Hyperrealistic neural decoding: Linear reconstruction of face stimuli from fMRI measurements via the GAN latent space

Thirza Dado, Yağmur Güçlütürk,
Luca Ambrogioni, Gabriëlle Ras, Sander E. Bosch,
Marcel van Gerven, Umut Güü̧lü
Radboud University, Donders Institute for Brain, Cognition and Behaviour
Nijmegen, Netherlands
t.dado@student.ru.nl

Abstract

We introduce a new framework for hyperrealistic reconstruction of perceived naturalistic stimuli from brain recordings. To this end, we embrace the use of generative adversarial networks (GANs) at the earliest step of our neural decoding pipeline by acquiring functional magnetic resonance imaging data as subjects perceived face images created by the generator network of a GAN. Subsequently, we used a linear decoding approach to predict the latent state of the GAN from brain data. Hence, latent representations that are needed for stimulus (re-)generation are obtained, leading to ground-breaking image reconstructions. Altogether, we have developed a highly promising approach for decoding neural representations of real-world data, which may pave the way for systematically analyzing neural information processing in the functional brain.

Disclaimer: This manuscript contains no real face images; all faces are artificially generated by a generative adversarial network.

1 Introduction

In recent years, the field of neural decoding has been gaining more and more traction as advanced computational methods became increasingly available for application on neural data. This is a very welcome development in both neuroscience and neurotechnology since reading neural information will not only help understand and explain human brain function, but also find applications in brain computer interfaces and neuroprosthetics to help people with disabilities.

Neural decoding can be conceptualized as the inverse problem of mapping brain responses back to sensory stimuli via a latent space [18]. Such a mapping can be idealized as a composite function of linear and nonlinear transformations. The linear transformation models the mapping from brain responses to the latent space. The latent space should effectively capture the defining properties of the underlying neural representations. The nonlinear transformation models the mapping from the latent space to sensory stimuli.

The systematic correspondences between latent representations of discriminative convnets and neural representations of sensory cortices is well established [20, 12, 2, 6, 7, 5]. As such, exploiting these systematic correspondences by adapting discriminative convnets for generative modeling has pushed the state-of-the-art in neural decoding forward [9, 8, 15]. Yet, there is still much room for improvement since state-of-the-art results still fall short of providing photorealistic reconstructions.

Preprint. Under review.

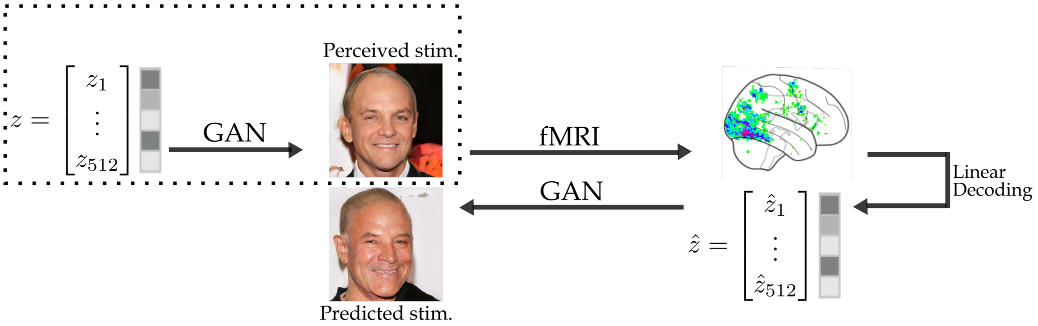
Perceived stim.
$z = \begin{bmatrix} z_1 \\ \vdots \\ z_{512} \end{bmatrix}$
GAN
fMRI
GAN
Linear Decoding
$\hat{z} = \begin{bmatrix} \hat{z}_1 \\ \vdots \\ \hat{z}_{512} \end{bmatrix}$
Predicted stim.

Figure 1: Schematic illustration of the HYPER framework. Face images are generated from randomly sampled latent features $z \in Z$ by a face-generating GAN, as denoted by the dotted box. These faces are then presented as visual stimuli during brain scanning. Next, a linear decoding model learns the mapping from brain responses to the original latent representation, after which it predicts latent features $\hat{z}$ for unseen brain responses. Ultimately, these predicted latent features are fed to the GAN for image reconstruction.

