischinski, D. & Shechtman, E. Stylespace analysis: Disentangled controls for stylegan image generation in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021), 12863–12872.

16. Xia, W. et al. Gan inversion: A survey. IEEE transactions on pattern analysis and machine intelligence 45, 3121–3138 (2022).

17. Abdal, R., Qin, Y. & Wonka, P. Image2stylegan: How to embed images into the stylegan latent space? in Proceedings of the IEEE/CVF international conference on computer vision (2019), 4432–4441.

18. Karras, T. et al. Alias-free generative adversarial networks. Advances in Neural Information Processing Systems 34, 852–863 (2021).

19. Sauer, A., Schwarz, K. & Geiger, A. Stylegan-xl: Scaling stylegan to large diverse datasets in ACM SIGGRAPH 2022 conference proceedings (2022), 1–10.

20. Dado, T. et al. Hyperrealistic neural decoding for reconstructing faces from fMRI activations via the GAN latent space. Scientific reports 12, 141 (2022).

21. Dado, T. et al. Brain2GAN; Reconstructing perceived faces from the primate brain via StyleGAN3 (2023).

22. Karras, T., Aila, T., Laine, S. & Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017).

23. Liu, Z., Luo, P., Wang, X. & Tang, X. Deep learning face attributes in the wild in Proceedings of the IEEE international conference on computer vision (2015), 3730–3738.

24. Hebart, M. N. et al. THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. PloS one 14, e0223792 (2019).

25. Deng, J. et al. Imagenet: A large-scale hierarchical image database in 2009 IEEE conference on computer vision and pattern recognition (2009), 248–255.

26. Russakovsky, O. et al. Imagenet large scale visual recognition challenge. International journal of computer vision 115, 211–252 (2015).

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

28. Roich, D., Mokady, R., Bermano, A. H. & Cohen-Or, D. Pivotal tuning for latent-based editing of real images. ACM Transactions on graphics (TOG) 42, 1–13 (2022).

29. Pedregosa, F. et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research 12, 2825–2830 (2011).

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

31. DiCarlo, J. J. & Cox, D. D. Untangling invariant object recognition. Trends in cognitive sciences 11, 333–341 (2007).

32. Hassabis, D. & Maguire, E. A. The construction system of the brain. Philosophical Transactions of the Royal Society B: Biological Sciences 364, 1263–1271 (2009).

33. Xu, R., Wang, X., Chen, K., Zhou, B. & Loy, C. C. Positional encoding as spatial inductive bias in gans in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021), 13569– 13578.

34. Wang, Y., Huang, H., Rudin, C. & Shaposhnik, Y. Understanding how dimension reduction tools work: an empirical approach to deciphering t-SNE, UMAP, TriMAP, and PaCMAP for data visualization. The Journal of Machine Learning Research 22, 9129–9201 (2021).

35. Tsao, D. Y., Freiwald, W. A., Tootell, R. B. & Livingstone, M. S. A cortical region consisting entirely of face-selective cells. Science 311, 670–674 (2006).

36. Loffler, G., Yourganov, G., Wilkinson, F. & Wilson, H. R. fMRI evidence for the neural representation of faces. Nature neuroscience 8, 1386–1391 (2005).

37. Pinsk, M. A. et al. Neural representations of faces and body parts in macaque and human cortex: a comparative FMRI study. Journal of neurophysiology 101, 2581–2600 (2009).

38. David, S. V., Hayden, B. Y. & Gallant, J. L. Spectral receptive field properties explain shape selectivity in area V4. Journal of neurophysiology 96, 3492–3505 (2006).

39. Connor, C. E., Brincat, S. L. & Pasupathy, A. Transformation of shape information in the ventral pathway. Current opinion in neurobiology 17, 140–147 (2007).

40. Kourtzi, Z. & Connor, C. E. Neural representations for object perception: structure, category, and adaptive coding. Annual review of neuroscience 34, 45–67 (2011).

41. Creswell, A. et al. Generative adversarial networks: An overview. IEEE signal processing magazine 35, 53–65 (2018).

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

43. Ahlheim, C. & Love, B. C. Estimating the functional dimensionality of neural representations. NeuroImage 179, 51–62 (2018).

44. Ienca, M., Haselager, P. & Emanuel, E. J. Brain leaks and consumer neurotechnology. Nature biotechnology 36, 805–810 (2018).

45. Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O. & Cohen-Or, D. Designing an encoder for stylegan image manipulation. ACM Transactions on Graphics (TOG) 40, 1–14 (2021).

46. Shen, Y., Gu, J., Tang, X. & Zhou, B. Interpreting the latent space of gans for semantic face editing in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2020), 9243–9252.

Appendix

6.1 Pivotal Tuning Inversion

Pivotal Tuning Inversion (PTI) is a novel approach aimed at enabling realistic image editing for out-of-domain images by addressing the distortioneditability tradeoff in GAN inversion [28]. The main challenge in editing real images using GANs arises from the fact that real images do not belong to the trained generator's latent space, making it difficult to directly edit them. PTI overcomes this limitation by introducing an innovative methodology that combines GAN inversion and fine-tuning of the generator.

