Hyperrealistic Neural Decoding: Reconstruction of Face Stimuli From FMRI Measurements via the GAN Latent Space

Thirza Dado[1], Yagmur Güçlütürk[1], Luca Ambrogioni[1],
Gabriëlle Ras[1], Sander E. Bosch[1], Marcel van Gerven[1] & Umut Güçlü[1]

[1]Radboud University Nijmegen, Donders Institute for Brain, Cognition and Behaviour, The Netherlands

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 decoding approach to predict the latent state of the GAN from brain data. Hence, latent representations for stimulus (re-)generation were obtained, leading to state-of-the-art image reconstructions.

Keywords: deep learning, face generation, functional magnetic resonance imaging, generative adversarial network, neural decoding

Corresponding author: Thirza Dado; E-mail: thirza.dado.1@donders.ru.nl

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 will also find applications in brain computer interfaces and neuroprosthetics to help people with disabilities.

Neural decoding can be conceptualized as the inverse problem where brain responses are mapped back to sensory stimuli via a latent space (Van Gerven, Seeliger, Güçlü, & Güçlütürk, 2019). 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 convolutional networks (convnets) and neural representations of sensory cortices are well established (Yamins et al., 2014; Seyed-Mahdi, Khaligh-Razavi & Kriegeskorte, 2014; Cadieu et al., 2014; Güçlü & Van Gerven, 2015; Güçlü & Van Gerven, 2017; Güçlü, Thielen, Hanke, & Van Gerven, 2016). As such, exploiting these systematic correspondences in neural decoding of visual experience has pushed the state-of-the-art forward (Van Gerven et al, 2019) . This includes linear reconstruction of perceived handwritten characters (Schoenmakers, Barth, Heskes, & Van Gerven, 2013), neural decoding of perceived and imagined object categories (Horikawa & Kamitani, 2017), and reconstruction of natural images (Seeliger, Güçlü, Ambrogioni, Güçlütürk & Van Gerven, 2018; Shen, Horikawa, Majima, & Kamitani, 2019) and faces (Güçlütürk et al., 2017; VanRullen & Reddy, 2019). Yet, there is still room for improvement since these state-of-the-art results still fall short of providing photorealistic reconstructions.

At the same time, generative adversarial networks (GANs) have emerged as perhaps the most powerful generative models to date that can potentially bring neural decoding to the next level (Brock, Donahue, & Simonyan, 2018; Goodfellow et al., 2014; Karras, Aila, Laine, & Lehtinen, 2017; Karras, Laine, & Aila, 2019). A GAN is a is a deep learning architecture for generative modelling, consisting of two competing neural networks, as described by Goodfellow et al. (2014). In short, a generator network is pitted against a discriminator network that learns to distinguish reconstructed "fake" data samples from real data samples. In turn, the generator's goal is to fool the discriminator by generating new and unique, real-looking data samples from randomly sampled low-dimensional latent features. Competition is the drive between both neural networks to improve their methods in tandem until the generated samples are indistinguishable from the real ones. However, since the *true* latent representations of GANs are not readily available for pre-existing neural data (unlike those of the aforementioned discriminative convnets), the adoption of GANs in neural decoding has been relatively slow (see (Seeliger et al., 2018) for an earlier attempt with GANs and (VanRullen & Reddy, 2019) for a related attempt with variational autoencoders-GAN [VAE-GANs]).

In this study, we introduce a very powerful yet simple framework for HYperrealistic reconstruction of PERception (HYPER), which elegantly

