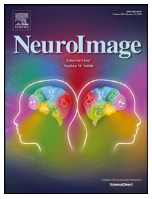
NeuroImage 247 (2022) 118812

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage



NeuroGen: Activation optimized image synthesis for discovery neuroscience

Zijin Gu[a], Keith Wakefield Jamison[b], Meenakshi Khosla[a], Emily J. Allen[c,d], Yihan Wu[c], Ghislain St-Yves[c,d], Thomas Naselaris[c,d], Kendrick Kay[c], Mert R. Sabuncu[a], Amy Kuceyeski[b,*]

[a] School of Electrical and Computer Engineering, Cornell University, Ithaca, New York, USA
[b] Department of Radiology, Weill Cornell Medicine, New York, New York, USA
[c] Center for Magnetic Resonance Research(CMRR), Department of Radiology, University of Minnesota, Minneapolis, Minnesota, USA
[d] Department of Neuroscience, University of Minnesota, Minneapolis, Minnesota, USA

ARTICLE INFO

Keywords:
Function MRI
Neural encoding
Image synthesis
Deep learning

ABSTRACT

Functional MRI (fMRI) is a powerful technique that has allowed us to characterize visual cortex responses to stimuli, yet such experiments are by nature constructed based on a priori hypotheses, limited to the set of images presented to the individual while they are in the scanner, are subject to noise in the observed brain responses, and may vary widely across individuals. In this work, we propose a novel computational strategy, which we call NeuroGen, to overcome these limitations and develop a powerful tool for human vision neuroscience discovery. NeuroGen combines an fMRI-trained neural encoding model of human vision with a deep generative network to synthesize images predicted to achieve a target pattern of macro-scale brain activation. We demonstrate that the reduction of noise that the encoding model provides, coupled with the generative network's ability to produce images of high fidelity, results in a robust discovery architecture for visual neuroscience. By using only a small number of synthetic images created by NeuroGen, we demonstrate that we can detect and amplify differences in regional and individual human brain response patterns to visual stimuli. We then verify that these discoveries are reflected in the several thousand observed image responses measured with fMRI. We further demonstrate that NeuroGen can create synthetic images predicted to achieve regional response patterns not achievable by the best-matching natural images. The NeuroGen framework extends the utility of brain encoding models and opens up a new avenue for exploring, and possibly precisely controlling, the human visual system.

1. Introduction

Light rays reaching the retina are converted into bioelectrical signals and carried through the ophthalmic projections to the brain, where incoming signals are represented by corresponding neural activation patterns in the visual cortex (Wandell et al., 2007). The specific patterns of neural activation in response to visual stimuli are determined in part by the texture, color, orientation and content of the visual stimuli. The visual system has provided a rich model for understanding how brains receive, represent, process and interpret external stimuli, and has led to advances in understanding how the human brain experiences the world (Thorpe et al., 1996; Van Essen et al., 1992).

Much is known about how regions in the visual cortex activate in response to different image features or content. Our knowledge of stimulus-response maps has mostly been derived from identifying features that maximally activate various neurons or populations of neurons (Hubel and Wiesel, 1962; 1968). Non-invasive techniques such as functional MRI (fMRI), are now one of the most utilized approaches for measuring human brain responses to visual (and other) stimuli (Allen, 2021; Van Essen, 2013). The responses of early visual areas such as primary visual cortex (V1) have been studied using population receptive field (pRF) experiments wherein a participant fixates on a central dot while patterned stimuli continuously moved in the visual field (Wandell et al., 2007). Neurons in early visual areas have been found to be selective for stimulus location, but also other low-level stimulus properties such as orientation, direction of motion, spatial and temporal frequency (De Valois and De Valois, 1980; DeAngelis et al., 1995; Hubel and Wiesel, 2020; Movshon et al., 1978). Recently, intermediate visual areas, like V2 or V4, were found to be responsive to textures, curved contours or shapes (Nandy et al., 2013; Ziemba et al., 2016). Late visual area activations have typically been explored by contrasting response patterns to images with varied content, e.g. to faces, bodies, text, and places. For example, the fusiform face area (FFA) (Kanwisher et al., 1997) involved in face perception, the extrastriate body area (EBA) (Downing et al., 2001) involved in human body and body part perception, and the parahippocampal place area (PPA) (Epstein and Kanwisher, 1998) involved in perception of indoor and outdoor scenes, have been defined by con-

* Corresponding author.
E-mail address: amk2012@med.cornell.edu (A. Kuceyeski).

https://doi.org/10.1016/j.neuroimage.2021.118812.
Received 20 July 2021; Received in revised form 11 October 2021; Accepted 12 December 2021
Available online 20 December 2021.
1053-8119/© 2021 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

trasting patterns of brain activity evoked by images with different content. However, this approach has several limitations: the contrasts 1) are constructed based on a priori hypotheses about stimulus-response mappings, 2) are by nature limited to the set of images presented to the individual while they are in the scanner, 3) are subject to noise in the observed brain responses, and 4) may vary widely across individuals (Benson and Winawer, 2018; Seymour et al., 2018).