Early attempts on finding the latent representation of a real image in the generator's latent space involved inverting the image to the generator's native latent space W. However, this resulted in distorted images. Recent studies have introduced an extended latent space, W+, which allows for less distortion and better visual quality for out-of-domain images [17]. However, W+ suffers from weaker editability compared to W, creating a distortioneditability tradeoff. PTI was introduced as a solution to the distortion-editability tradeoff conflict [45].

PTI enables realistic image editing for out-of-domain images in a two-step process. In the first step, PTI inverts the input image to find an editable latent code that represents the image in the generator's latent space. This results in an image that closely resembles the original input.

The second step of PTI involves tuning the pretrained StyleGAN using the editable latent code found in the first step. Instead of projecting the input image into the generator's latent space, PTI augments the learned manifold to include the image by making subtle modifications to the generator [28]. This approach maintains the editing qualities of the latent code while achieving unprecedented reconstruction quality (see Figure 6.1).

PTI provides a compromise between distortion and editability, allowing for more realistic and high-quality image editing compared to conventional GAN inversion methods. By incorporating the input image directly into the learned manifold, PTI ensures that the latent code preserves its editing capabilities, leading to improved reconstruction quality (see Figure 6.2).

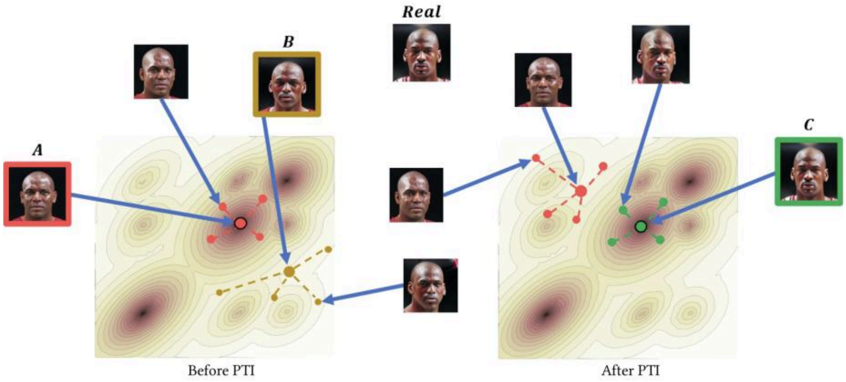
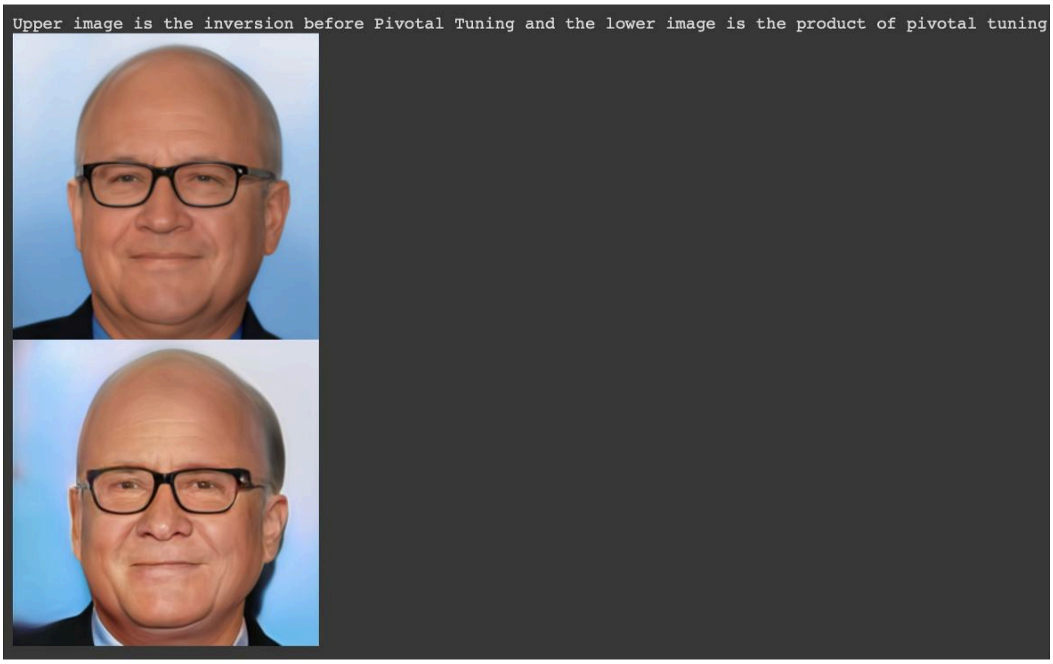


Figure 6.1: Illustration of the PTI method, retrieved from [28]. In StyleGAN's latent space, warmer colors indicate greater editability. On the left, before Pivotal Tuning, we face an Editability-Distortion tradeoff between two identities, ”A” and ”B.” ”A” is highly editable but differs from the original image, while ”B” is less editable with fewer artifacts but more distortion. After Pivotal Tuning on the right, Roich et al. [28] introduce ”C,” which retains ”A's” editing capabilities while improving similarity to the original image compared to ”B”.