At the same time, generative adversarial networks have emerged as perhaps the most powerful generative models to date [4, 10, 11, 1] that can potentially bring neural decoding to the next level. However, since the *true* latent representations of GANs are not readily available for preexisting neural data (unlike those of the aforementioned discriminative convnets), the adoption of GANs in neural decoding has been relatively slow (see [14] for an earlier attempt with GANs and [19] for a related attempt with VAEs) and still falls short of providing high-fidelity reconstruction of naturalistic stimuli.

In this study, we introduce a very powerful yet simple framework for HYperrealistic reconstruction of PERception (HYPER), which elegantly integrates GANs in neural decoding by combining the following components (Figure 1).

First, we used a pretrained generative adversarial network, which allows for the generation of meaningful data samples from randomly sampled latent vectors. This model is used both for generating the stimulus set and for the ultimate reconstruction of perceived stimuli. In the current study, we used the progressive growing of GANs (PGGAN) model [10], which generates photorealistic faces that resemble celebrities.

Second, we made use of neural data with a known latent representation, obtained by presenting the stimulus set produced using the above-mentioned generative model, and recording the brain responses to these stimuli. In the current study, we collected fMRI recordings in response to the images produced using the PGGAN. We created a dataset consisting of a separate training and test set.

Third, we used a linear model, mapping the neural data to the latent space of the generative model. In the current study, we trained a ridge regression model using the data from the training set. Using this model, we then obtained latent vectors for the neural responses corresponding to the stimulus images in the test set. Feeding these latent vectors back into the generative model resulted in the hyperrealistic reconstructions of perception.

2 Related work

Deep neural networks (DNNs) have previously been used for neural decoding of visual experience [18]. Examples include linear reconstruction of perceived handwritten characters [13], decoding of perceived and imagined object categories [9], and reconstruction of perceived and imagined natural images [15, 14]. In the light of face perception, the most closely related work is probably that of Güçlütürk et al. [8] and VanRullen & Reddy [19]. HYPER distinguishes itself from these approaches by using artificially generated face stimuli of which the latent representations needed for (re-)generation are known, instead of estimating latent features from the images by dimensionality reducing techniques.

Güçlütürk et al. [8] have used a deep convolutional neural network in combination with principal component analysis (PCA) to perform the nonlinear encoding transformation from face images from the CelebA dataset (224 × 224 pixels) into 699-dimensional latent features. Next, latent representations are predicted from brain activations by maximum a priori (MAP) estimation. Ultimately, face reconstructions (64 × 64 pixels) are obtained by nonlinear decoding of the predicted latent features using a the pre-trained generator network of a GAN that takes the adversarial loss, feature loss, and stimulus loss into account.

VanRullen & Reddy [19] trained a variational autoencoder-generative adversarial network (VAE-GAN) on face images from CelebA. The encoder network encodes face images (128 × 128 pixels) into 1024-dimensional latent representations. Next, the linear relationship between latent features and brain responses is determined by linear regression to predict the latent features from unseen brain responses. And eventually, face images are reconstructed by feeding these latent vectors to the VAE-GAN's decoder network.

3 Methods

3.1 Training on synthetic images with known latent features

State-of-the art face reconstruction techniques use deep neural networks to encode vectors of latent features for the images presented during the fMRI experiment [8, 19]. These feature vectors have been shown to have a linear relation with measured brain responses that can be captured using a multivariate linear regression. However, this approach entails information loss since the target images need to be reconstructed from the linear prediction using an approximate inversion network such as a variational decoder. This reconstruction is imperfect even when the ground truth latent is known, leading to a severe bottleneck to the maximum possible reconstruction quality.

In this paper, we avoid this sub-optimality by presenting to the participants photorealistic synthetic images generated using a PGGAN. This allows us to store the ground-truth latents corresponding to the generated images which can be perfectly reconstructed using the generative model.

3.2 Linear neural decoding

The linear relationship between voxel responses and latent features is estimated using ridge regression. For each latent feature z we minimize a loss term of the form:

$$\mathcal{L} = \sum_{i=1}^{M} \left(z_i - \mathbf{w}^T \mathbf{x}_i\right)^2 + \lambda ||\mathbf{w}||_2^2 \tag{1}$$

where $\mathbf{x}_i$ and z_i represent the measured voxel responses and latent feature value for the i-th observation, respectively, $\mathbf{w}$ are the model coefficients, and λ the regularization parameter. As we are dealing with with multi-target output, ridge regression returns a combination of 512 single-output models (i.e. one model per latent value). Generalized (leave-one-out) cross-validation is used to select the most appropriate regularization parameter ($\lambda = 100, 500, 1000, 2000$, or 5000).