The recent explosion of machine learning literature has centered largely around Artificial Neural Networks (ANNs). These networks, originally inspired by how the human brain processes visual information (Rosenblatt, 1958), have proved remarkably useful for classification or regression problems of many types (Belagiannis et al., 2015; He et al., 2016; Krizhevsky et al., 2012; Simonyan and Zisserman, 2014; Toshev and Szegedy, 2014). Common applications of ANNs are in the field of computer vision, including image segmentation (Girshick et al., 2014), feature extraction (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014) and object recognition (Sermanet, 2013). Meanwhile, in the field of neuroscience, researchers have incorporated ANNs into "encoding models" that predict neural responses to visual stimuli and, furthermore, have been shown to reflect structure and function of the visual processing pathway (Cichy et al., 2016; Khaligh-Razavi and Kriegeskorte, 2014; Khosla et al., 2020; St-Yves and Naselaris, 2018a). Encoding models are an important tool in sensory neuroscience, as they can perform "offline" mapping of stimuli to brain responses, providing a computational stand-in for a human brain that also smooths measurement noise in the stimuli-response maps. ANNs' internal "representations" of visual stimuli have also been shown to mirror biological brain representations of the same stimuli, a finding replicated in early, mid and high-level visual regions (Cadena, 2019; Yamins, 2014). This observation has led to speculation that primate ventral visual stream may have evolved to be an optimal system for object recognition/detection in the same way that ANNs are identifying optimal computational architectures.

An alternative approach to understanding and interpreting neural activation patterns is decoding, in which the stimulus is reconstructed based on its corresponding neural activity response pattern. The presence of distinct semantic content in natural movies, e.g. object and action categories, has previously been decoded from fMRI responses with high accuracy using a hierarchical logistic regression graphical model (Huth, 2016). Beyond semantic content, natural scenes and even human faces can be reliably reconstructed from fMRI using generative adversarial network (GAN) approaches (Mozafari et al., 2020; Shen et al., 2019; St-Yves and Naselaris, 2018b; VanRullen and Reddy, 2019). Encoding and decoding models, in conjunction with state-of-the-art generative networks, may also allow single neuron or neural population control. Recent work in macaques used an ANN-based model of visual encoding and closed-loop physiological experiments recording neurons to generate images specifically designed to achieve maximal activation in neurons of V4; the resulting synthetic images achieved higher firing rates beyond what was achieveable by natural images (Bashivan et al., 2019). Moreover, by adopting a pretrained deep generative network and combining it with a genetic algorithm, realistic images were evolved to maximally activate target neurons in monkey's inferotemporal cortex (Ponce, 2019). Both studies' results suggested the synthetic images uncovered some encoded information in the observed neurons that was consistent with previous literature, and, furthermore, that they evoked higher responses than any of the natural images presented. One recent study applied generative networks to synthesize preferred images for functionally-defined regions of interest in the human brain, specifically FFA, EBA and PPA, and were able to replicate regional category selectivity (Ratan Murty et al., 2021). However, generative networks have not yet been applied to investigate 1) inter-individual and inter-regional differences in image features that maximize activation in single regions of the human visual cortex or 2) image features that are designed to achieve more complex optimizations of activation patterns over multiple regions of the human visual cortex.

In this work, we build upon three recent advances in the literature. The first is the existence of the Natural Scenes Dataset (NSD), which consists of densely-sampled fMRI in eight individuals who each participated in 30–40 fMRI scanning sessions in which responses to 9,000–10,000 natural images were measured (Allen, 2021). The second is in an interpretable and scalable encoding model, based on the NSD data, that performs accurate individual-level mapping from natural images to brain responses (St-Yves and Naselaris, 2018a). The third is in the development of generative networks which are able to synthesize images with high fidelity and variety (Brock et al., 2018; Nguyen et al., 2016). Here, we propose a state-of-the-art generative framework, called NeuroGen, which allows synthesis of images that are optimized to achieve specific, predetermined brain activation responses in the human brain. We then apply this framework as a discovery architecture to amplify differences in regional and individual brain response patterns to visual stimuli.

