BRAIN2GAN: FEATURE-DISENTANGLED NEURAL CODING OF VISUAL PERCEPTION IN THE PRIMATE BRAIN

Thirza Dado[1*] **Paolo Papale**[2] **Antonio Lozano**[2] **Lynn Le**[1] **Feng Wang**[2]

Marcel van Gerven[1] **Pieter Roelfsema**[2,3,4,5] **Yağmur Güçlütürk**[1] **Umut Güçlü**[1*]

[1] Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, Netherlands. [2] Department of Vision and Cognition, Netherlands Institute for Neuroscience, Amsterdam, Netherlands. [3] Laboratory of Visual Brain Therapy, Sorbonne University, Paris, France. [4] Department of Integrative Neurophysiology, VU Amsterdam, Amsterdam, Netherlands. [5] Department of Psychiatry, Amsterdam UMC, Amsterdam, Netherlands.

*Correspondence: `thirza.dado@donders.ru.nl`, `u.guclu@donders.ru.nl`

ABSTRACT

A challenging goal of neural coding is to characterize the neural representations underlying visual perception. To this end, we analyzed the relationship between multi-unit activity of macaque visual cortex and latent representations of state-of-the-art deep generative models, including feature-disentangled representations of generative adversarial networks (i.e., w-latents of StyleGAN) and language-contrastive representations of latent diffusion networks (i.e., CLIP-latents of Stable Diffusion). A mass univariate neural encoding analysis of the latent representations showed that feature-disentangled representations explain increasingly more variance than the alternative representations over the ventral stream. Subsequently, a multivariate neural decoding analysis of the feature-disentangled representations resulted in state-of-the-art spatiotemporal reconstructions of visual perception. Taken together, our results not only highlight the important role of feature-disentanglement in shaping high-level neural representations underlying visual perception but also serve as an important benchmark for the future of neural coding.

***K*eywords** feature disentanglement · generative adversarial networks · macaque visual cortex · multi-unit activity · neural coding

Figure 1: **Example results.** Stimulus (top) reconstructions (bottom) from brain activity.

NB The images in the figures are AI-generated and do not depict real objects or real subjects.

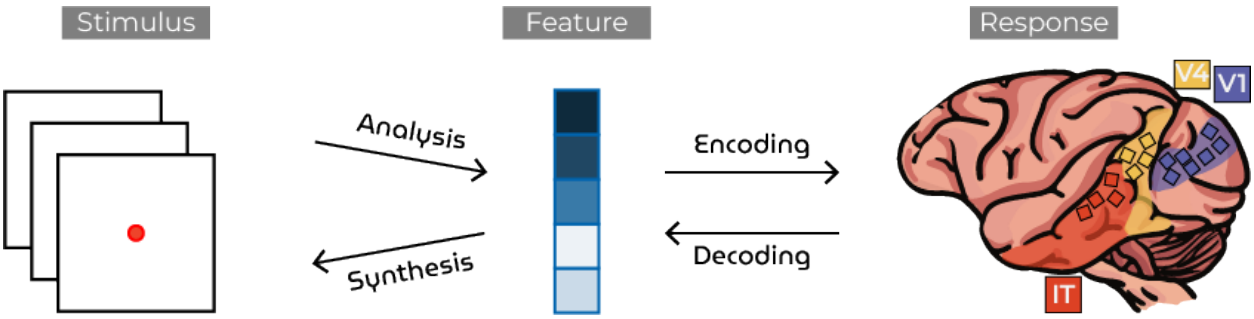

Figure 2: **Neural coding.** The transformation between sensory stimuli and brain responses via an intermediate feature space. Neural encoding is factorized into a nonlinear "analysis" and a linear "encoding" mapping. Neural decoding is factorized into a linear "decoding" and a nonlinear "synthesis" mapping.

1 Introduction

The brain is adept at recognizing a virtually unlimited variety of different visual inputs depicting different faces, objects and scenes in the world. Each unique stimulus creates a unique pattern of brain activity that carries information about that stimulus in some shape or form. However, this stimulus-response transformation remains largely unsolved due to the complexity of multi-layered visual processing in the brain. In the field of neural coding, we aim to characterize this relationship that underlies the brain's ability to recognize the statistical invariances of structured yet complex naturalistic environments. *Neural encoding* seeks to find how properties of external phenomena are stored in the brain [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14], and vice versa, *neural decoding* aims to find what information about the original stimulus is present in and can be retrieved from the recorded brain activity by classification [15, 16, 17], identification [18, 19, 20, 21] or reconstruction [22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34]. In classification, brain activity is taken to predict the category to which the original stimulus belongs, based on a predefined set of categories. In identification, brain activity is utilized to identify the most probable stimulus from a given set of available stimuli. In reconstruction, a literal replica of the original stimulus is recreated which involves the extraction of *specific stimulus characteristics* from neural data. Note that the latter problem is considerably harder as its solution exists in an infinitely large set of possibilities whereas those of classification and identification can be selected from a finite set. In both neural encoding and -decoding, it is common to factorize the direct transformation into two by invoking an in-between feature space (Figure 2). The rationale behind this is twofold:

1. *Efficiency*: modeling the *direct* stimulus-response relationship from scratch requires large amounts of training data (up to the order of millions) which is challenging because neural data is scarce. To work around the problem of data scarcity, we can leverage the knowledge of computational models (typically, deep neural networks that are pretrained on huge datasets) by extracting their feature activations to images and then aligning these with the elicited neural activity to those images during neuroimaging experiments, based on the systematic correspondence between the two.
2. *Interpretability*: the computational model whose (emerged) features align best with neural activity can be informative about what drives the neural processing of the same stimulus. As such, alternative hypotheses can be tested about what drives neural representations themselves (e.g., alternative objective functions and training paradigms). Note that we lose this explanatory property when optimizing a model directly on neural data itself.

The main aim of this study is to characterize high-level neural representations underlying perception, for which we analyzed the relationship between brain responses and various feature representations of recent generative models with different properties such as feature disentanglement and language regularisation, each of which captured a specific set of features and patterns about the visual stimuli. We used the most similar representations to reconstruct perceived stimuli with state-of-the-art quality.

1.1 A feature-disentangled candidate of neural representations

Although neural representations are constructed from experience, an infinite amount of visual phenomena can be represented by the brain to successfully interact with the environment. That is, novel yet plausible situations that respect the regularities of the natural environment can also be mentally simulated or *imagined* [35]. From a machine learning perspective, generative models achieve the same objective by capturing the probability density underlying a huge set of observations. We can sample from this modeled distribution and synthesize new instances that appear as

if they belong to the real data distribution yet are suitably different from the observed instances thereof. Particularly, generative adversarial networks (GANs) [36] are among the most impressive generative models to date which can synthesize novel yet realistic-looking images (e.g., natural images and images of human faces, bedrooms, cars and cats [37, 38, 39, 40] from latent vectors. A GAN consists of two neural networks: a generator network that synthesizes images from randomly-sampled latents and a discriminator network that distinguishes synthesized from real images. During training, these networks are pitted against each other until the generated data are indistinguishable from the real data. The bijective latent-to-image relationship of the generator can be exploited in neural coding to disambiguate the synthesized images since visual content is specified by their underlying latent [41] and perform *analysis by synthesis* [42]. Importantly, traditional GANs are known to suffer from feature entanglement where the generator strictly follows the data distribution it has been trained on (e.g., visual semantics "gender" and "hair length" may be entangled when, during training, the generator has predominantly been exposed to feminine-looking faces with long hair) [43]. As such, biases are inherited from the training dataset that do not necessarily reflect reality (e.g., feminine faces do not always have long hair, and faces with long hair do not always look feminine).