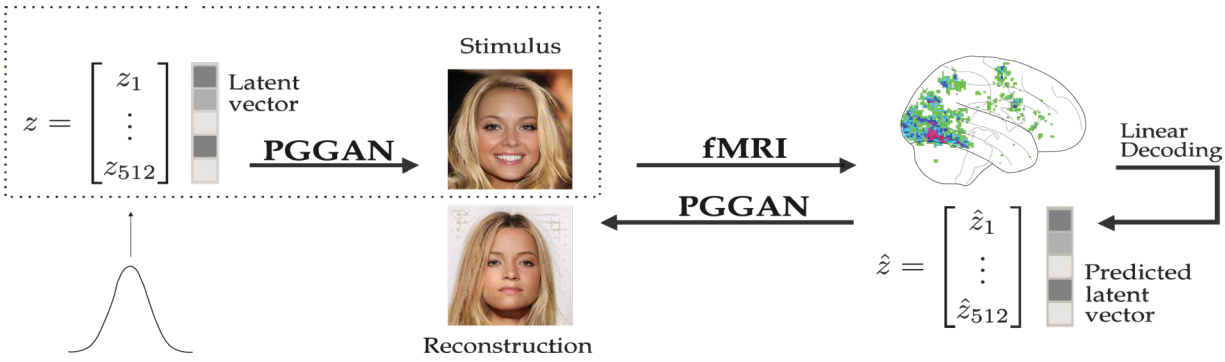

Figure 1. Schematic illustration of the HYPER framework. Face images are generated from randomly sampled latent features $z \in Z$ by a face-generating GAN, as denoted by the dotted box. These faces are then presented as visual stimuli during brain scanning. Next, a linear decoding model learns the mapping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

integrates GANs in neural decoding by combining the following components (Fig. 1):

i GAN. We used a pretrained GAN, 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 (Karras et al., 2017), which generates photorealistic faces that resemble celebrities.

ii Functional magnetic resonance imaging (fMRI). 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 of participants 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.

iii Decoding model. We used a decoding model, mapping the neural data to the latent space of the generative model. 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.

Method

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 (Güçlütürk et al., 2017; VanRullen & Reddy, 2019). These feature vectors have been shown to have a linear relation with measured brain responses. 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, leading to a severe bottleneck to the maximum possible reconstruction quality.

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

Neural Decoding

Progressive GAN. To achieve the generation of high-resolution images, a training procedure was developed that grows the generator and discriminator network in a progressive fashion (Karras et al., 2017). More specifically, training on face images from the CelebA-HQ dataset started at a low resolution of 4×4 pixels and layers were added incrementally. To avoid shocks to the well-trained lower-resolution layers, these additional layers were "faded in" smoothly by linear interpolation of the weights from 0 to 1. In the end, a mapping was established from 512-dimensional latent features to hyper-realistic face images with a final resolution of 1024×1024 pixels. At this point, both the generator and discriminator network consisted of nine phases and 23.1M trainable parameters.

Predicting latent vectors from brain data. We adapted the deep generative network of PGGAN by adding a dense layer at the beginning to transform brain data into latent vectors. This layer was trained by minimizing the Euclidean distance between true and predicted latent representations ($batchsize = 30$, $lr = 0.00001$, Adam optimization) with weight decay ($alpha = 0.01$) to reduce complexity and multicollinearity of the model. The remainder of the generative network was kept fixed.

Datasets

Visual Stimuli. High-resolution face images (1024×1024 pixels) were generated by the generator network of a Progressive GAN (PGGAN) model (Karras et al., 2017) from randomly sampled latent vectors. Each generated face image was cropped and resized to 224×224 pixels. In total, 1050 unique faces were presented once for the training set, and 36 faces were repeated 14 times for the test set of which the average brain response was taken. This ensured that the training set covered a large stimulus space to fit a general face model, whereas the voxel responses from the test set contained less noise and higher statistical power.

Brain responses. fMRI data was collected, consisting of blood oxygen level dependent (BOLD) responses that corresponded to the perceived face stimuli. The BOLD responses (TR = 1.5 s, voxel size = 2×2×2 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) (Thaler, Schütz, Goodale, & Gegenfurtner, 2013) superimposed on the stimuli (15×15 degrees) to minimize involuntary eye movements.

During preprocessing, the obtained brain volumes were realigned to the first functional scan and the mean functional scan, respectively, after which the volumes were normalized to MNI space. A general linear model was fit to deconvolve task-related neural activation with the canonical hemodynamic response function (HRF). Next, for each voxel, we computed its t-statistic and converted these t-scores to z-statistics to obtain a brain map in terms of z per perceived stimulus. Ultimately, most-active 4096 voxels were selected from the training set to define a voxel mask (Fig. 2). Most of these mask voxels are located in the downstream brain regions. Voxel responses from the test set are not used to create the voxel 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 for both subjects and used models are openly accessible.

Evaluation

Model performance is assessed in terms of three metrics: latent similarity, feature similarity, and structural similarity. First, latent similarity is the Euclidean similarity between predicted and true latent vectors. Second, feature similarity is the Euclidean similarity between feature extraction layer outputs (n=2048) of the ResNet50 model, pretrained for face recognition, which we feed stimuli and reconstructions. Lastly, structural similarity is used to measure the spatial interdependence between pixels of stimuli and reconstructions (Wang, Bovik, Sheikh, & Simoncelli, 2004).