3.3 Datasets

3.3.1 Visual stimuli

High-resolution face images (1024×1024 pixels) are generated by the generator network of a Progressive GAN (PGGAN) model [10] from randomly sampled latent vectors. Each generated face image is cropped and resized to 224×224 pixels. In total, 1050 unique faces are presented once for the training set, and 36 faces are repeated 14 times for the testing set. In this way, it is ensured that the training set covers a large stimulus space to fit a general face model, whereas the voxel responses from the testing set contain less noise and higher statistical power.

3.3.2 Brain responses

An fMRI dataset was collected, consisting of BOLD responses that correspond to the perceived face stimuli. The BOLD responses (TR = 1.5 s, voxel size = $2 \times 2 \times 2$ mm^3, whole-brain coverage) of two healthy subjects were measured (S1: 30-year old male; S2: 32-year old male) while they were fixating on a target (0.6×0.6 degrees) [17] superimposed on the stimuli (15×15 degrees) to minimize involuntary eye movements.

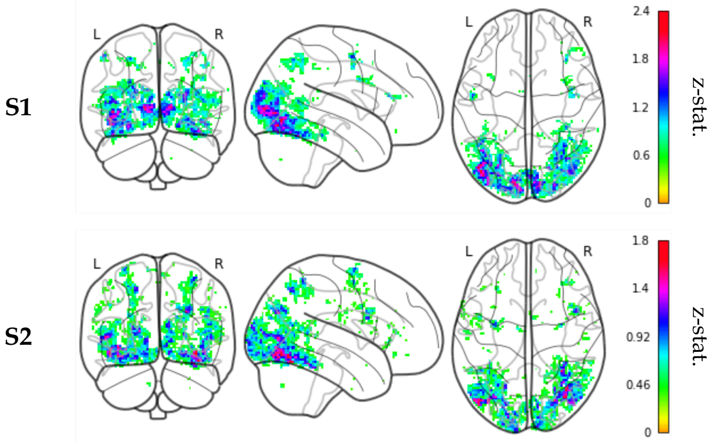

Figure 2: Functional ROI: 4096 most active voxels based on highest z-statistics within the averaged z-map from the training set responses, resulting in a distributed network of activity. Top: fROI from Subject 1. Bottom: fROI from Subject 2.

During preprocessing, the obtained brain volumes are realigned to the first functional scan and the mean functional scan, respectively, after which the volumes are normalized to MNI space. A general linear model is fit to deconvolve task-related neural activation with the canonical hemodynamic response function (HRF). Next, for each voxel, we computed its t-statistic (the probability of true activation corresponds to a high t-statistic), and converted these t-scores to z-statistics to obtain a brain map in terms of z per perceived stimulus. Ultimately, the average z-map is taken from all training set responses to cancel out random activations. From this average brain map, the most active 4096 voxels are selected as a functional region of interest (fROI) for the remainder of this study (Figure 2). After all, training images are only presented once such that these individual recordings contain a high signal-to-noise ratio. As expected, most of these fROI voxels are located in the visual ventral stream. Voxel responses from the testing set are not used to create this mask to avoid double-dipping.

The experiment was approved by the local ethics committee (CMO Regio Arnhem-Nijmegen). Subjects provided written informed consent in accordance with the Declaration of Helsinki. The fMRI dataset is available from the first author on request and the code is linked in the Supplementary Materials.

3.4 Evaluation

3.4.1 Performance metrics

Performance of the linear decoding model on the test set of the fMRI dataset is assessed by four metrics: (i) the average Euclidean distance and (ii) average Pearson correlation coefficient between each of the 512 feature dimensions of true and predicted latent vectors, and (iii) the average feature similarity and (iv) average Pearson correlation coefficient between stimuli and their reconstructions. Specifically, feature similarity is defined as the Euclidean similarity between feature extraction layer outputs ($n = 2048$) of the ResNet50 model, pretrained for face recognition.

3.4.2 Feature scores

Based on the assumption that there exists a hyperplane in latent space for binary semantic attributes (e.g. male vs. female), [16] have identified the decision boundaries for five semantic face attributes in PGGAN's latent space: gender, age, the presence of eyeglasses, smile, and pose, by training five independent linear support vector machines (SVMs). We used these five decision boundaries to compute feature scores by taking the dot product between latent representation and decision boundary. In this way, model performance with regard to specific visual features can be assessed.