Feature-disentangled GANs have been designed to separate the factors of variation (e.g., color, texture, and shape) in the generated images. One member of the family of feature-disentangled GANs is StyleGAN [39] - which maps the conventional z-latent via an 8-layered multilayer perceptron to an intermediate and less entangled w-latent space. Here, we propose feature-disentangled w-latents as a promising feature candidate to explain neural responses during visual perception. To test this, visual stimuli were synthesized by a feature-disentangled GAN and presented to a macaque with cortical implants in a passive fixation task. In contrast to many previous studies that relied on noninvasive fMRI signals with limited temporal resolution and low signal-to-noise ratio, the current use of intracranial recordings provided opportunities for spatiotemporal analysis of brain activity in unprecedented detail. We performed neural encoding to predict brain activity from StyleGAN's z- and w-latents, as well as latents from Contrastive Language-Image Pre-training (CLIP; ViT-L/14@336px) which represent images and text in a shared representational space that captures their semantic relationships [44]. CLIP latents are used in recent latent diffusion models, such as stable diffusion [45].

First, our encoding analysis revealed that w-latents were the most successful at predicting high-level brain activity in the inferior temporal (IT) cortex, which is located at the end of the visual ventral pathway. Second, neural decoding using w-latents resulted in highly accurate reconstructions that matched the stimuli in their specific visual characteristics. This was done by fitting a decoder to the recorded brain responses and the ground-truth w-latents of the training stimuli. We then used this decoder to predict the w-latents from responses of the held-out test set, and fed these to the generator of the GAN for reconstruction [34]. The high reconstruction performance indicates the importance of feature disentanglement in explaining the neural representations underlying perception, offering a new way forward for the previously limited yet biologically more plausible unsupervised models of brain function. Third, time-based neural decoding showed how meaningful reconstructions unfold in time. Finally, we applied Euclidean vector arithmetic to w-latents as well as brain activity. Taken together, the high quality of the neural recordings and feature representations resulted in novel experimental findings that not only demonstrate how advances in machine learning extend to neuroscience but also will serve as an important benchmark for future research.

1.2 Earlier work

The visual cortex can represent a virtually unlimited number of different images. At an earlier stage, decoding studies largely relied on retinotopy to infer visual content, since the spatial organization of images is reflected in the stimulus-invoked brain responses in the primary visual cortex (V1) [46, 22, 23, 25, 27, 30]. As such, visual content was mainly inferred from the stimulus-invoked brain responses in those early cortical areas. However, these stimuli had to consist of low-resolution contrast patterns or digits that were far from the complex naturalistic images that the brain perceives in daily life. In addition, it is known that visual stimuli are represented with increased complexity across the visual ventral stream. That is, visual experience is partially determined by the selective responses of neuronal populations along the visual ventral "what" pathway [47] where the receptive fields of neurons in early cortical regions are selective for simple features (e.g., local edge orientations [48]) whereas those in more downstream regions respond to more complex patterns of combined features [49, 50]. This complexity gradient in visual processing from low- to high-level features is also identified in DNNs: aligning layer- with neural activations revealed that early layers were mainly predictive of responses in early visual areas whereas deeper layers were more predictive of more downstream visual areas in humans [3, 4, 5, 6, 7, 8, 11] as well as in the primate brain [1, 12].

At present, neural coding has advanced beyond conventional retinotopy, and DNNs are commonly used to decode neural activity during visual perception, imagery and dreaming [51, 17, 52, 32, 53, 31, 33, 34]. To our knowledge, [31, 33, 34] are the most similar studies that also attempted to decode images from brain activity. [31] used the feature representations from VGG16 pretrained on face recognition (i.e., trained in a supervised setting). Although more biologically plausible, unsupervised learning paradigms seemed to appear less successful in modeling neural

representations in the primate brain than their supervised counterparts [5] with the exception of [33] and [34] who used adversarially learned latent representations of a variational autoencoder-GAN (VAE-GAN) and a GAN, respectively. Importantly, [34] used synthesized stimuli to have direct access to the ground-truth latents instead of using post-hoc approximate inference, as VAE-GANs do by design. Here, we adopted and improved this experimental paradigm to study more high-level neural representations during visual perception.

2 Results

We used two datasets of visual stimuli: (i) face images synthesized by StyleGAN3 (pretrained on FFHQ) and (ii) high-variety natural images synthesized by StyleGAN-XL (pretrained on ImageNet).

2.1 Neural encoding

We predicted the neural activity of the individual units in the multi-unit microelectrodes in visual areas V1, V4 and IT from five feature representations: the conventional z-latents, feature-disentangled w-latents, language-regularized CLIP latents and the activations of layers 2 and 7 (after max pooling) of VGG16 as control models for low- and mid-level features. The latter two representations will be referred to as layers 1 and 3 for the remainder of the paper; this numbering system corresponds to the order of the max pooling layers in VGG16, which has a total of five max pooling layers. The feature representation that resulted in the highest encoding performance was assigned to each unit, revealing the well-established complexity gradient where lower-level representations are mainly predictive of responses in earlier visual areas and higher-level representations of more downstream areas (Figure 3). We found that w-latents outperformed the other feature representations in predicting the brain activity in downstream area IT. As such, this representation captures the features or patterns from the visual stimuli that are particularly relevant to high-level neural activity.