2. Materials and methods

2.1. Natural scenes data set

We used the Natural Scenes Dataset (NSD; http://naturalscenesdataset.org (Allen, 2021) to train the encoding model. In short, the NSD dataset contains densely-sampled functional MRI (fMRI) data from 8 participants collected over approximately a year. Over the course of 30–40 MRI scans, each subject viewed 9,000–10,000 distinct color natural scenes (22,000–30,000 trials with repeats) while undergoing fMRI. Scanning was conducted at 7T using whole-brain gradient-echo EPI at 1.8-mm iso–voxel resolution and 1.6s TR. Images were sourced from the Microsoft Common Objects in Context (COCO) database (Lin, 2014), square cropped, and presented at a size of 8.4° × 8.4°. A set of 1000 images were shared across all subjects; the remaining images for each individual were mutually exclusive across subjects. Images were presented for 3s on and 1s off. Subjects fixated centrally and performed a long-term continuous recognition task on the images in order to encourage maintenance of attention. The fMRI data were pre-processed by performing one temporal interpolation (to correct for slice time differences) and one spatial interpolation (to correct for head motion). A general linear model (GLM) was used to estimate single-trial beta weights, representing the voxel-wise activation in response to the image presented. Specifically, there are three components in the GLM: first, a library of hemodynamic response functions (HRFs) derived from an initial analysis of the dataset was used as an efficient and well-regularized method for estimating voxel-specific HRFs; second, the GLMdenoise technique (Charest et al., 2018; Kay et al., 2013) was adapted to the single-trial GLM framework, thereby enabling the use of data-driven nuisance regressors; third, an efficient ridge regression (Rokem and Kay, 2020) was used to regularize and improve the accuracy of the betas. Cortical surface reconstructions were generated using FreeSurfer, and both volume- and surface-based versions of the response maps were created.

In addition to viewing the COCO images, individuals all underwent the same image functional localizer (floc) task to define the visual region boundaries (Stigliani et al., 2015). In short, regions of interest were defined by contrasting activation maps for different types of localizer images (floc-bodies, floc-faces, floc-places, floc-words). Several regions exhibiting preference for the associated category were defined (e.g., floc-faces was based on t-values for the contrast of faces>non-faces). Regions were defined by drawing a polygon around a given patch of cortex and then restricting the region to vertices within the polygon that satisfy t>0. See Supplementary Figure S10-18 for the 8 individuals' early and late visual region maps, along with quantification of the regions' overlaps using Dice coefficient.

2.2. Deepnet feature-weighted receptive field encoding model

We trained a Deepnet feature-weighted receptive (fwRF) encoding model (St-Yves and Naselaris, 2018a) using the paired NSD images and fMRI response maps described above. Here, instead of the voxel-wise model previously created, we trained a model for each of the 24 early and late visual regions for each of the 8 individuals in the NSD dataset that predicts a single number - the average response over the voxels in that region for that individual. There are three components in the Deepnet-fwRF model: K feature maps, a vector of feature weights w_k, and a feature pooling field. The output of each fitted encoding model is the predicted activation $\hat{r}$ for a given region for a given individual in response to an image S:

$$\hat{r}(S) = \sum_{k}^{K} w_k \int_{-D/2}^{D/2} \int_{-D/2}^{D/2} g(x, y; \mu_x, \mu_y, \sigma_g) f^k_{i(x)j(y)}(S) dx dy$$

where w_k is the feature weight for kth feature map f^k, g is the feature pooling field (described below), D is the total visual angle sustained by the image, $i(x) = \lfloor (2x + D)/2\Delta \rfloor$ (likewise for $j(y)$) is the discretization depending on $\Delta = D/n_k$ which is the visual angle sustained by one pixel of a feature map with resolution, and $n_k \times n_k$ is the resolution of kth feature map.

The feature maps f^k were obtained from Alexnet (Krizhevsky et al., 2012), a deep convolutional neural network containing 5 convolutional layers (interleaved with max-pool layers) and 3 fully-connected layers. AlexNet was originally trained for classification of images in ImageNet (Russakovsky, 2015), and is often used to extract salient features from images. The exact structure and trained network can be downloaded as part of the Pytorch library, and is also available at https://github.com/pytorch/vision/blob/master/torchvision/models/alexnet.py. The feature maps can be drawn from all convolutional layers and fully-connected layers in Alexnet. To limit the total number of feature maps, we first set the maximum feature maps for each layer to 512. For those layers whose dimension exceeded 512, we calculated the variance of the layer values across the image set and retained those 512 feature maps with the highest variance. We then concatenated the selected feature maps having the same spatial resolution, which resulted in three feature maps of size $(256, 27, 27)$, $(896, 13, 13)$ and $(1536, 1, 1)$.