Next, based on the assumption that there exists a hyperplane in latent space for binary semantic attributes (e.g. male vs. female), Shen, Gu, Tang and Zhou (2019) 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 decision boundaries to compute feature scores per image, by taking the dot product between latent representation and decision boundary, resulting in a scalar. In this way, model performance with regard to specific visual features could be captured along a continuous spectrum and could be compared across images.

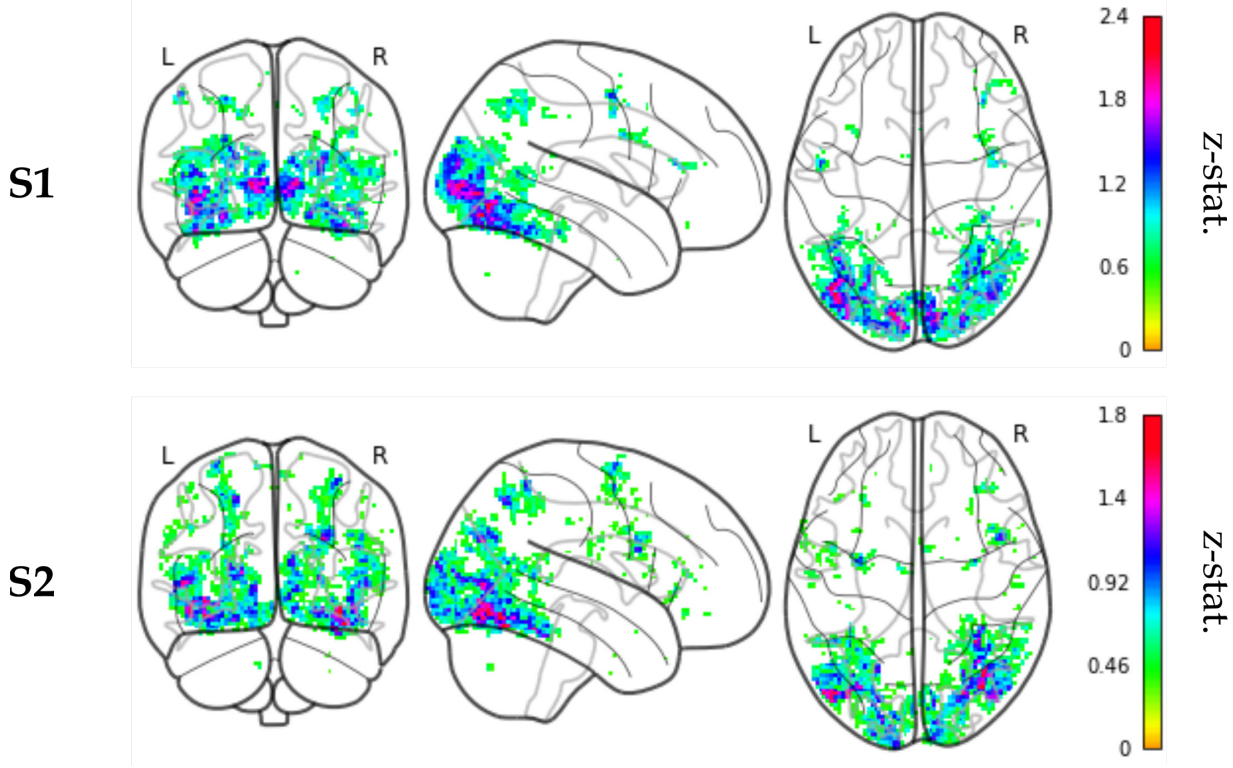

Figure 2. Voxel mask: 4096 most active voxels based on highest z-statistics within the averaged z-map from the training set responses, resulting in a distributed network of activity.

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.

Results

Linear decoding of fMRI recordings using PGGAN's latent space has led to unprecedented stimuli reconstructions. Figure 3 presents all 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 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. A few cases, however, demonstrate that there is still room for improvement (e.g. number 31, 34, and 35). Correlating the feature scores for stimuli and reconstructions resulted in significant ($p < 0.05$; Student's t-test) results for gender, age, eyeglasses, and pose, but not for smile (Fig. 4). We would like to point out that using feature scores quantifies model performance as continuous rather than binary, explaining the significant correlation for eyeglasses despite lack of reconstruction in number 1 and 8.