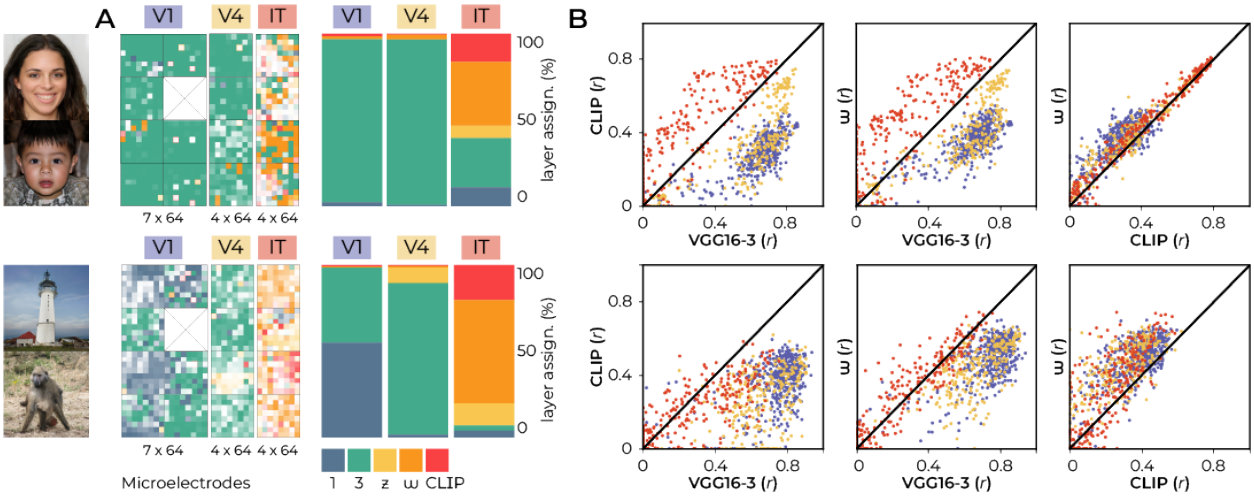

Figure 3: **Neural encoding.** For each individual microelectrode unit, we fit five encoding models to predict its neural response to visual stimuli from low- and mid-level feature representations of VGG16 (pretrained for face or object recognition), z-, w- and CLIP latent representations of the images. **A.** The representation that resulted in the highest encoding performance (Student's t-test) was assigned to each microelectrode unit. Note that there are seven, four and four microelectrode arrays (64 units each) for V1, V4 and IT, respectively. The distribution of assigned features shows a complexity gradient where the more low-level representations are assigned to early brain regions V1 and V4 and more high-level representations to IT with a preference for w-latents. **B.** The scatterplots show the correlations of one model on the X-axis and that of the other model on the Y-axis to show the relationship between the two. Each dot represents the performance of one modeled microelectrode unit in terms of both encoding models. The diagonal represents equal performance between both models. It is clear to see that w and CLIP outperform L3 in IT but not for V1 and V4, and w slightly outperforms CLIP for all three brain regions as the dots lie above the diagonal in the direction of its axis (here, the Y-axis in the second and third plot).

Figure 4: **Qualitative reconstruction results:** The 100 test set stimuli (top row) and their reconstructions from brain activity (bottom row).

Table 1: **Quantitative results.** The upper and lower block display model performance when reconstructing face- and natural images, respectively, in terms of six metrics of perceptual cosine similarity using the five MaxPool layer outputs of VGG16 for face recognition (faces images) / object recognition (natural images) and latent cosine similarity between w-latents of stimuli and their reconstructions ($mean \pm std.error$). The rows display decoding performance when using the recordings from "*all*" recording sites or from a specific brain area.).

		VGG16-1 sim.	VGG16-2 sim.	VGG16-3 sim.	VGG16-4 sim.	VGG16-5 sim.	Lat. sim.
Face images	*All*	**0.7871 ± 0.0102**	**0.7681 ± 0.0075**	**0.5874 ± 0.0075**	**0.6170 ± 0.0085**	**0.5940 ± 0.0104**	**0.5548 ± 0.0045**
	V1	0.6382 ± 0.0079	0.6758 ± 0.0064	0.4891 ± 0.0064	0.5041 ± 0.0083	0.4442 ± 0.0092	0.5022 ± 0.0047
	V4	0.6303 ± 0.0101	0.6729 ± 0.0068	0.4890 ± 0.0068	0.5006 ± 0.0085	0.4191 ± 0.0091	0.5026 ± 0.0040
	IT	*0.7123 ± 0.0110*	*0.7133 ± 0.0073*	*0.5093 ± 0.0073*	*0.5253 ± 0.0087*	*0.4434 ± 0.0096*	*0.5176 ± 0.0039*
Natural images	*All*	**0.4083 ± 0.0036**	**0.3322 ± 0.0036**	**0.2555 ± 0.0025**	**0.2192 ± 0.0043**	**0.2497 ± 0.0066**	**0.8032 ± 0.0032**
	V1	*0.3929 ± 0.0031*	*0.3147 ± 0.0031*	0.2223 ± 0.0019	0.1511 ± 0.0023	0.1367 ± 0.0037	0.7336 ± 0.0036
	V4	0.3790 ± 0.0029	0.3132 ± 0.0029	0.2270 ± 0.0019	0.1641 ± 0.0027	0.1617 ± 0.0045	0.7614 ± 0.0034
	IT	0.3798 ± 0.0026	0.3127 ± 0.0026	*0.2302 ± 0.0020*	*0.1790 ± 0.0039*	*0.1692 ± 0.0057*	*0.7653 ± 0.0039*

2.2 Neural decoding

Neural decoding of neural activity via feature-disentangled w-latents resulted in highly accurate reconstructions that closely resembled the stimuli in their specific characteristics; Figure 4, 5. Perceptually, we see high similarity between stimuli and their reconstructions in terms of their specific attributes (e.g., gender, age, pose, haircut, lighting, hair color, skin tone, smile and eyeglasses for faces; shapes, colors, textures, object locations, (in-)animacy for natural images). In addition, we also repeated the experiment with another macaque that had silicon-based electrodes in V1, V2, V3 and V4; see Appendix A.

The quantitative metrics in Table 1 show the similarity between stimuli and their reconstructions from brain activity in terms of six metrics that evaluated reconstruction quality at different levels of abstraction. Specifically, a stimulus and its reconstruction were both fed to VGG16 (pretrained on face- and object recognition for faces and natural images, respectively) and we extracted five intermediate activations (the five MaxPool layers) thereto. The early layers capture more low-level features (e.g., edges and orientations) whereas deeper layers capture increasingly higher-level features