6.1.1 Applying PTI on Neural Decoding

Despite its advantages, the performance of PTI, when applied to image reconstruction from brain activity, has shown limitations. The reconstructed images from brain activity, using a trained linear decoder on inverted images and brain activity, could not be tuned closely enough to the original images (see Figure 6.3). This suggests that PTI, which was originally designed for better image editing, may not be directly applicable for the specific task of neural decoding using inverted latents for image reconstruction from brain activity.



```
Upper image is the inversion before Pivotal Tuning and the lower image is the product of pivotal tuning
```

Figure 6.2: PTI applied solely on inverted image.

In conclusion, PTI represents an innovative approach that addresses the distortion-editability tradeoff in GAN inversion, enabling realistic image editing for out-of-domain images. While PTI has demonstrated promising results in various image editing tasks, its applicability to neural decoding from brain activity requires further investigation and adaptation to fit with the specific challenges of image reconstruction from brain activity.

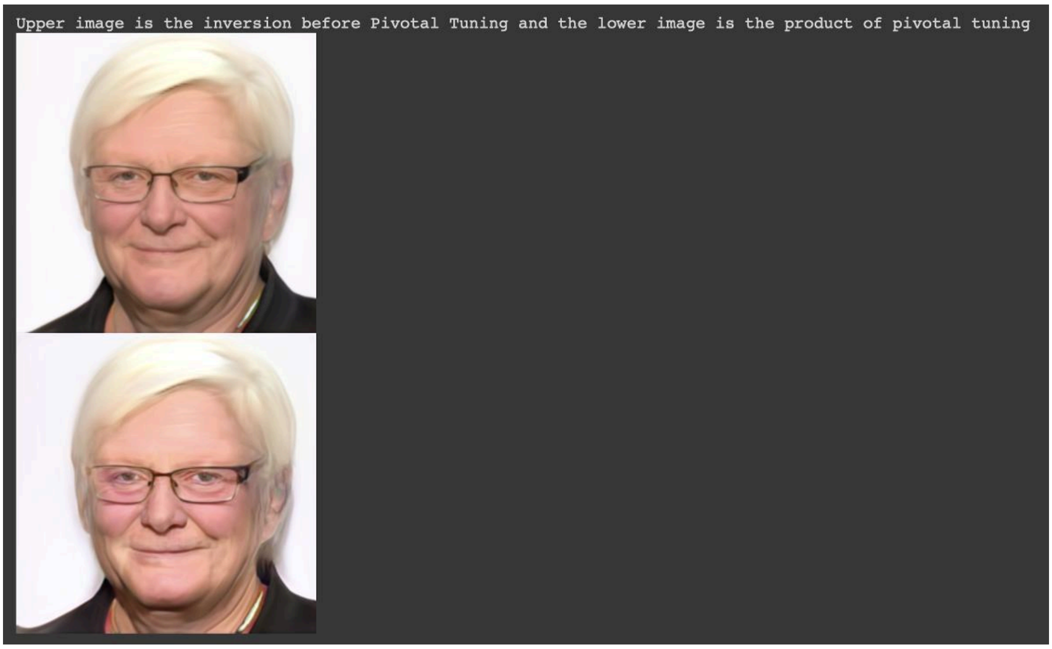

Upper image is the inversion before Pivotal Tuning and the lower image is the product of pivotal tuning

Figure 6.3: PTI applied on reconstructed image from trained neural decoder predictions on the inverted image and brain activity.

6.2 Feature Entanglement

Feature entanglement refers to the phenomenon where specific image features, such as facial attributes or object characteristics, are not clearly separated in the GAN latent space [46]. We investigated feature entanglement in StyleGAN3 (for face-centered image datasets) and StyleGAN-XL (for object-centered image datasets) using the extended latent space W+.

GAN generators are susceptible to feature entanglement because to biases present in the training data. For example, editing a latent vector to make the generated face wear eyeglasses might simultaneously affect other facial attributes, such as making the face appear older [20]. Such entanglement arises from the complex interactions between different visual features in the GAN's latent space.