3.5 Implementation details

fMRI preprocessing is implemented in SPM12 after which first-order analysis is carried out in Python's Nipy environment. NVIDIA's PGGAN TensorFlow source code is used in combination with CUDA V10.0.130, CuDNN, and Anaconda3 (Python 3.6). Keras' pretrained implementation of VGGFace (ResNet50 model) is used to evaluate similarities between feature maps of the perceived and reconstructed images. Linear decoding is implemented using ScikitLearn.

4 Results

Linear decoding of fMRI recordings using PGGAN's latent space has led to unprecedented stimulus reconstructions. Figure 3 presents all of the image reconstructions together with the originally perceived stimuli.

To keep the presentation concise, the first half of the images (1-18) are reconstructed from brain activations from Subject 1 and the second half (19-36) from Subject 2. The complete collection of reconstructions can be found in the supplement, including interpolations and feature scores. The interpolations visualize the distance between predicted and true latent representations that underlie the (re)generated faces. It demonstrates which features are being retained or change. The bar graphs next to the perceived and reconstructed images show the scores of each image in terms of five semantic face attributes in PGGAN's latent space: gender, age, the presence of eyeglasses, smile, and pose. Looking at the similarities and differences in the graphs for perceived and reconstructed images is a way to evaluate how well each semantic attribute is captured by our model. For most reconstructions, the two graphs match in terms of directionality. There are a few cases, however, demonstrating that there is still room for improvement, e.g. number 31, 33 and 35.

We further assessed the model performance of HYPER with respect to the consistency of these five semantic features by correlating the values for the reconstructions and perceived stimuli. We found high correlations for gender, pose, and age, but no significant correlation for the smile attribute (Figure 4).

Next, we compared the performance of the HYPER framework to the state-of-the-art VAE-GAN approach [19] and the eigenface approach [3]. We compared the Euclidean distance and Pearson correlation coefficient between the predictions and ground truth images. All quantitative and qualitative comparisons showed that the HYPER framework outperformed the baselines and had significantly above-chance latent and feature reconstruction performance (p « 0.01, permutation test) (Table 1).

Table 1: Reconstruction performance of our method compared to the state-of-the-art VAE-GAN [19] and the eigenface approach [3] is assessed in terms of the Euclidean distance and Pearson correlation coefficient between prediction and ground truth. The first two columns involve true and predicted latent vectors whereas the last two columns concern the features of stimuli and their reconstructions. The table displays mean values ± standard errors. Given that all three methods require different image resolutions, all images are resized to 224 × 224 pixels for a fair comparison. In addition, statistical significance of our method is evaluated by permutation tests.

		Eucl. dist. latents	**Corr. coef. latents**	**Eucl. dist. feats.**	**Corr. coef. feats.**
S1	HYPER	1.1615 ± 0.0525	0.0283 ± 0.0284	5.5285 ± 0.1796	0.2326 ± 0.0234
		$(p < 0.001; perm.test)$	$(p < 0.001; perm.test)$	$(p < 0.001; perm.test)$	$(p < 0.001; perm.test)$
	VAE-GAN	-	-	6.4644 ± 0.1083	0.0951 ± 0.0140
	Eigenface	-	-	6.6013 ± 0.0868	-0.0024 ± 0.0086
S2	HYPER	1.1934 ± 0.0556	0.0261 ± 0.0287	5.5669 ± 0.1748	0.1718 ± 0.0226
		$(p < 0.001; perm.test)$	$(p < 0.001; perm.test)$	$(p < 0.001; perm.test)$	$(p < 0.001; perm.test)$
	VAE-GAN	-	-	6.6233 ± 0.1103	0.0928 ± 0.0124
	Eigenface	-	-	7.2894 ± 0.1180	0.0108 ± 0.0102

We also present arbitrarily chosen but representative reconstruction examples from VAE-GAN and eigenface approaches, again demonstrating that the HYPER framework resulted in markedly better reconstructions (Figure 5). It is important to note that the reconstructions by the VAE-GAN approach appear to be of lower quality than those presented in the original study. One likely explanation for this result could be that the number of training images in our dataset was not sufficient to effectively train their model. The training set used in [19] consisted of ∼8000 face images, whereas ours consist of