The fwRF model was designed based on the hypothesis that the closer the feature map pixel is to the center of the voxel's feature pooling field, the more it contributes to the voxel's response. We also assume this is the case for clusters of voxels (regions in our case). The feature pooling field was modeled as an isotropic 2D Gaussian blob:

$$g(x, y; \mu_x, \mu_y, \sigma_g) = \frac{1}{\sqrt{2\pi}\sigma_g} \exp\left[-\frac{(x-\mu_x)^2 + (y-\mu_y)^2}{2\sigma_g^2}\right]$$

where μ_x, μ_y are the feature pooling field center, and σ_g is the feature pooling field radius. The feature pooling field center and radius were considered hyperparameters and learned during training of the encoding model. By definition, when the radius of the feature pooling field is very large (e.g. $\sigma_g \gg \Delta$), the predicted activation from a single layer reduces to a weighted sum of all pixels within the field; otherwise, it reduces to just one single spatial unit. In experiments, the grid of candidate receptive fields included 8 log-spaced receptive field sizes between 0.04 and 0.4 relative to $1/n_k$, and the candidate feature pooling field centers were spaced 1.4 degrees apart (regardless of size), resulting in a total of 875 candidate feature pooling fields. We searched the regularization parameter over 9 log-spaced values between $10^3 \sim 10^7$ and chose the one that had the best performance on the held-out validation set of 3000 images.

Each of the 2688 image feature maps has an associated feature weight, w_k, indicating how important the kth feature map was for predicting that region's activity. Once the feature pooling hyperparameters were determined, a set of feature weights were then learned via ridge regression for each visual region in each of the 8 individuals. Since the images that each individual viewed were different and the individuals had a slightly different number of image-activation pairs, we used the individual-specific images for training and validating the ridge models and tested the ridge models on the shared 1000 images where we calculated the predicted accuracy. During training, 3000 randomly selected image-activation pairs were held-out and used as a validation set with which to identify the optimal hyperparameters, namely those that maximized the prediction accuracy within this validation set.

2.3. The BigGAN-deep image generator network

We used the generator from BigGAN-deep, a pretrained deep generative adversarial network (GAN), utilizing the ImageNet image set, whose goal is to synthesize images of a given category that look natural enough to fool an automated fake/real classifier (Brock et al., 2018). The BigGAN-deep was built upon the self-attention GAN (SA-GAN) (Zhang et al., 2019) with some differences: i) a shared class embedding was used for the conditional BatchNorm layers to provide class information, ii) the entire noise vector z was concatenated with the conditional vector to allow the generator G to use the latent space to directly influence features at different resolutions and levels of hierarchy, iii) the noise vector z was truncated by resampling the values with magnitude above a chosen threshold to fall inside that range instead of using a Normal or Uniform distribution and iv) orthogonal regularization was used to enforce the models' amenability to truncation. BigGAN-deep's truncation threshold controls the balance of fidelity and variety of the synthetic images: larger thresholds lead to higher variety but lower fidelity images while lower thresholds lead to higher fidelity images of lower variety. In our experiments, the truncation parameter was set to 0.4 to achieve a balance of both fidelity and variety. The PyTorch version of the pretrained model can be publicly downloaded at https://github.com/huggingface/pytorch-pretrained-BigGAN.

2.4. Synthesizing images to optimize regional predicted activation

We aimed to synthesize images using our generator network G that either maximized the activation in single or pairs of regions or maximized activation in one region while minimizing activation in another region, depending on the experiment. For the sake of simplicity, we will describe the optimization procedure using the example goal of maximizing activation in a single region. Because our generator network (BigGAN-deep) is a conditional GAN, it requires identification of an image class via a one-hot encoded class vector c, then uses a noise vector z to generate an image of that class. We performed the optimization in two steps; first, we identified the 10 most optimal classes, then, for each class we optimized over the noise vector space. To identify optimal classes, we generated 100 images from each of the 1000 classes in the ImageNet database using 100 different random noise vectors. We input the resulting images into the encoding model to obtain associated predicted regional activation, which was averaged over the 100 images per class. The 10 classes that gave the highest average predicted activation for the region of interest were identified and encoded in c_i $(i = 1, \ldots 10)$, and corresponding image generator noise vectors z_{ij} $(i = 1, \ldots 10,\ j = 1, \ldots 10)$ were further optimized via backpropagation with 10 random initialization seeds per class. The result of this optimization is a set of 100 images $G(c_i, z_{ij})$ that yielded maximal predicted activation of the target region. Formally, and including a regularization term, we posed the optimization problem as finding the codes $\hat{z}_{ij}$ such that:

$$\hat{z}_{ij}(c_i) = \arg\max_{z_{ij}} (\hat{r}_t(G(c_i, z_{ij})) - \lambda \|z_{ij}\|)$$

where $\lambda = 0.001$ was the regularization parameter. For the maximization/minimization of region pairs, the cost function aimed to maximize/minimize the sum of the two regions' activations together, with both regions considered equally. For the maximization of one region and the minimization of the other, the cost function was the sum of the