Next, we compared the performance of the HYPER framework to the state-of-the-art VAE-GAN approach (VanRullen & Reddy, 2019) and the traditional eigenface approach (Cowen, Chun, & Kuhl, 2014) which maps the brain recordings onto different latent spaces. For a fair comparison, we used the same voxel mask to evaluate all the methods presented in this study without any optimization to a particular decoding approach. The VAE-GAN approach predicts 1024-dimensional latent representations which are fed to the VAE's decoder network for stimulus reconstruction (128×128 pixels). The eigenface approach predicts the first 512 principal components (or ' eigenfaces') after which stimulus reconstruction (64×64 pixels) is achieved by applying an inverse principal component analysis (PCA) transform. All quantitative and qualitative comparisons showed that the HYPER framework outperformed the baselines and had significantly above-chance latent and feature reconstruction performance ($p < 0.001$, permutation test), indicating the probability that a random latent vector or image would be more similar to the original stimulus (Table 1).

We also present arbitrarily chosen but representative reconstruction examples from the VAE-GAN and eigenface approach, again demonstrating that the HYPER framework resulted in markedly better reconstructions (Fig. 5).

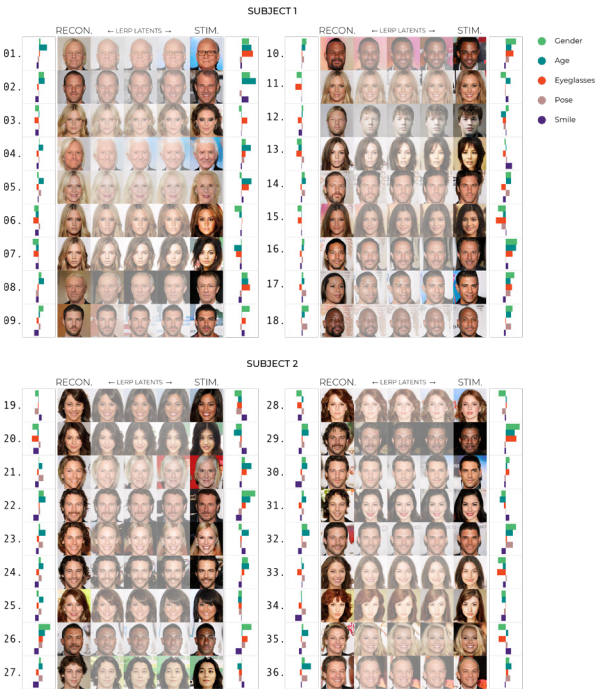

Figure 3. Results of model 0 that is trained on only the latent vectors. Here, we display the testing set samples 1-18 for Subject 1 and 19-36 for Subject 2. Image reconstructions **(left)** versus perceived images **(right)**. Interpolations visualize similarity regarding the underlying latent representations. Next to each reconstruction and perceived stimulus, a rotated bar graph displays the corresponding feature scores for gender (g), age (a), eyeglasses (e), pose (p), and smile (s).

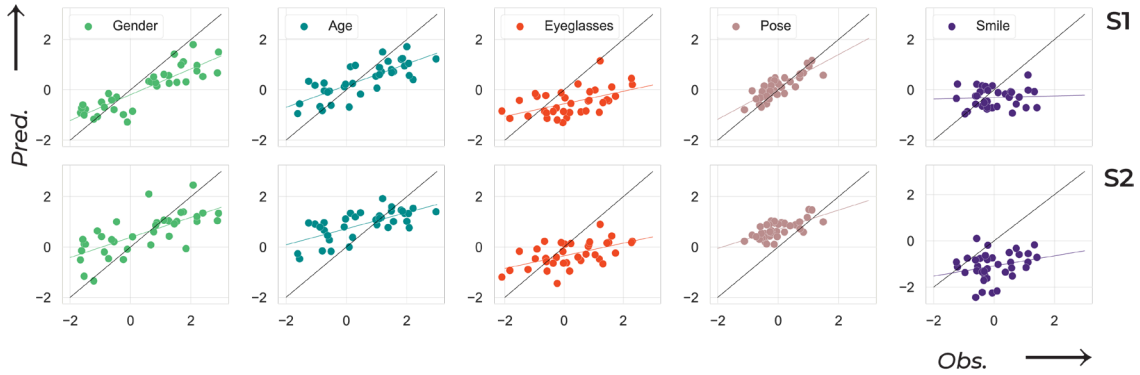

Figure 4. Reconstruction performance on five features. The x-axis denotes the true scores with respect to the perceived stimuli whereas the y-axis represents the predicted scores with respect to the reconstructions. Additionally, the Pearson correlation coefficient (r) and corresponding p-value (p) are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

Discussion

We have decoded brain recordings during perception of face photographs using the presented HYPER method, leading to state-of-the-art stimulus reconstructions. Consequently, this work serves as a proof-of-concept of using generative modelling to approximate neural manifolds of real-world data, possibly bringing our understanding of human brain function forward in the process. The success of this approach is due to the astonishing performance of PGGAN. At the same time, PGGAN puts (potential) bottlenecks on what can be reconstructed: the generator network 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.