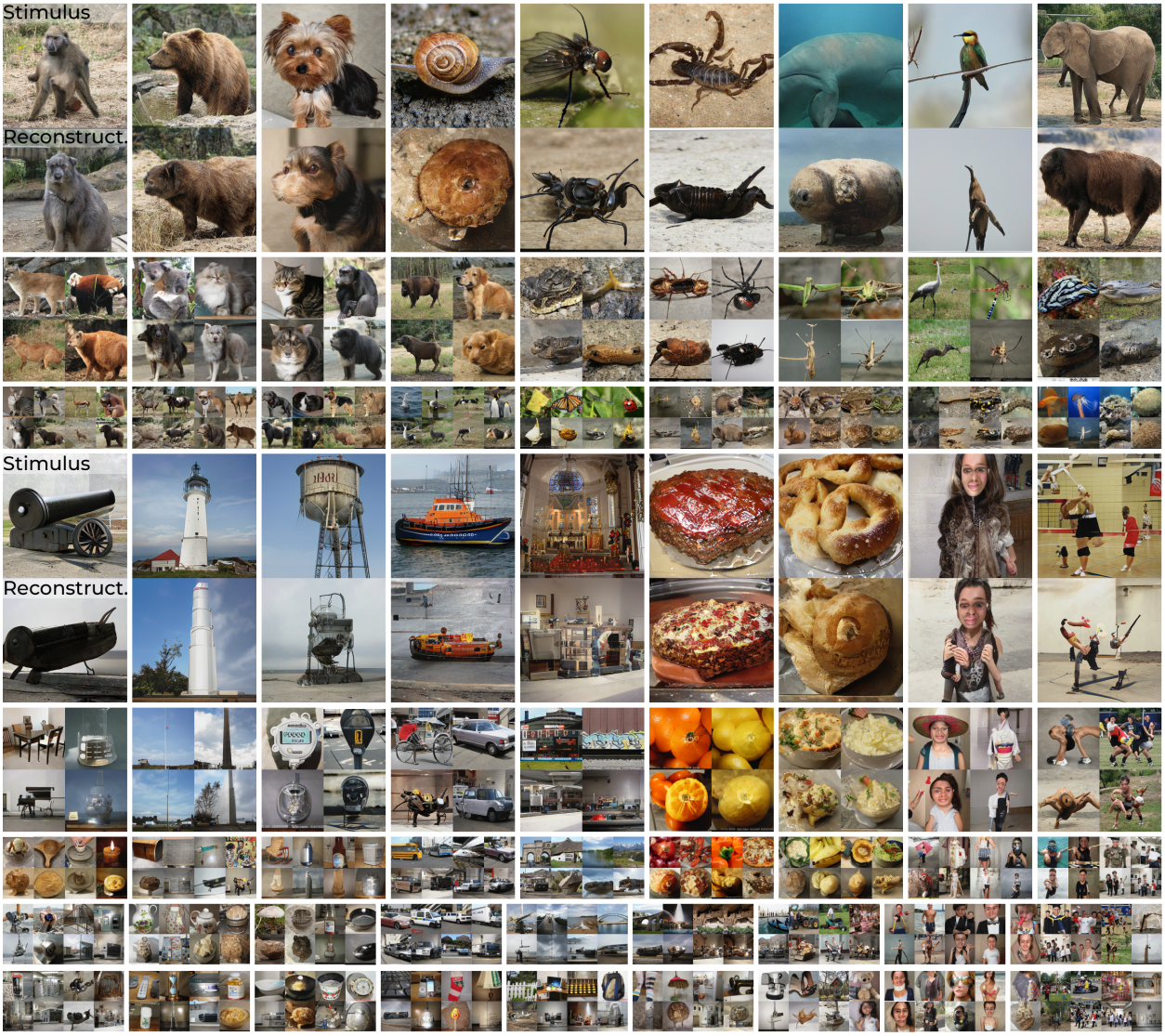

Figure 5: **Qualitative reconstruction results:** The 200 test set stimuli (top row) and their reconstructions from brain activity (bottom row).

(e.g., textures to object parts to entire objects). We then compared the cosine similarity between these extracted representations of stimulus and reconstruction. Next, to study the decoder that resulted in these accurate reconstructions, the contribution of each visual area was determined by the occlusion of the electrode recordings in the other two brain areas (rather than fitting three independent decoders on subsets of brain activity). It is reasonable to say that, of the three cortical areas, the area that resulted in the highest similarity contains the most information about that representation. For faces, decoding performance was for the largest part determined by responses from IT - which is the most downstream site we recorded from. For natural images, we found that the lower-level representations (VGG16 layers 1-2) were most similar when decoded from V1 and the higher-level representations (VGG16 layers 3-5) and latent space were most similar when decoded from area IT. We validated our quantitative results with a permutation test as follows: per iteration, we sampled a hundred/two-hundred random latents from the same distribution as our original test set and generated their corresponding images. We assessed whether these random latents and images were closer to the ground-truth latent and images than our predictions from brain activity, and found that our predictions from brain activity were always closer to the original stimuli than the random samples for all metrics, yielding statistical significance ($p < 0.001$). The charts showing the six similarity metrics over iterations for the random samples and our predictions based on brain activity can be found in Appendix B. We also provided a visual guide for all nine evaluation metrics in Appendix C by showing the stimulus-reconstruction pairs with the five highest- and the five lowest similarities for each metric.

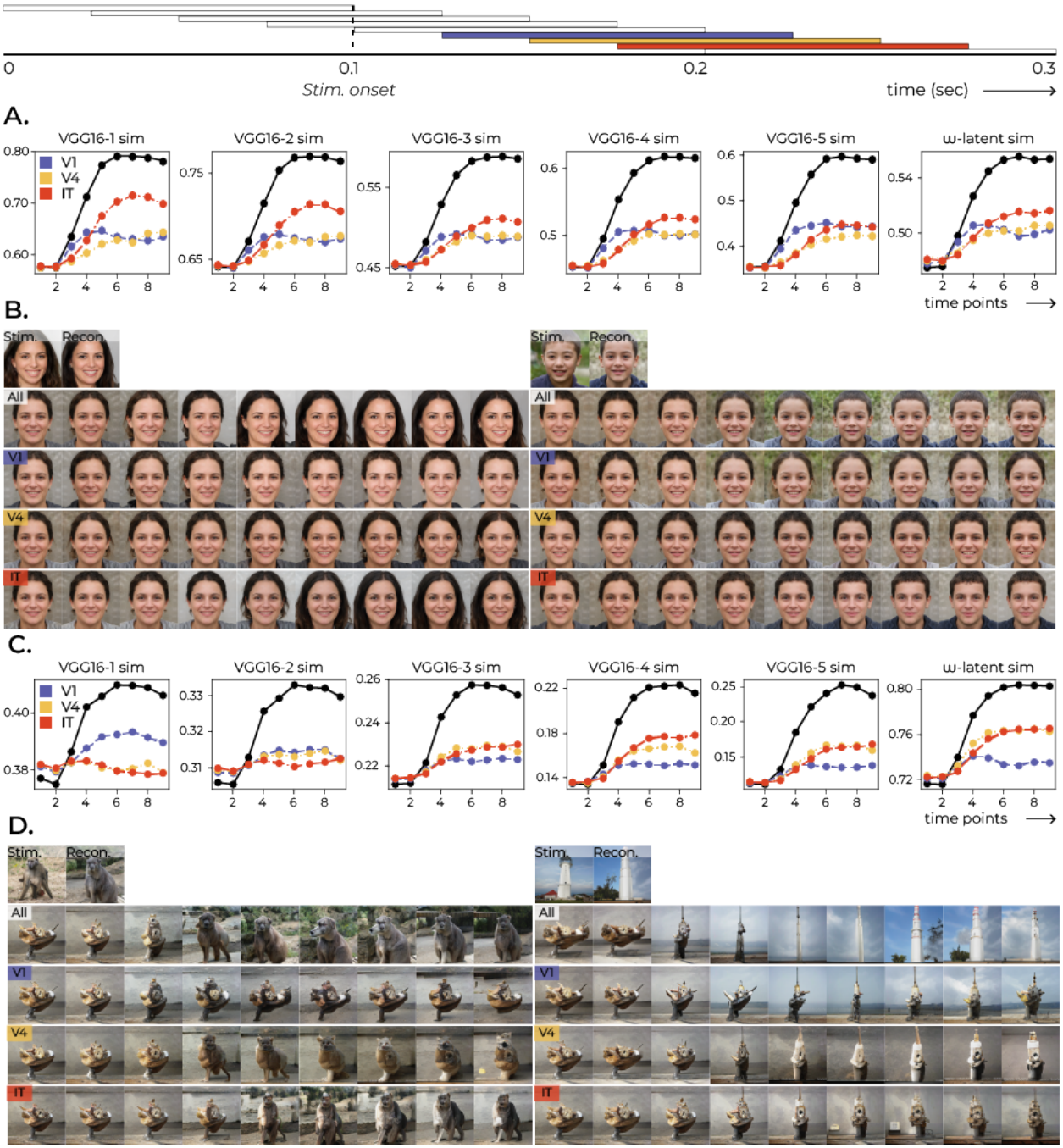

Figure 6: **Time-based decoding.** For each trial, responses were recorded for 300 ms with stimulus onset at 100 ms. Rather than taking the average response within the *original* time windows (see the three color-coded windows for V1, V4 and IT), we applied a sliding window of 100 ms with a stride of 25 ms over the entire time course which resulted in nine average responses in time. **A.** and **C.** show decoding performance over time for faces and natural images, respectively. It can be noted how V1 performance climbs the slightly earlier than the other two visual areas. For faces, IT outperforms V1 and V4 in most instances. For natural images, V1 outperforms the others for low-level feature similarity, after which V4 and IT climb up together and outperform V1 for the more high-level feature similarity metrics. **B.** and **D.** show how two stimulus-reconstruction examples evolve over time.