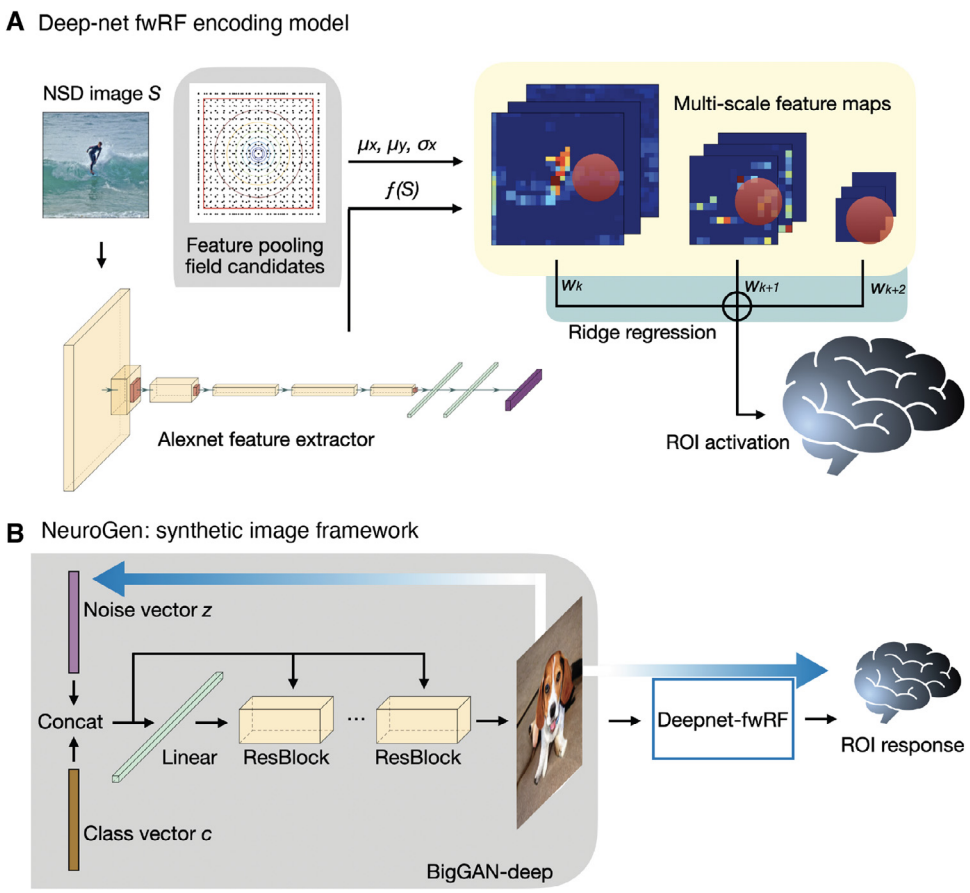


Fig. 1. Deepnet-fwRF encoding model and NeuroGen framework for synthetic image generation. **A** The deepnet-fwRF encoding model begins by passing an image through the Alexnet feature extractor, then applying a 2D Gaussian pooling receptive field to obtain multi-scale feature maps. Finally, ridge regression is applied to the multi-scale features to predict brain region-specific responses to the image. **B** The deepnet-fwRF encoding model is concatenated with a pretrained conditional generative network (BigGAN-deep) to synthesize images that are predicted to optimally match a desired response pattern (e.g. maximizing predicted activation in the fusiform face area). The optimized synthetic images are created in three steps. First, a single image for each of the 1000 classes in the conditional GAN is created from an initial truncated Gaussian noise vector; the resulting images are provided to the encoding model to obtain their predicted activation responses. Second, the 10 classes that give the predicted activation best matching the target activation are identified (e.g. those that give maximal predicted activation in the fusiform face area). Third, fine tuning of the noise vectors for each of the 10 synthetic images via gradient descent is performed; gradients flow from the encoding model's predicted response back to the synthetic image and to the noise vector that initializes the conditional GAN.

max and min, with both regions considered equally. Once this optimization was performed, we selected the top 10 images by taking the most optimal image from each of the 10 distinct classes.

3. Results

3.1. Encoding models accurately map images to brain responses and remove noise present in fMRI activation maps

First, subject and region-specific encoding models were built using the Deepnet-fwRF framework (St-Yves and Naselaris, 2018a) and the NSD data, which contained between 20k to 30k paired images and fMRI-based brain response patterns for each of eight individuals (Fig. 1A) (Allen, 2021). Other than a shared set of 1000, the images were mutually exclusive across subjects. The encoding models first extract image features using AlexNet, a pre-trained deep neural network for image classification (Krizhevsky et al., 2012). The encoding model then applies a 2D Gaussian pooling field to the image features to obtain a set of multi-scale feature maps. Ridge regression is used in the final step to predict individual regional activations (which is the average activation over all voxels within that region) from the multi-scale feature maps. For each of the eight subjects, we trained a separate model for each early/late visual region, of which there were at most 24. Regions were defined using functional localizer tasks (Allen, 2021) and some were missing in certain individuals.