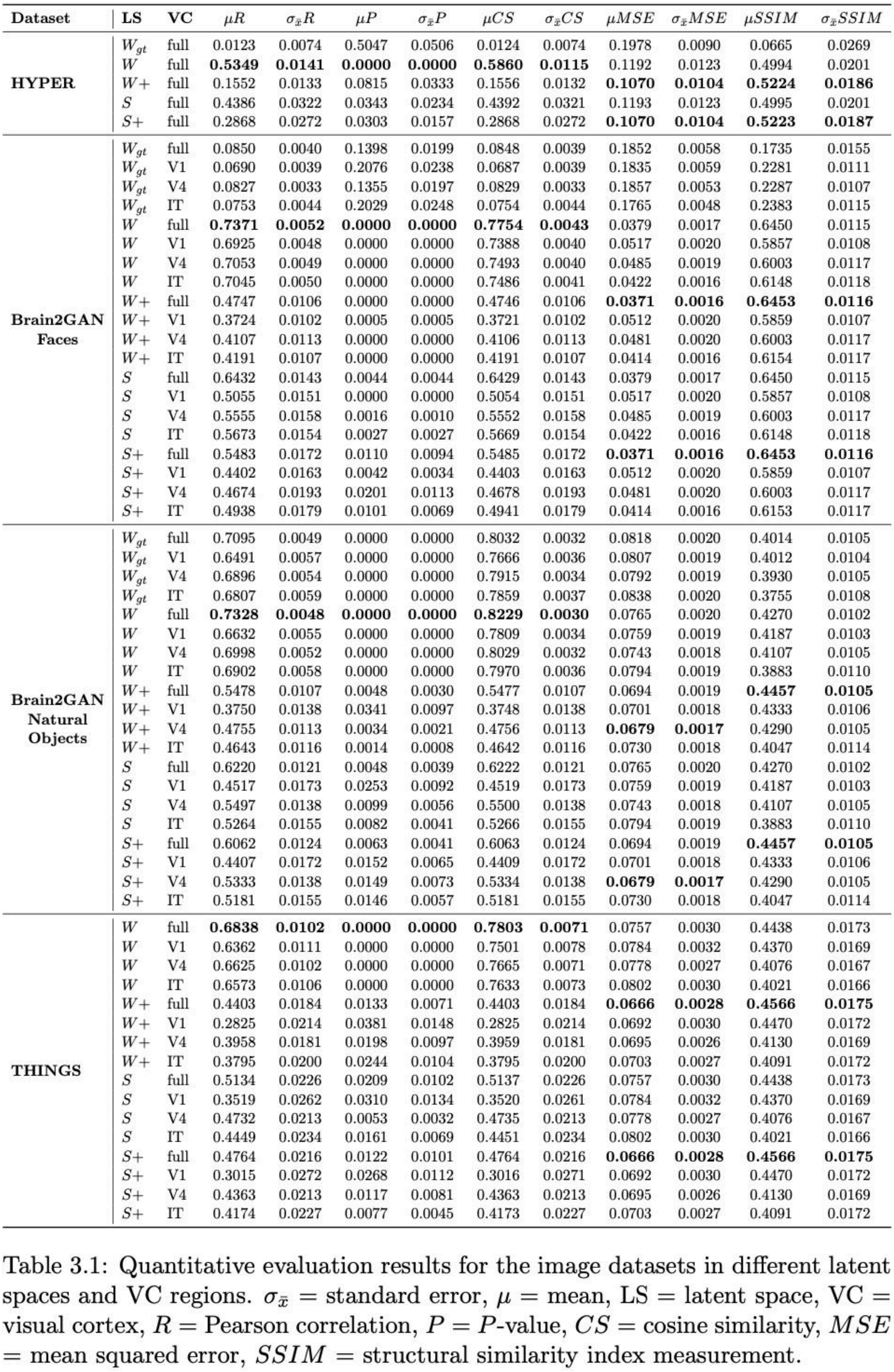
6.2.1 Using Distinct W+ Channels for Feature Disentanglement

To study feature entanglement, we reconstructed images for individual input channels of both StyleGAN3 and StyleGAN-XL, using the extended latent space W+. Each input channel within W+ corresponds to a unique set of image features, allowing us to examine distinct feature representations. We employed two reconstruction methods for this analysis. The first method involved individual reconstructions from specific W+ latents (see Figure 6.5), while the second method included cumulative reconstructions from combined W+ latents (see Figure 6.4 and Figure 6.6). In both cases, for channels not under investigation, we used w latents from the W space to fill those channels.

Based on our results, we did not observe clear and distinct patterns per channel. This suggests that the extended latent space W+ might not facilitate feature disentanglement in the current implementation. Currently, we used w latents from the latent space W as filling for the channels not specifically under investigation within the extended latent space W+. It is worth noting that the W space contains identical latent values across all channels, effectively representing the same set of image features. This uniformity across channels in W can potentially introduce a challenge in terms of feature separation and independence. When we apply w latents from W to channels within W+, we introduce latent values that inherently carry information about multiple image features. This intermingling of latent values from W into the extended latent space W+ may result in feature mixing and entanglement. In other words, the latent values meant to represent specific image attributes within the channels under investigation may become influenced by latent values representing unrelated features. This phenomenon can complicate the disentanglement process and make it challenging to isolate and manipulate individual features accurately.

Our findings underscore the complexity of disentangling image features in high-dimensional latent spaces, especially when the latent spaces have shared or uniform components. Future research may explore alternative strategies to reduce feature mixing and enhance feature disentanglement in such settings.