Figure 6 illustrates time-based neural decoding. Prior to stimulus onset, average-looking images were reconstructed from brain activity at baseline (no visual stimulus). After stimulus presentation ($t = 100$ ms), the reconstructions

Figure 7: *(row 1)* Linear interpolation between two ground-truth w-latents, *(row 2)* two predicted w-latents (decoded from brain activity), *(row 3)* two recorded neural responses, which we then decoded into w-latents, and *(row 4)* two predicted neural responses from VGG16 layer 4 features (neural encoding) which we then decoded into w-latents (neural decoding). The (interpolated) w-latents were fed to the generator to obtain the corresponding images. (a) face images, also contains vector arithmetic. (b) As for (a) but for natural images.

started to take on an appearance over time that resembled the stimulus as meaningful information was extracted from the neural activity. The area-based reconstructions and performance graphs revealed that V1 generally displayed stimulus-like visual features earlier whereas IT consistently outperformed the other two in the final reconstruction of stimulus information.

Lastly, GAN-based decoding enables us to traverse and explore the neural manifold via linear operations applied to latent space (Figure 7). These operations directly translate to meaningful perceptual changes in the generated images such that visual data that look perceptually similar in terms of certain features are also closely positioned in latent space. As such, pathways through the latent landscape were explored by *interpolation* between two (distinct) latents which resulted in an ordered set of images whose semantics varied smoothly with latent codes [54] and simple *arithmetic* operations [55] (Figure 7B) for the faces dataset. We applied spherical rather than regular linear interpolation to account for the spherical geometry of the latent space (multimodal Gaussian distribution). In addition to the original and predicted w-latents, we also applied these linear operations to the recorded and predicted brain responses: we linearly interpolated two brain responses to two stimuli, fed the resulting responses to our decoder, and reconstructed

the corresponding images. The predicted brain activity was encoded from VGG16 layer 4 activations. We found that interpolation and arithmetic in neural space led to faces that were perceptually similar to those from w-latent space. This suggests that the neural- and w-latent space are organized similarly such that we can navigate the former via the latter and access neural responses to arbitrary visual features via linear operation.

3 Discussion

This study characterized the high-level neural representations of visual perception using feature-disentangled latent representations. Our encoding analysis showed w-latents conditioned on StyleGAN3/StyleGAN-XL to outperform the other candidate representations in explaining the variance of neural activity in IT. We then used the same w-latent space for neural decoding of the recorded brain activity, which resulted in reconstructions that strongly resembled the original stimuli in their specific characteristics. Given the virtually infinite number of possible candidate representations to encode the same image, finding a representation that accurately reflects the information in brain activity is not a trivial task. In our approach, the decoded w-latents resulted in image reconstructions that closely matched the stimuli in their semantic as well as structural features. Overall, this work highlights the importance of feature disentanglement in explaining high-level neural responses and demonstrates the potential of aligning unsupervised generative models with biological processes. These findings have implications for the advancements of computational models and the development of clinical applications for people with disabilities. For instance, neuroprosthetics to restore vision in blind patients as well as brain-computer interfaces (BCIs) to enable nonmuscular communication with individuals who are locked-in.

3.1 Uncovering principles of neural coding

The primary goal of our study was to uncover the principles that govern neural coding of the visual world and gain a more interpretable understanding of high-level neural representations underlying visual perception using deep generative modeling. First of all, StyleGAN was designed to *disentangle* the different visual semantics into separate w-latent features. The superior performance of StyleGAN in neural coding highlights the crucial role of feature disentanglement in explaining the high-level neural representations underlying perception, and the ability to disentangle the object manifold [56]. Note that StyleGAN itself has never been optimized on neural data which implies a general principle of shared encoding of real-world phenomena. Second, the similarity between w-latents and the brain could provide further insights into what drives the organization of visual processing in the brain. For instance, GANs are trained in an unsupervised setting; they learn directly from raw visual data without explicit labels or annotations. Not only does this make GANs more biologically plausible than their supervised counterparts since it resembles more closely how the brain learns from its environment but they may also lead to more flexible and generalizable representations that are better able to capture the underlying structure and patterns in the observed data. In contrast, supervised models can categorize but may not capture the full range of visual features and nuances that are present in the data. Finally, there is a conceptual analogy between the adversarial training of GANs and the predictive coding theory of perception where the brain uses top-down predictions, based on prior knowledge and experience, to guide bottom-up sensory processing and adjusts its internal models based on the mismatch between expectations and actual observations. In GANs, the discriminator and generator engage in a similar process with the discriminator evaluating the "real" sensory input and the "predicted/imagined" instances by the generator. Based on the mismatch, as determined by the discriminator, the internal model of the generator is refined such that its outputs match the real-world data closer; the generator harnesses the knowledge of the discriminator to learn how to represent the world in its latent space. So, while the exact mechanisms used by the brain and GANs differ significantly, their conceptual similarities could provide insights into the nature of perception and the potential of machine learning to capture some of the same principles underlying this ability.

3.2 Limitations and future directions

This study solely used synthesized stimuli with known latent representations generated by StyleGAN. While this allowed for a controlled and systematic examination of neural representations of visual information, future studies should also include real photographs to see how this method generalizes. That said, this study still performed valid neural encoding and reconstruction from brain activity despite the nature of the presented images themselves. Another limitation is the small sample size of one subject (note that we did include face reconstructions from a second subject with different cortical implants in the supplementary materials). Although small sample sizes are common in studies using invasive recordings - larger sample sizes are needed to further confirm the robustness of our findings . Finally, it is worth noting that the use of deep neural networks to model brain activity is still a developing field and the models used in this study are not flawless representations of the underlying neural processes.

4 Methods and Materials

4.1 Stimuli

StyleGAN [39, 57] was developed to optimize control over the semantics in the synthesized images in single-category datasets (e.g., only-faces, -bedrooms, -cars or -cats) [40]. This generative model maps z-latents via an 8-layer MLP to an intermediate w-latent space in favor of feature disentanglement. That is, the original z-latent space is restricted to follow the data distribution that it is trained on (e.g., old-looking faces wear eyeglasses more often than young-looking faces), and such biases are entangled in the z-latents. The less entangled w-latent space overcomes this such that unfamiliar latent elements can be mapped to their respective visual features.