To objectively evaluate the regional encoding models' predictive abilities, we compared their prediction accuracies against the reliability of repeated fMRI measurements of responses to the same image as a noise ceiling estimate. A region's observed and predicted activations are the mean of those quantities over all voxels in that region. We first selected the images in the held-out shared image set that were viewed twice by each subject, totalling between 700–1000 images per person (see Supplementary Figure S1 for the prediction accuracy for all images, including repeats, in the held-out shared image set). Model prediction accuracy was calculated as the Pearson correlation between the predicted activation for images that viewed twice with their measured fMRI responses, where the two repeated responses/predictions per image were concatenated. Note that the measured responses are different for the two repeats of the same image, but the predictions are the same. The noise ceiling was calculated by correlating the two measured fMRI responses for the same set of images (see Fig. 2A). Distributions of the prediction accuracies and noise ceilings for each subject and each region's encoding models are shown in Fig. 2B–F; the prediction accuracy and noise ceiling for the same subject are connected by a dashed line. Generally, regions with higher noise ceilings had higher accuracies. In all cases, prediction accuracies were greater than noise ceilings, reflecting the noise in fMRI measurements as well as the good predictive performance of the encoding models.

One approach we explore here is to assume that images with similar predicted responses can be treated as repeated measurements under similar conditions. Thus, we first rank-order the images in the held-out shared image set by their predicted activations, then apply non-overlapping windows of fixed size to group the images. We then average the predicted and measured responses within these windows and calculate the Pearson correlation of the two resulting averages across all windows, see Fig. 3A. As the window size in the smoothed accuracy calculation increases, the correlation also increases (see in Fig. 3B–F), thus validating the noise-smoothing ability and relatively high accuracy of the encoding model. It is clear from this analysis that the encoding model accuracy estimates are limited by the noise present in the fMRI measurements. For comparison in the null case of no correlation between predicted and measured responses, we calculated the smoothed accuracy when the measurement-response pairs were uncoupled (see Supplementary Figure S2 for the smoothed ac-