6.3 Full Quantitative Results



Dataset	LS	VC	μR	$\sigma_{\bar{x}} R$	μP	$\sigma_{\bar{x}} P$	μCS	$\sigma_{\bar{x}} CS$	μMSE	$\sigma_{\bar{x}} MSE$	$\mu SSIM$	$\sigma_{\bar{x}} SSIM$
HYPER	W_{gt}	full	0.0123	0.0074	0.5047	0.0506	0.0124	0.0074	0.1978	0.0090	0.0665	0.0269
	W	full	**0.5349**	**0.0141**	**0.0000**	**0.0000**	**0.5860**	**0.0115**	0.1192	0.0123	0.4994	0.0201
	$W+$	full	0.1552	0.0133	0.0815	0.0333	0.1556	0.0132	**0.1070**	**0.0104**	**0.5224**	**0.0186**
	S	full	0.4386	0.0322	0.0343	0.0234	0.4392	0.0321	0.1193	0.0123	0.4995	0.0201
	$S+$	full	0.2868	0.0272	0.0303	0.0157	0.2868	0.0272	**0.1070**	**0.0104**	**0.5223**	**0.0187**
Brain2GAN Faces	W_{gt}	full	0.0850	0.0040	0.1398	0.0199	0.0848	0.0039	0.1852	0.0058	0.1735	0.0155
	W_{gt}	V1	0.0690	0.0039	0.2076	0.0238	0.0687	0.0039	0.1835	0.0059	0.2281	0.0111
	W_{gt}	V4	0.0827	0.0033	0.1355	0.0197	0.0829	0.0033	0.1857	0.0053	0.2287	0.0107
	W_{gt}	IT	0.0753	0.0044	0.2029	0.0248	0.0754	0.0044	0.1765	0.0048	0.2383	0.0115
	W	full	**0.7371**	**0.0052**	**0.0000**	**0.0000**	**0.7754**	**0.0043**	0.0379	0.0017	0.6450	0.0115
	W	V1	0.6925	0.0048	0.0000	0.0000	0.7388	0.0040	0.0517	0.0020	0.5857	0.0108
	W	V4	0.7053	0.0049	0.0000	0.0000	0.7493	0.0040	0.0485	0.0019	0.6003	0.0117
	W	IT	0.7045	0.0050	0.0000	0.0000	0.7486	0.0041	0.0422	0.0016	0.6148	0.0118
	$W+$	full	0.4747	0.0106	0.0000	0.0000	0.4746	0.0106	**0.0371**	**0.0016**	**0.6453**	**0.0116**
	$W+$	V1	0.3724	0.0102	0.0005	0.0005	0.3721	0.0102	0.0512	0.0020	0.5859	0.0107
	$W+$	V4	0.4107	0.0113	0.0000	0.0000	0.4106	0.0113	0.0481	0.0020	0.6003	0.0117
	$W+$	IT	0.4191	0.0107	0.0000	0.0000	0.4191	0.0107	0.0414	0.0016	0.6154	0.0117
	S	full	0.6432	0.0143	0.0044	0.0044	0.6429	0.0143	0.0379	0.0017	0.6450	0.0115
	S	V1	0.5055	0.0151	0.0000	0.0000	0.5054	0.0151	0.0517	0.0020	0.5857	0.0108
	S	V4	0.5555	0.0158	0.0016	0.0010	0.5552	0.0158	0.0485	0.0019	0.6003	0.0117
	S	IT	0.5673	0.0154	0.0027	0.0027	0.5669	0.0154	0.0422	0.0016	0.6148	0.0118
	$S+$	full	0.5483	0.0172	0.0110	0.0094	0.5485	0.0172	**0.0371**	**0.0016**	**0.6453**	**0.0116**
	$S+$	V1	0.4402	0.0163	0.0042	0.0034	0.4403	0.0163	0.0512	0.0020	0.5859	0.0107
	$S+$	V4	0.4674	0.0193	0.0201	0.0113	0.4678	0.0193	0.0481	0.0020	0.6003	0.0117
	$S+$	IT	0.4938	0.0179	0.0101	0.0069	0.4941	0.0179	0.0414	0.0016	0.6153	0.0117
Brain2GAN Natural Objects	W_{gt}	full	0.7095	0.0049	0.0000	0.0000	0.8032	0.0032	0.0818	0.0020	0.4014	0.0105
	W_{gt}	V1	0.6491	0.0057	0.0000	0.0000	0.7666	0.0036	0.0807	0.0019	0.4012	0.0104
	W_{gt}	V4	0.6896	0.0054	0.0000	0.0000	0.7915	0.0034	0.0792	0.0019	0.3930	0.0105
	W_{gt}	IT	0.6807	0.0059	0.0000	0.0000	0.7859	0.0037	0.0838	0.0020	0.3755	0.0108