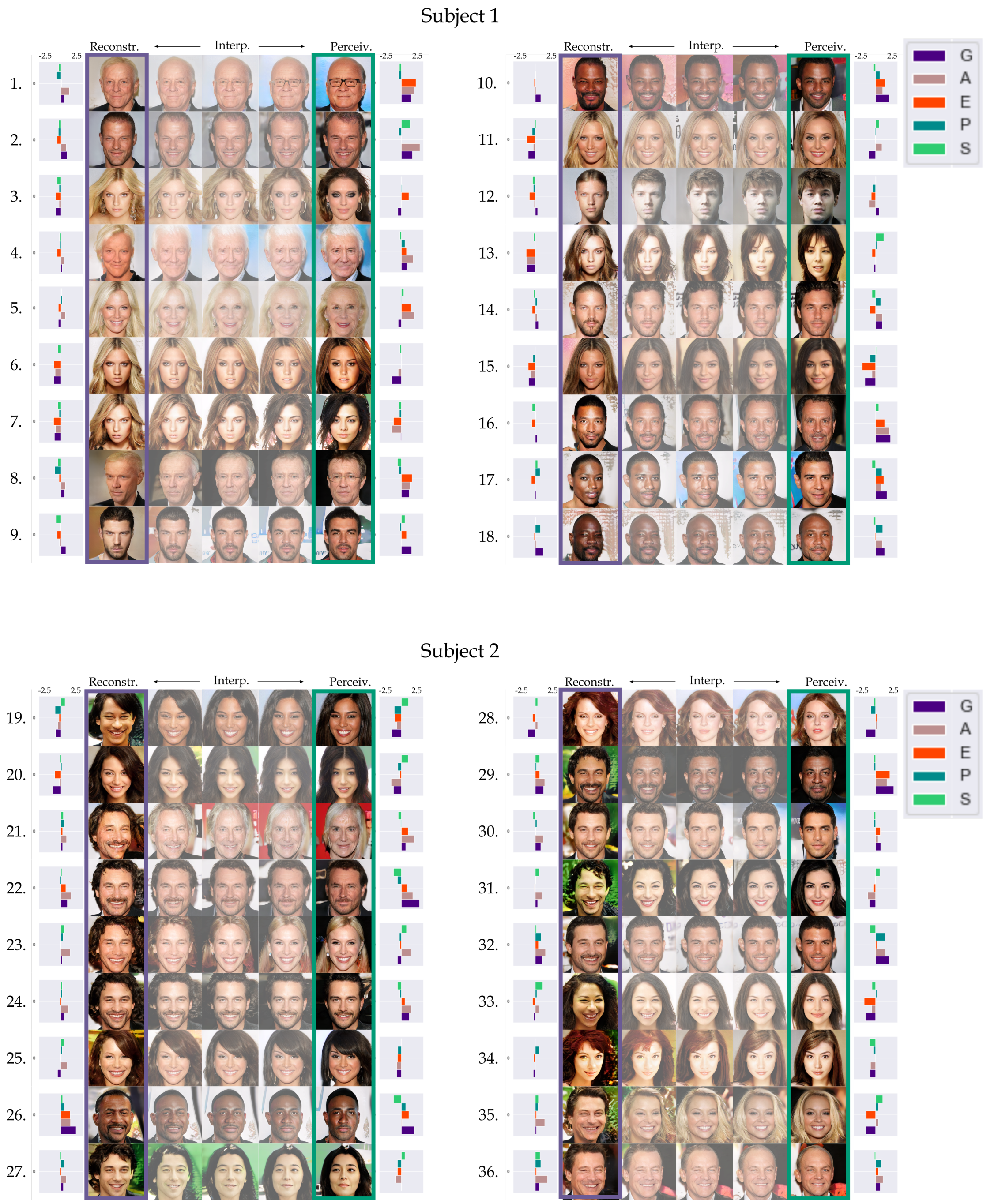

Figure 3: Results of testing set samples 1-18 for Subject 1 and 19-36 for Subject 2. Image reconstructions (left) versus perceived images (right), where interpolations visualize similarity regarding the underlying latent representations. Next to each reconstruction and perceived stimulus, a rotated bar graph displays the corresponding feature scores for gender, age, eyeglasses, pose, and smile.

1050 face images only. For a fair comparison, we used the present dataset to evaluate all the methods presented in this study.