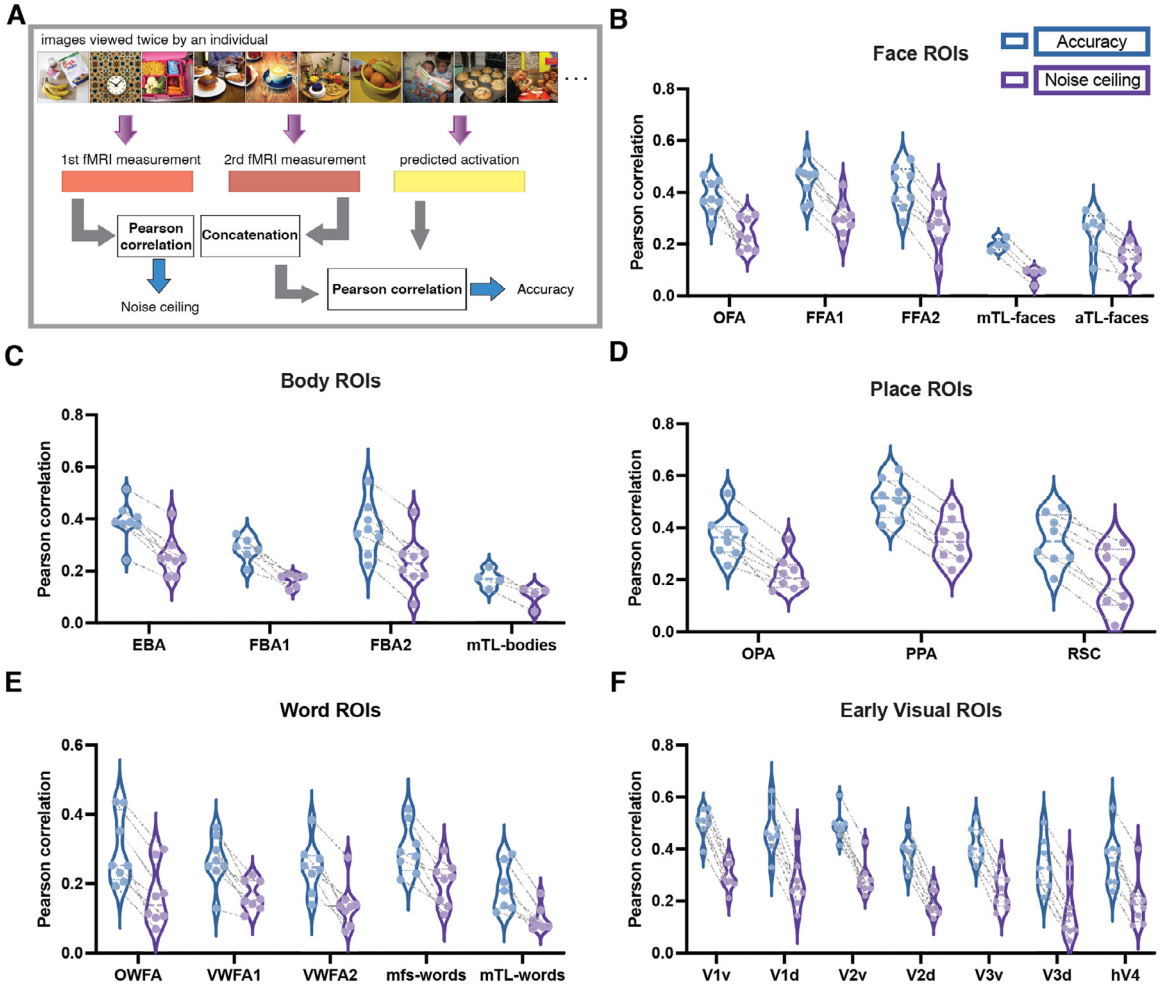


Fig. 2. The Deepnet-fwRF encoding model accurately predicts brain response to visual stimuli. **A** Accuracy and noise ceiling calculation. For comparison, we selected those images in the held-out shared image set that were viewed twice by each subject to quantify the reliability of repeated measurements as a noise ceiling for our model predictions. The noise ceiling thus was calculated as the Pearson correlation between the two fMRI measurements, while the prediction accuracy was calculated as the Pearson correlation between the predicted activation and the concatenated fMRI measurements. Note the measured responses are different for the two repeats of the same image, but the predictions are the same. **B, C, D, E** and **F** Distribution of the individuals' prediction accuracies and noise ceilings for faces, body, place, word and early visual ROIs, respectively. Each point in the violin plot represents an individual; accuracy and noise ceiling of the same subject are connected by dashed line. Face ROIs: OFA - occipital face area; FFA - fusiform face area; mTLfaces - medial temporal lobe face area; aTLfaces - anterior temporal lobe face area. Body ROIs: EBA - extrastriate body area; FBA - fusiform body area; mTLbodies - medial temporal lobe body area. Place ROIs: OPA - occipital place area; PPA - parahippocampal place area; RSC - retrosplenial cortex. Word ROIs: OWFA - occipital word form area; VWFA - visual word form area; mfswords - mid-fusiform sulcus word area; mTLwords - medial temporal lobe word area. Early visual ROIs: v - ventral; d - dorsal.

curacy calculations when the predicted activation vector is randomly permuted).

3.2. Optimized synthetic images more consistently align with expectations of regional feature selectivity than natural images.

After validating the encoding model, we sought to generate images whose predicted activation optimally achieved certain criteria, e.g. had a maximal predicted activation in OFA (as quantified by the average activation over all voxels in the OFA). A conditional deep generative network (BigGAN-deep), based on the diverse set of natural images in ImageNet, was adopted as NeuroGen's synthetic image generator. By constructing the activation maximization scheme (Fig. 1B), which connects the encoding model and the image generator, gradients flowed from the regional predicted activation back to the noise vector to allow optimization. Because BigGan-deep is conditional, the class of the image is identified prior to synthesis which allows generation of more realistic synthetic images. We began optimization by first identifying the top 10 classes that gave the best match to our desired activation pattern. Once the top 10 classes were identified, the noise vector was then fine-tuned as previously described. We began by using this NeuroGen framework to generate images that give maximal predicted activation for a single region in turn. Fig. 4A–E, shows the i) 10 natural images that had highest measured fMRI activation (top row) ii) 10 natural images that had the highest predicted activation from the encoding model (middle row) and iii) the 10 synthetic images optimized to achieve maximal predicted activation from the encoding model (bottom row) for a subject whose encoding model prediction accuracy was closest to the median accuracy. Word clouds representing the semantic content of all the top 10 images from the three sources are also provided by collecting the labels from the natural images and obtaining labels for the synthetic images using an automated image classifier (Ren et al., 2015). The images and word clouds largely show alignment with expected image content and agreement across the three sources; however, when looking at the individual images, the content and features of both the natural and synthetic images with highest *predicted* activations are more aligned with a priori expectations than the content and features of images with highest *measured* activations. For example, the images with the highest observed