	W	full	**0.7328**	**0.0048**	**0.0000**	**0.0000**	**0.8229**	**0.0030**	0.0765	0.0020	0.4270	0.0102
	W	V1	0.6632	0.0055	0.0000	0.0000	0.7809	0.0034	0.0759	0.0019	0.4187	0.0103
	W	V4	0.6998	0.0052	0.0000	0.0000	0.8029	0.0032	0.0743	0.0018	0.4107	0.0105
	W	IT	0.6902	0.0058	0.0000	0.0000	0.7970	0.0036	0.0794	0.0019	0.3883	0.0110
	$W+$	full	0.5478	0.0107	0.0048	0.0030	0.5477	0.0107	0.0694	0.0019	**0.4457**	**0.0105**
	$W+$	V1	0.3750	0.0138	0.0341	0.0097	0.3748	0.0138	0.0701	0.0018	0.4333	0.0106
	$W+$	V4	0.4755	0.0113	0.0034	0.0021	0.4756	0.0113	**0.0679**	**0.0017**	0.4290	0.0105
	$W+$	IT	0.4643	0.0116	0.0014	0.0008	0.4642	0.0116	0.0730	0.0018	0.4047	0.0114
	S	full	0.6220	0.0121	0.0048	0.0039	0.6222	0.0121	0.0765	0.0020	0.4270	0.0102
	S	V1	0.4517	0.0173	0.0253	0.0092	0.4519	0.0173	0.0759	0.0019	0.4187	0.0103
	S	V4	0.5497	0.0138	0.0099	0.0056	0.5500	0.0138	0.0743	0.0018	0.4107	0.0105
	S	IT	0.5264	0.0155	0.0082	0.0041	0.5266	0.0155	0.0794	0.0019	0.3883	0.0110
	$S+$	full	0.6062	0.0124	0.0063	0.0041	0.6063	0.0124	0.0694	0.0019	**0.4457**	**0.0105**
	$S+$	V1	0.4407	0.0172	0.0152	0.0065	0.4409	0.0172	0.0701	0.0018	0.4333	0.0106
	$S+$	V4	0.5333	0.0138	0.0149	0.0073	0.5334	0.0138	**0.0679**	**0.0017**	0.4290	0.0105
	$S+$	IT	0.5181	0.0155	0.0146	0.0057	0.5181	0.0155	0.0730	0.0018	0.4047	0.0114
THINGS	W	full	**0.6838**	**0.0102**	**0.0000**	**0.0000**	**0.7803**	**0.0071**	0.0757	0.0030	0.4438	0.0173
	W	V1	0.6362	0.0111	0.0000	0.0000	0.7501	0.0078	0.0784	0.0032	0.4370	0.0169
	W	V4	0.6625	0.0102	0.0000	0.0000	0.7665	0.0071	0.0778	0.0027	0.4076	0.0167
	W	IT	0.6573	0.0106	0.0000	0.0000	0.7633	0.0073	0.0802	0.0030	0.4021	0.0166
	$W+$	full	0.4403	0.0184	0.0133	0.0071	0.4403	0.0184	**0.0666**	**0.0028**	**0.4566**	**0.0175**
	$W+$	V1	0.2825	0.0214	0.0381	0.0148	0.2825	0.0214	0.0692	0.0030	0.4470	0.0172
	$W+$	V4	0.3958	0.0181	0.0198	0.0097	0.3959	0.0181	0.0695	0.0026	0.4130	0.0169
	$W+$	IT	0.3795	0.0200	0.0244	0.0104	0.3795	0.0200	0.0703	0.0027	0.4091	0.0172
	S	full	0.5134	0.0226	0.0209	0.0102	0.5137	0.0226	0.0757	0.0030	0.4438	0.0173
	S	V1	0.3519	0.0262	0.0310	0.0134	0.3520	0.0261	0.0784	0.0032	0.4370	0.0169
	S	V4	0.4732	0.0213	0.0053	0.0032	0.4735	0.0213	0.0778	0.0027	0.4076	0.0167
	S	IT	0.4449	0.0234	0.0161	0.0069	0.4451	0.0234	0.0802	0.0030	0.4021	0.0166
	$S+$	full	0.4764	0.0216	0.0122	0.0101	0.4764	0.0216	**0.0666**	**0.0028**	**0.4566**	**0.0175**
	$S+$	V1	0.3015	0.0272	0.0268	0.0112	0.3016	0.0271	0.0692	0.0030	0.4470	0.0172
	$S+$	V4	0.4363	0.0213	0.0117	0.0081	0.4363	0.0213	0.0695	0.0026	0.4130	0.0169
	$S+$	IT	0.4174	0.0227	0.0077	0.0045	0.4173	0.0227	0.0703	0.0027	0.4091	0.0172