Dataset i: Face images. We synthesized photorealistic face images of 1024×1024 px resolution from (512-dim.) z-latent vectors with the generator network of StyleGAN3 (Figure 8) which is pretrained on the high-quality Flickr-Faces-HQ (FFHQ) dataset [39]. The z-latents were randomly sampled from the standard Gaussian. We specified a truncation of 0.7 so that the sampled values are ensured to fall within this range to benefit image quality. During synthesis, learned affine transformations integrate w-latents into the generator network with adaptive instance normalization (like *style transfer* [58]) as illustrated in Figure 8. Finally, we synthesized a training set of 4000 face images that were each presented once to cover a large stimulus space to fit a general model. The test set consisted of 100 synthesized faces that were averaged over twenty repetitions.

Dataset ii: Natural images. Recently, StyleGAN-XL (three times larger in depth and parameter count than a standard StyleGAN3) was developed to scale up to larger and less-structured datasets using a new training strategy [59]. Concretely, the new training strategy combined (i) *the progressive growing paradigm* where architecture size is gradually increased by adding new layers, (ii) *the projected GAN paradigm* where both synthesized and real samples are mapped to four fixed feature spaces before being fed to four corresponding and independent discriminator networks, and (iii) *classifier guidance* where the cross-entropy loss of a pretrained classifier is added as a term to the generator loss. As such, StyleGAN-XL has been successfully trained on ImageNet [60] to generate high-resolution images of a thousand different categories, resulting in a complex and diverse stimulus dataset. StyleGAN-XL was used to synthesize 512×512 px resolution RGB images from (512-dim.) z-latent vectors. The training set z-latents were randomly sampled from the standard Gaussian, mapped to w-latents that were truncated at 0.7 to support image quality as well as diversity. The average w-latent of each category was utilized for the test set due to the high quality and because variation was not required as we only used one image per category. In total, the training and test set consisted of 4000 and 200 stimuli, respectively. We synthesized images from the 200 classes from Tiny ImageNet (a subset rather than all thousand classes from ImageNet) so that each class was represented by twenty training set stimuli and one test set stimulus. The 200 category labels are listed in Appendix D.

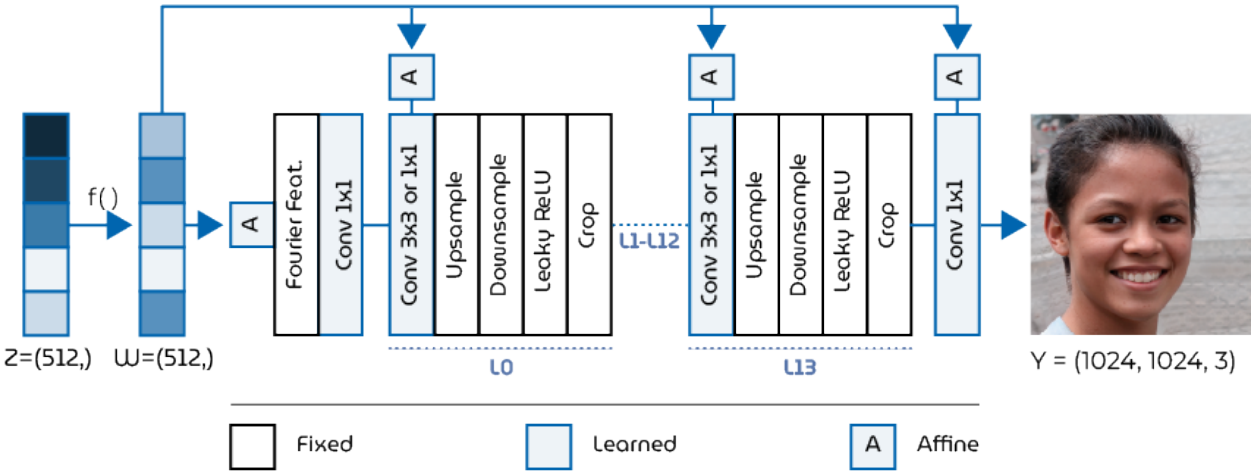

Figure 8: **StyleGAN3 generator architecture.** The generator takes a 512-dim. latent vector as input and transforms it into a 1024×1024 px resolution RGB image. We collected a dataset of 4000 training set images and 100 test set images.

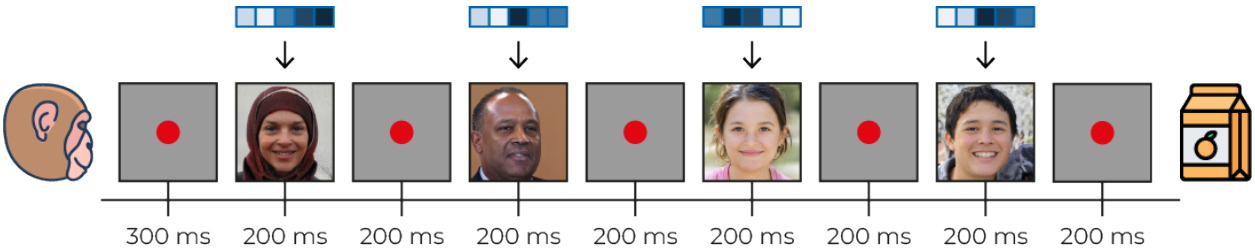

Figure 9: **Passive fixation task.** The monkey was fixating a red dot with gray background for 300 ms followed by a fast sequence four face images (500^2 pixels): 200 ms stimulus presentation and 200 ms inter-trial interval. The stimuli were slightly shifted to the lower right such that the fovea corresponded with pixel (150, 150). The monkey was rewarded with juice if fixation was kept for the whole sequence.

4.2 Features

As the in-between feature candidates we used z-, $w-$, and CLIP latent representations as well as layer activations of VGG16 for face recognition [61] and object recognition [62]. Because the layer 1 and 2 features from VGG16 were very large ($\sim 10^6$), we performed downsampling, as done in [11]. That is, for each channel in the activation, the feature map was spatially smoothed with a Gaussian filter and subsampled with a factor 2. The kernel size was set to be equal to the downsampling factor.

4.3 Responses

We recorded multi-unit activity [63] with 15 chronically implanted electrode arrays (64 channels each) in one macaque (male, 7 years old) upon presentation with images in a passive fixation experiment (Figure 9). Neural responses were recorded in V1 (7 arrays), V4 (4 arrays) and IT (4 arrays) leading to a total of 960 channels (see electrode placings in Figure 2). For each trial, we averaged the early response of each channel using the following time-windows: 25-125 ms for V1, 50-150 ms for V4 and 75-175 ms for IT. The data was normalized as in [64] such that for each channel, the mean was subtracted from all the responses which were then divided by the standard deviation. All procedures complied with the NIH Guide for Care and Use of Laboratory Animals and were approved by the local institutional animal care and use committee of the Royal Netherlands Academy of Arts and Sciences.

To determine the contribution of the activity in each brain region to the overall model performance, we evaluated the decoder using partially occluded test set data. Concretely, we used our main decoder which was trained on neural data from all three brain areas and evaluated it using test set recordings from one brain area. To do this, the responses from the other two areas were occluded by the average response of all but the corresponding response. Alternatively, one could also evaluate the contribution per region by training three *independent* decoders on subsets of neural data (V1-only, V4-only and IT-only) which would allow for evaluation of the contribution of each brain area independently of one another. But in our case, we used the occlusion approach to investigate the area-specific contribution to the *same* decoder's performance by keeping the contributions from the other two areas constant.