Next, the HYPER framework resulted in considerably better reconstructions than the two benchmark approaches. 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. A likely explanation for this result could be that the number of training images in our dataset was not sufficient to effectively train their model (8000 vs 1050) and the different voxel selection procedure.

Importantly, image reconstructions by HYPER appear to contain biases. That is, the model predicts

Table 1. Model performance of the HYPER framework compared to the state-of-the-art VAE-GAN (VanRullen & Reddy, 2019) and the eigenface approach (Cowen et al., 2014) is assessed in terms of the feature similarity (column 2) and structural similarity (column 3) between stimuli and reconstructions (mean ± std error). The first column displays latent similarity which is only applicable to the HYPER method because the true and predicted latent vectors are known. Because of resolution differences, all images were resized to 224 × 224 pixels and smoothed with a Gaussian filter (kernel size = 3) for a fair comparison. Also, the backgrounds of the images were removed. In addition, statistical significance of the HYPER method was evaluated against randomly generated latent vectors and their reconstructions.

		Lat. Sim.	Feat. Sim	Struct. Sim
S1	HYPER	0.4521 ± 0.0026 ($p < 0.001$; perm. test)	0.1745 ± 0.0038 ($p < 0.001$; perm. test)	0.6663 ± 0.0115 ($p < 0.001$; perm. test)
	VAE-GAN	-	0.1416 ± 0.0025	0.5598 ± 0.0151
	Eigenface	-	0.1319 ± 0.0016	0.5877 ± 0.0115
S2	HYPER	0.4447 ± 0.0020 ($p < 0.001$; perm. test)	0.1715 ± 0.0049 ($p < 0.001$; perm. test)	0.6035 ± 0.0128 ($p < 0.001$; perm. test)
	VAE-GAN	-	0.1461 ± 0.0022	0.5832 ± 0.0141
	Eigenface	-	0.1261 ± 0.0019	0.5616 ± 0.0097

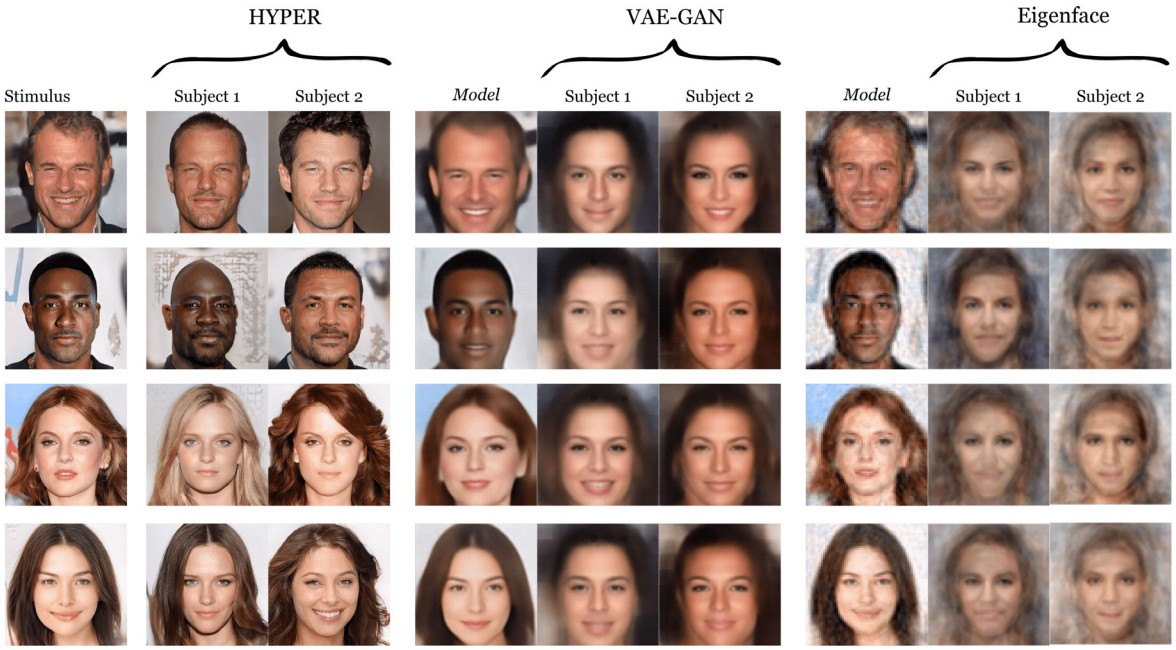

Figure 5. Qualitative results of our approach compared to VanRullen and Reddy (2019) (the VAE-GAN approach) and the eigenface approach in reconstructing image 26, 28, and 36 (arbitrarily chosen). The model columns display the best possible results. For VanRullen and Reddy (2019), this displays reconstructions 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.