Table 3.1: Quantitative evaluation results for the image datasets in different latent spaces and VC regions. $\sigma_{\bar{x}}$ = standard error, μ = mean, LS = latent space, VC = visual cortex, R = Pearson correlation, P = P-value, CS = cosine similarity, MSE = mean squared error, $SSIM$ = structural similarity index measurement.

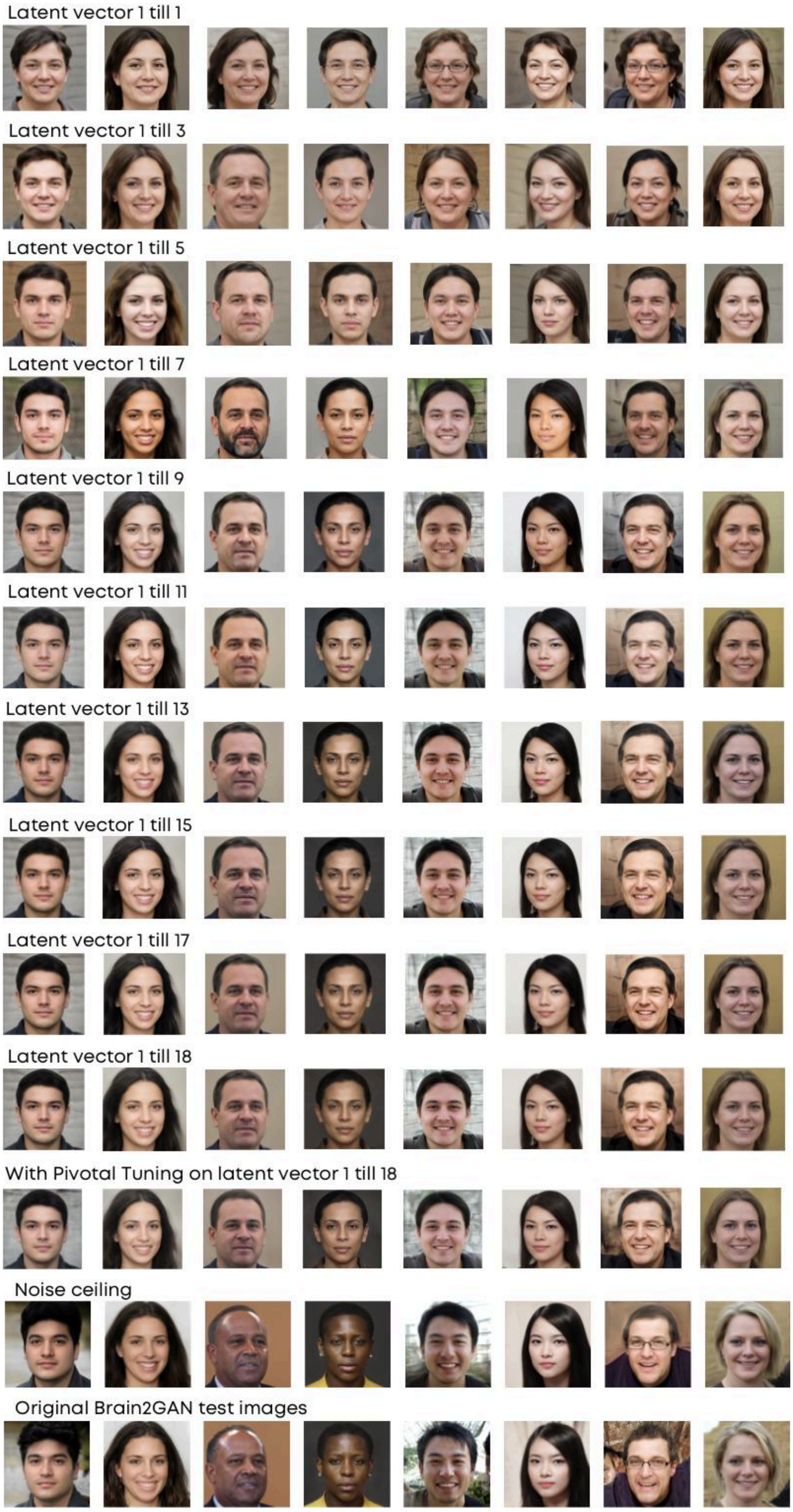


Figure 6.4: W+ cumulative channel reconstruction for Brain2GAN Faces stimuli.