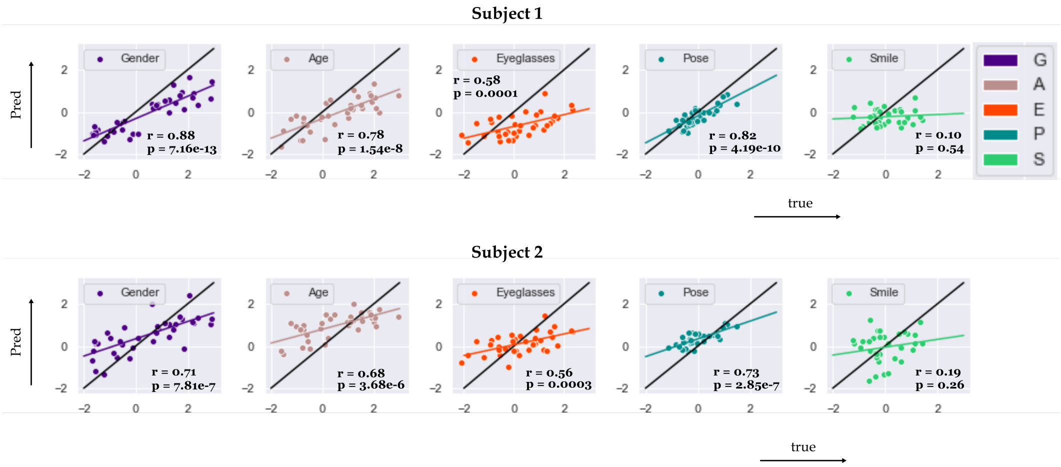

Figure 4: Reconstruction performance on five features. The X axis denotes the true scores with respect to the perceived stimuli whereas the Y axis represents the predicted scores with respect to the reconstructions. Additionally, the Pearson correlation coefficient (r) and corresponding p-value (p) are displayed.

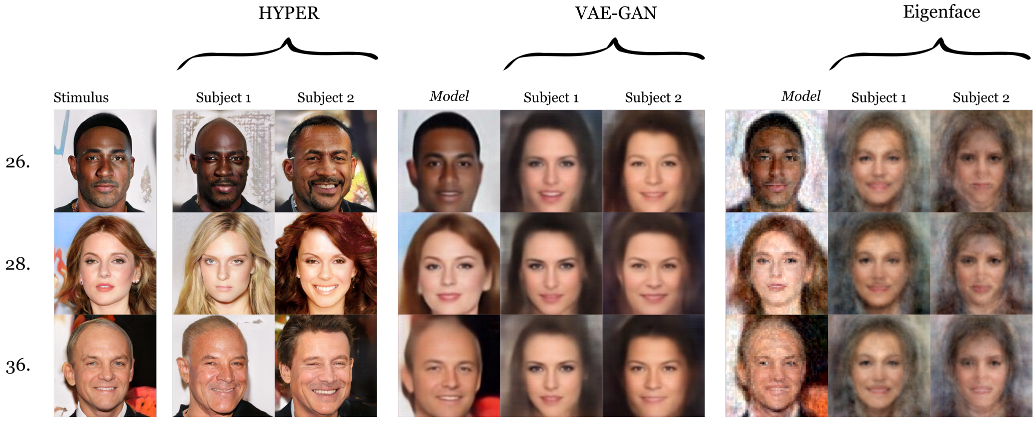

Figure 5: Qualitative results of our approach compared to [19] and the eigenface approach in reconstructing image 26, 28, and 36 (arbitrarily chosen). The *reference* columns display the best possible results. For [19], this displays reconstructions directly decoded from the 1024-dimensional latent representation of this method. For the eigenfaces approach, this shows reconstructions directly obtained from the 512 principal components.

5 Discussion

We have shown how the latent space learned by a generative network can be used for linear decoding of brain activations, thereby serving as a proof-of-concept of using generative modeling to approximate neural manifolds of real-world data. The success of this approach is due to the astonishing performance of PGGAN for generating human faces. At the same time, this also puts (potential) bottlenecks on what can be reconstructed. In this study, PGGAN's generator had to regenerate face images that it had already generated before, guaranteeing its competence. The next step is verifying whether a linear decoding model trained on brain responses with regard to generated face images generalizes to brain responses to real faces. The true latent representations of real images are not accessible, but would no longer be required if the decoding model has learned to accurately predict them from the artificial data samples. This would result in a great leap forward within the field of neural coding.