primarily latent representations corresponding to young, western-looking faces without eyeglasses because predictions tend to follow the image statistics of the (celebrity) training set. PGGAN's generator network is also known to suffer from this problem – referred to as "feature entanglement" – where manipulating one particular feature in latent space affects other features as well (Shen et al., 2018). For example, editing a latent vector to make the generated face wear eyeglasses simultaneously makes the face look 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 images of real faces.

A modified version of PGGAN, called StyleGAN (Karras et al., 2019; Karras et al., 2020), is designed to overcome 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. Compared to PGGAN, the generated face photographs by StyleGAN have improved considerably in quality and variation, of which the latter is likely to alleviate current biases. Replacing the PGGAN with StyleGAN would therefore be a logical next step for studies concerned with the neural decoding of faces.

Furthermore, neural decoding can reveal what 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. 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.

Finally, besides the large scientific potential, this research could also have societal impacts when enabling 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.

Conclusion

We have presented a framework for HYperrealistic reconstruction of PERception (HYPER) by 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 modelling, 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 allowing for better understanding the mechanisms of human brain function.

References

Brock, A., Donahue, J., & Simonyan, K. (2018). Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.

Cadieu, C. F., Hong, H., Yamins, D. L., Pinto, N., Ardila, D., Solomon, E. A., ... DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS Comput Biol*, *10*(12), e1003963.

Cowen, A. S., Chun, M. M., & Kuhl, B. A. (2014). Neural portraits of perception: reconstructing face images from evoked brain activity. *Neuroimage*, *94*, 12–22.

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., ... others (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, *536*(7615), 171–178.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672–2680).

Güçlü, U., Thielen, J., Hanke, M., & Van Gerven, M. (2016). Brains on beats. In *Advances in neural information processing systems* (pp. 2101–2109).

Güçlü, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014.

Güçlü, U., & van Gerven, M. A. (2017). Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. *NeuroImage*, *145*, 329–336.

Güçlütürk, Y., Güçlü, U., Seeliger, K., Bosch, S., van Lier, R., & van Gerven, M. A. (2017). Reconstructing perceived faces from brain activations with deep adversarial neural decoding. *Advances in Neural Information Processing Systems*, 4246-4257.

Horikawa, T., & Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, *8*(1), 1–15.

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

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4401–4410).

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of stylegan. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 8110–8119).

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Comput Biol*, *10*(11).

Schoenmakers, S., Barth, M., Heskes, T., & Van Gerven, M. (2013). Linear reconstruction of perceived images from human brain activity. *NeuroImage*, *83*, 951–961.

Seeliger, K., Güçlü, U., Ambrogioni, L., Güçlütürk, Y., & van Gerven, M. A. (2018). Generative adversarial networks for reconstructing natural images from brain activity. *NeuroImage*, *181*, 775–785.

Shen, G., Horikawa, T., Majima, K., & Kamitani, Y. (2019). Deep image reconstruction from human brain activity. *PLoS Comput Biol*, *15*(1), e1006633.

Shen, Y., Gu, J., Tang, X., & Zhou, B. (2019). Interpreting the latent space of gans for semantic face editing. *arXiv preprint arXiv:1907.10786*.

Thaler, L., Schütz, A. C., Goodale, M. A., & Gegenfurtner, K. R. (2013). What is the best fixation target? the effect of target shape on stability of fixational eye movements. *Vision Research*, *76*, 31–42.

van Gerven, M. A., Seeliger, K., Güçlü, U., & Güçlütürk, Y. (2019). Current advances in neural decoding. In *Explainable ai: Interpreting, explaining and visualizing deep learning* (pp. 379–394). Springer.

VanRullen, R., & Reddy, L. (2019). Reconstructing faces from fmri patterns using deep generative neural networks. *Communications biology*, *2*(1), 193.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, *13*(4), 600–612.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.