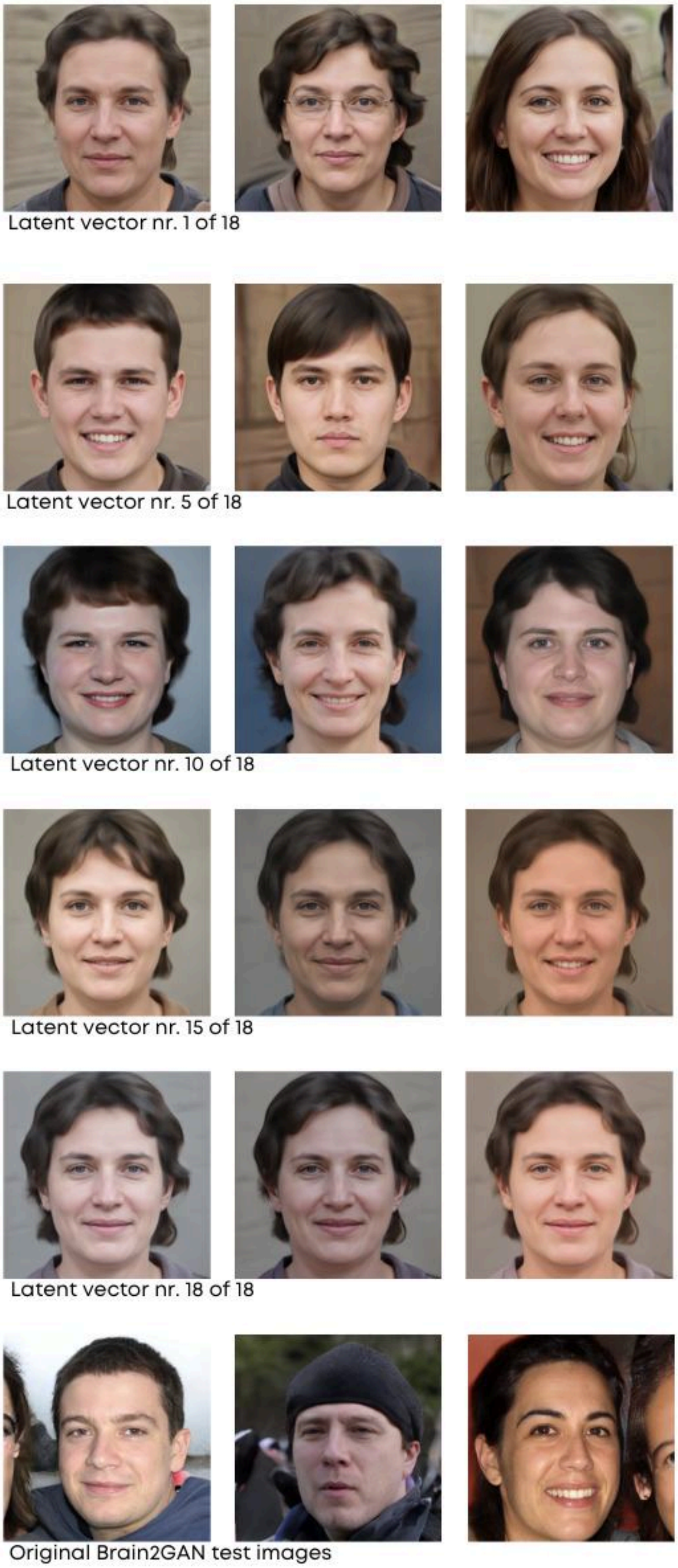

Latent vector nr. 1 of 18
Latent vector nr. 5 of 18
Latent vector nr. 10 of 18
Latent vector nr. 15 of 18
Latent vector nr. 18 of 18
Original Brain2GAN test images

Figure 6.5: W+ individual channel reconstruction for Brain2GAN Faces stimuli.

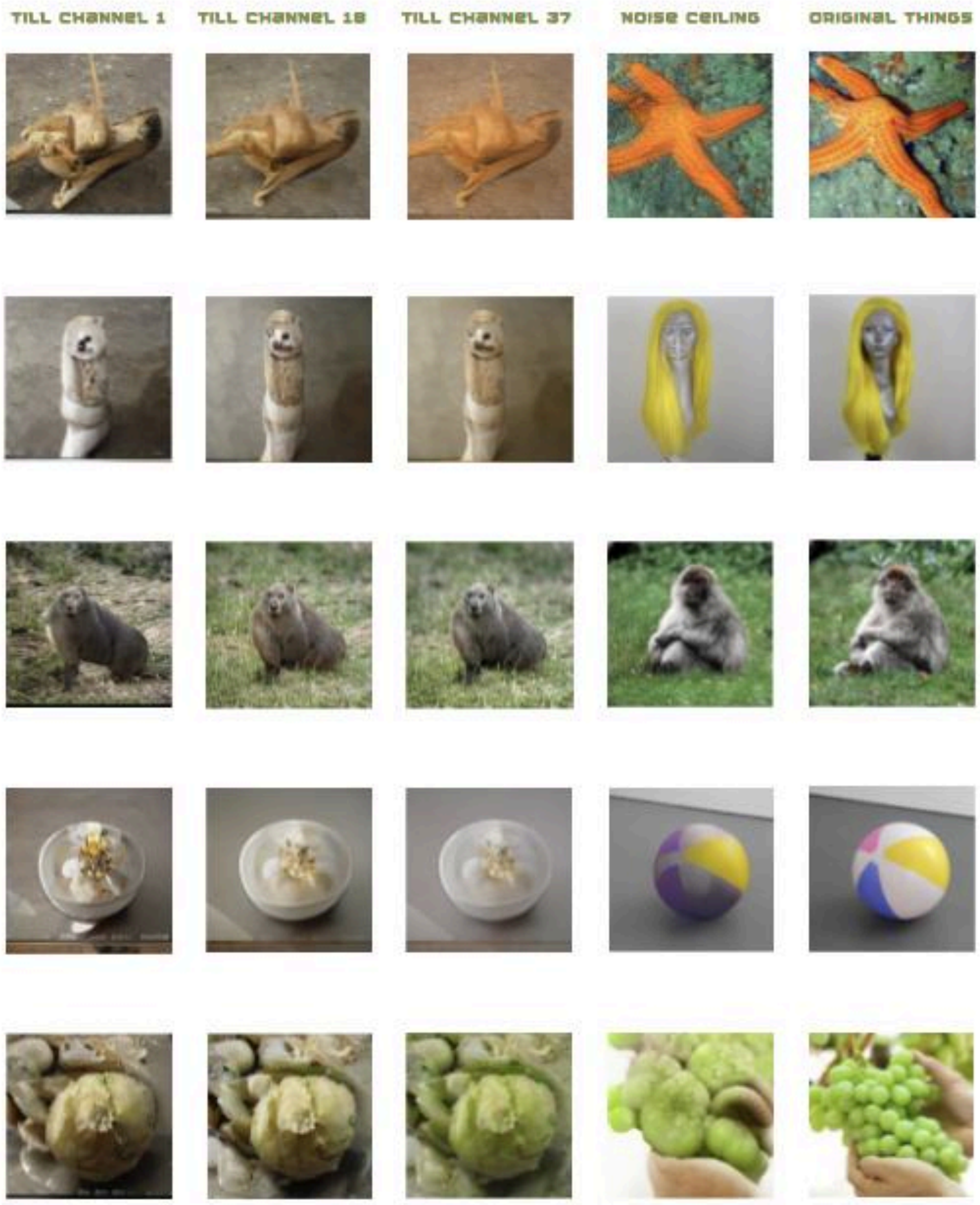


Figure 6.6: W+ cumulative channel reconstruction for THINGS stimuli.