4.4 Models

We used linear mapping to evaluate our claim that the feature- and neural representation effectively encode the same stimulus properties, as is standard in neural coding [65, 6]. A more complex nonlinear transformation would not be valid to support this claim since nonlinearities will fundamentally change the underlying representations.

4.4.1 Encoding

Kernel ridge regression was used to model how every recording site in the visual cortex is linearly dependent on the stimulus features. That is, an encoding model is defined for each electrode. Encoding required regularization to avoid overfitting since we predicted from feature space $\mathbf{x}_i \rightarrow \phi(\mathbf{x}_i)$ where $\phi()$ is the feature extraction model. Hence we used ridge regression where the norm of $\mathbf{w}$ is penalized to define encoding models by a weighted sum of $\phi(\mathbf{x_i})$:

$$\mathcal{L} = \frac{1}{2}\sum_{i=1}^{N}\left(y_i - \mathbf{w}^T\phi(\mathbf{x}_i)\right)^2 + \frac{1}{2}\lambda||\mathbf{w}||^2 \tag{1}$$

where $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)^T \in \mathbb{R}^{N \times d}$, $y = (y_1, y_2, \ldots, y_N)^T \in \mathbb{R}^{N \times 1}$, N the number of stimulus-response pairs, d the number of pixels, and $\lambda \geq 0$ the regularization parameter. We then solved for $\mathbf{w}$ by applying the "kernel trick" [66]:

$$\mathbf{w} = (\lambda \mathbf{I}_q + \Phi\Phi^T)^{-1}\Phi y \tag{2}$$

where $\Phi = (\phi(x_1), \phi(x_2), \ldots, \phi(x_N))^T \in \mathbb{R}^{N \times q}$ (i.e., the design matrix) where q is the number of feature elements and $y = (y_1, y_2, \ldots, y_N)^T \in \mathbb{R}^{N \times 1}$. This means that $\mathbf{w}$ must lie in the space induced by the training data even when $q \gg N$. The optimal λ is determined with grid search, as in [2]. The grid is obtained by dividing the domain of λ in M values and evaluating model performance at every value. This hyperparameter domain is controlled by the capacity of the model, i.e., the effective degrees of freedom dof of the ridge regression fit from $[1, N]$:

$$\text{dof}(\lambda_j) = \sum_{i=1}^{N} \frac{s_i^2}{s_i^2 + \lambda_j} \tag{3}$$

where s are the non-zero singular values of the design matrix Φ as obtained by singular value decomposition. We can solve for each λ_j with Newton's method. Now that the grid of lambda values is defined, we can search for the optimal λ_j that minimizes the 10-fold cross validation error.

4.4.2 Decoding

Multiple linear regression was used to model how the individual units within feature representations y_i (e.g., w_i-latents) are linearly dependent on brain activity $\mathbf{x_i}$ per electrode:

$$\mathcal{L} = \frac{1}{2}\sum_{i=1}^{N} \left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2 \tag{4}$$

where i ranges over samples. We reconstructed the images by feeding the predicted latents to brain responses of the test set by feeding them to the generator without truncation.

4.5 Evaluation

Decoding performance was evaluated by six metrics that compared the stimuli from the held-out test set with their reconstructions from brain activity: perceptual cosine similarity using the five MaxPool layer outputs of VGG16 and latent cosine similarity. For *perceptual cosine similarity*, we computed the cosine similarity between layer activations (rather than pixel space which is the model input) extracted by VGG16 pretrained for object recognition. This metric reflects human perception of similarity better because it takes more high-level visual cues into account (e.g., color, texture, and spatial information) and human perception is often not directly related to the pixel values themselves. Specifically, we fed the stimuli and their reconstructions to the DNN and then considered the cosine similarity per activation unit:

$$S_p(x, \hat{x}) = \frac{f(\hat{x})_i \cdot f(x)_i)}{\sqrt{\sum_{i=1}^{n}(f(\hat{x})_i)^2}\sqrt{\sum_{i=1}^{n}(f(x)_i)^2}}$$

where x and $\hat{x}$ are the visual stimuli and their reconstructions, respectively, n the number of activation elements, and $f(.)$ the image-activation transformation. For *latent similarity*, we considered the cosine similarity per latent dimension between predicted and ground-truth latent vectors:

$$S_l(w, \hat{w}) = \frac{\hat{z}_i \cdot z_i}{\sqrt{\sum_{i=1}^{512}(\hat{z}_i)^2}\sqrt{\sum_{i=1}^{512}(z_i)^2}}$$

where $\hat{w}$ and w are the 512-dimensional predicted and ground-truth feature-disentangled latent vectors, respectively.

4.6 Implementation details

All analyses were carried out in Python 3.8 on a cloud-based virtual machine with Intel(R) Xeon(R) CPU @ 2.20GHz and NVIDIA Tesla T4 GPU (Driver Version: 510.47.03, CUDA Version: 11.6) on a Linux-based operating system. We used the original PyTorch implementations of StyleGAN3, StyleGAN-XL, VGG16 (face recognition, object recognition). Our implementations of neural encoding and -decoding can be found on the GitHub repository.

5 Ethics statement

In conjunction with the evolving field of neural decoding grows the concern regarding mental privacy [67]. Because we think it is likely that access to subjective experience will be possible in the foreseeable future, we want to emphasize that it is important to at all times strictly follow the ethical rules and regulations regarding data extraction, storage, and protection. It should never be possible to invade the subjective contents of the mind, as this would violate privacy rights and ethical standards. The importance of responsible and ethical use of neural decoding technology cannot be overstated, and all efforts must be made to ensure the protection of individual mental privacy.

References

[1] Winrich A Freiwald and Doris Y Tsao. Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science*, 330(6005):845–851, 2010.

[2] Umut Güçlü and Marcel van Gerven. Unsupervised feature learning improves prediction of human brain activity in response to natural images. *PLoS computational biology*, 10(8):e1003724, 2014.

[3] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

[4] Charles F Cadieu, Ha Hong, Daniel LK Yamins, Nicolas Pinto, Diego Ardila, Ethan A Solomon, Najib J Majaj, and James J DiCarlo. Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS computational biology*, 10(12):e1003963, 2014.

[5] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11):e1003915, 2014.

[6] Umut Güçlü and Marcel van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015.

[7] Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.

[8] Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, 6(1):1–13, 2016.

[9] Umut Güçlü, Jordy Thielen, Michael Hanke, and Marcel van Gerven. Brains on beats. *Advances in Neural Information Processing Systems*, 29, 2016.

[10] Marcel van Gerven. A primer on encoding models in sensory neuroscience. *Journal of Mathematical Psychology*, 76:172–183, 2017.

[11] Michael Eickenberg, Alexandre Gramfort, Gaël Varoquaux, and Bertrand Thirion. Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152:184–194, 2017.

[12] Le Chang and Doris Y Tsao. The code for facial identity in the primate brain. *Cell*, 169(6):1013–1028, 2017.

[13] Umut Güçlü and Marcel van Gerven. Probing human brain function with artificial neural networks. *Computational Models of Brain and Behavior*, pages 413–423, 2017.