However, even though the artificial face images look incredibly realistic, this does not guarantee the ability of the generator network to reconstruct any face perceived in real life. The current reconstructions are already observed to contain biases. That is, the model predicts primarily latent representations corresponding to young, western-looking faces without eyeglasses, because predictions tend to follow the image statistics of the training set, containing these feature imbalances. Also PGGAN's generator network is known to suffer from this problem - referred to as *feature entanglement* - as manipulating one particular feature in latent space affects other features as well [16]. For example, editing a latent vector to make the generated face wear eyeglasses simultaneously makes the face older because of such biases in the training data. Feature entanglement obstructs the generator to map unfamiliar latent elements to their respective visual features. It is easy to foresee the complications for reconstructing real face images.

A modified version of PGGAN, called StyleGAN [11], overcomes the feature entanglement problem. StyleGAN maps the entangled latent vector to an additional intermediate latent space - thereby reducing feature entanglement - which is then integrated into the generator network using adaptive instance normalization. This results in superior control over the semantic attributes in the reconstructed images and possibly the generator's competence to reconstruct unfamiliar features. Replacing the PGGAN with StyleGAN could therefore be a logical next step for studies concerned with the neural decoding of faces.

Finally, neural decoding can reveal what subject-specific information is (not) present in the observed brain activations. That is, even though participants are presented with identical stimuli, sensory information is likely to be integrated with subjective expectations and beliefs, causing subjective variations in reconstructions. This may include enhanced, diminished, missing, imagined, or transformed information. Eventually, the HYPER framework might allow us to bridge the gap between objective and subjective experience.

6 Conclusion

We have presented a framework for HYperrealistic reconstruction of PERception (HYPER) by linear neural decoding of brain responses via the GAN latent space, leading to unparalleled state-of-the-art stimulus reconstructions. Considering the speed of progress in the field of generative modeling, we believe that the HYPER framework that we have introduced in this study will likely result in even more impressive reconstructions of perception and possibly even imagery in the near future, ultimately also allowing to better understand mechanisms of human brain function.

Broader Impact

The neural decoding framework presented in this paper is offers access to the subjective contents of the human mind by linear reconstruction of encoded sensory stimuli, possibly bringing our understanding of human brain function forward in the process. Besides the large scientific potential, neural decoding can also enable various applications in the field of neurotechnology (e.g. brain computer interfacing and neuroprosthetics) to help people with disabilities. While the current work focuses on decoding of sensory perception, extensions of our framework to imagery could make it a preferred means for communication for locked-in patients through a brain computer interface. However, care must be taken as “mind reading” technologies also involve serious ethical concerns regarding mental privacy. Although current approaches to neural decoding, such as the one presented in this manuscript, would not allow for involuntary access to thoughts of a person, future developments may allow for extraction of information from the brain more easily, as the field is rapidly developing. As with all scientific and technological developments, ethical principles and guidelines as well as data protection regulations should be followed strictly to ensure the safety of (the data of) potential users of these technologies.

References

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

[2] 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 Comput Biol*, 10(12):e1003963, 2014.

[3] 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.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[5] Umut Güçlü, Jordy Thielen, Michael Hanke, and Marcel Van Gerven. Brains on beats. In *Advances in Neural Information Processing Systems*, pages 2101–2109, 2016.

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

[8] Y. Güçlütürk, U. Güçlü, K. Seeliger, S. Bosch, R. van Lier, and M. A. van Gerven. Reconstructing perceived faces from brain activations with deep adversarial neural decoding. *Advances in Neural Information Processing Systems*, pages 4246–4257, 2017.

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

[10] 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.

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

[12] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Comput Biol*, 10(11), 2014.

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

[14] Katja Seeliger, Umut Güçlü, Luca Ambrogioni, Yagmur Güçlütürk, and Marcel AJ van Gerven. Generative adversarial networks for reconstructing natural images from brain activity. *NeuroImage*, 181:775–785, 2018.

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

[16] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. *arXiv preprint arXiv:1907.10786*, 2019.

[17] Lore Thaler, Alexander C Schütz, Melvyn A Goodale, and Karl R Gegenfurtner. What is the best fixation target? the effect of target shape on stability of fixational eye movements. *Vision Research*, 76:31–42, 2013.

[18] Marcel AJ van Gerven, Katja Seeliger, Umut Güçlü, and Yağmur Güçlütürk. Current advances in neural decoding. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 379–394. Springer, 2019.

[19] Rufin VanRullen and Leila Reddy. Reconstructing faces from fmri patterns using deep generative neural networks. *Communications biology*, 2(1):193, 2019.

[20] 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.