[14] Katja Seeliger, Matthias Fritsche, Umut Güçlü, Sanne Schoenmakers, J Schoffelen, Sander Bosch, and M van Gerven. Convolutional neural network-based encoding and decoding of visual object recognition in space and time. *NeuroImage*, 180:253–266, 2018.

[15] James V Haxby, M Ida Gobbini, Maura L Furey, Alumit Ishai, Jennifer L Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430, 2001.

[16] Yukiyasu Kamitani and Frank Tong. Decoding the visual and subjective contents of the human brain. *Nature neuroscience*, 8(5):679–685, 2005.

[17] Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature communications*, 8(1):1–15, 2017.

[18] Tom M Mitchell, Svetlana V Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L Malave, Robert A Mason, and Marcel Adam Just. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195, 2008.

[19] Kendrick N Kay, Thomas Naselaris, Ryan J Prenger, and Jack L Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, 2008.

[20] Umut Güçlü and Marcel van Gerven. Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. *NeuroImage*, 145:329–336, 2017.

[21] Umut Güçlü and Marcel van Gerven. Modeling the dynamics of human brain activity with recurrent neural networks. *Frontiers in computational neuroscience*, 11:7, 2017.

[22] Bertrand Thirion, Edouard Duchesnay, Edward Hubbard, Jessica Dubois, Jean-Baptiste Poline, Denis Lebihan, and Stanislas Dehaene. Inverse retinotopy: inferring the visual content of images from brain activation patterns. *NeuroImage*, 33(4):1104–1116, 2006.

[23] Yoichi Miyawaki, Hajime Uchida, Okito Yamashita, Masa-aki Sato, Yusuke Morito, Hiroki C Tanabe, Norihiro Sadato, and Yukiyasu Kamitani. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, 60(5):915–929, 2008.

[24] Thomas Naselaris, Ryan J Prenger, Kendrick N Kay, Michael Oliver, and Jack L Gallant. Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63(6):902–915, 2009.

[25] Marcel van Gerven, Floris P de Lange, and Tom Heskes. Neural decoding with hierarchical generative models. *Neural computation*, 22(12):3127–3142, 2010.

[26] Shinji Nishimoto, An T Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current biology*, 21(19):1641–1646, 2011.

[27] Sanne Schoenmakers, Markus Barth, Tom Heskes, and Marcel Van Gerven. Linear reconstruction of perceived images from human brain activity. *NeuroImage*, 83:951–961, 2013.

[28] Umut Güçlü and Marcel van Gerven. Unsupervised learning of features for bayesian decoding in functional magnetic resonance imaging. In *Belgian-Dutch Conference on Machine Learning*, 2013.

[29] Alan S Cowen, Marvin M Chun, and Brice A Kuhl. Neural portraits of perception: reconstructing face images from evoked brain activity. *NeuroImage*, 94:12–22, 2014.

[30] Changde Du, Changying Du, and Huiguang He. Sharing deep generative representation for perceived image reconstruction from human brain activity. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 1049–1056. IEEE, 2017.

[31] Yağmur Güçlütürk, Umut Güçlü, Katja Seeliger, Sander Bosch, Rob van Lier, and Marcel van Gerven. Reconstructing perceived faces from brain activations with deep adversarial neural decoding. *Advances in neural information processing systems*, 30, 2017.

[32] Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLoS computational biology*, 15(1):e1006633, 2019.

[33] Rufin VanRullen and Leila Reddy. Reconstructing faces from fmri patterns using deep generative neural networks. *Communications biology*, 2(1):1–10, 2019.

[34] Thirza Dado, Yağmur Güçlütürk, Luca Ambrogioni, Gabriëlle Ras, Sander Bosch, Marcel van Gerven, and Umut Güçlü. Hyperrealistic neural decoding for reconstructing faces from fmri activations via the gan latent space. *Scientific reports*, 12(1):1–9, 2022.

[35] Nadine Dijkstra, Sander Bosch, and Marcel van Gerven. Shared neural mechanisms of visual perception and imagery. *Trends in Cognitive Sciences*, 23(5):423–434, 2019.

[36] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[37] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[38] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[39] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[40] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34, 2021.

[41] Nikolaus Kriegeskorte. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, 1:417–446, 2015.

[42] Alan Yuille and Daniel Kersten. Vision as bayesian inference: analysis by synthesis? *Trends in cognitive sciences*, 10(7):301–308, 2006.

[43] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9243–9252, 2020.

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[46] Martin I Sereno, AM Dale, JB Reppas, KK Kwong, JW Belliveau, TJ Brady, BR Rosen, and RBH Tootell. Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science*, 268(5212):889–893, 1995.

[47] L.G. Ungerleider and M. Mishkin. Two cortical visual systems. In *Analysis of visual behavior*, pages 549–586–. MIT Press, Cambridge, MA, 1982.

[48] David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106–154, 1962.

[49] Charles G Gross, CE de Rocha-Miranda, and DB Bender. Visual properties of neurons in inferotemporal cortex of the macaque. *Journal of neurophysiology*, 35(1):96–111, 1972.

[50] Chou P Hung, Gabriel Kreiman, Tomaso Poggio, and James J DiCarlo. Fast readout of object identity from macaque inferior temporal cortex. *Science*, 310(5749):863–866, 2005.

[51] Tomoyasu Horikawa and Yukiyasu Kamitani. Hierarchical neural representation of dreamed objects revealed by brain decoding with deep neural network features. *Frontiers in computational neuroscience*, 11:4, 2017.

[52] Ghislain St-Yves and Thomas Naselaris. Generative adversarial networks conditioned on brain activity reconstruct seen images. In *2018 IEEE international conference on systems, man, and cybernetics (SMC)*, pages 1054–1061. IEEE, 2018.

[53] Guohua Shen, Kshitij Dwivedi, Kei Majima, Tomoyasu Horikawa, and Yukiyasu Kamitani. End-to-end deep image reconstruction from human brain activity. *Frontiers in Computational Neuroscience*, page 21, 2019.

[54] Hang Shao, Abhishek Kumar, and P Thomas Fletcher. The riemannian geometry of deep generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 315–323, 2018.

[55] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[56] James J DiCarlo and David D Cox. Untangling invariant object recognition. *Trends in cognitive sciences*, 11(8):333–341, 2007.

[57] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.

[58] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017.

[59] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022.

[60] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009.

[61] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In Xianghua Xie, Mark W. Jones, and Gary K. L. Tam, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 41.1–41.12. BMVA Press, September 2015.

[62] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[63] Hans Super and Pieter R Roelfsema. Chronic multiunit recordings in behaving animals: advantages and limitations. *Progress in brain research*, 147:263–282, 2005.

[64] Pouya Bashivan, Kohitij Kar, and James J DiCarlo. Neural population control via deep image synthesis. *Science*, 364(6439), 2019.

[65] Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fmri. *NeuroImage*, 56(2):400–410, 2011.

[66] Max Welling. Kernel ridge regression. *Max Welling's classnotes in machine learning*, pages 1–3, 2013.

[67] Marcello Ienca, Pim Haselager, and Ezekiel J Emanuel. Brain leaks and consumer neurotechnology. *Nature biotechnology*, 36(9):805–810, 2018.