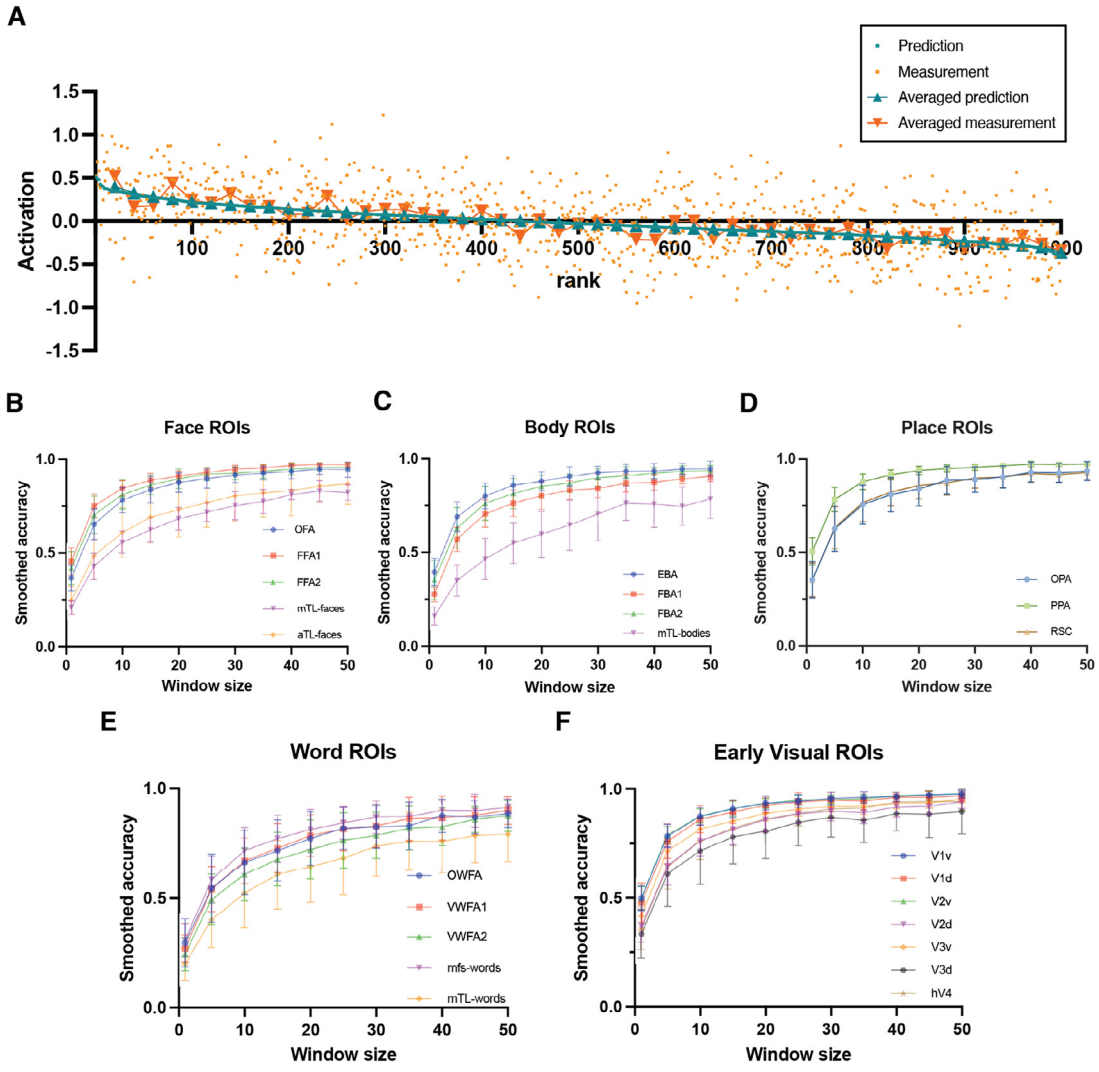


Fig. 3. The Deepnet-fwRF encoding model can smooth noise present in fMRI-based individual image response maps. **A** An example illustrating the relationship between predicted/measured activation for single images and, averaged predicted activation within non-overlapping windows of size 20, averaged measured activation within non-overlapping windows of size 20, with the rank of the images ordered by predicted activation. Smoothed accuracy is calculated by first ranking order the held-out images by their predicted activation, and then applying non-overlapping windows and correlating the averaged predicted activations within that window with the averaged measured activations within those windows. **B, C, D, E** and **F** The smoothed accuracy increases as the window size increases for face, body, place, word and early visual ROIs respectively. The "smoothed accuracy" all show the trend of approximating 1 when increasing the window size.

activation include a giraffe for the face area, a zebra for the body area and many of the word area top images do not contain text. We posit that this is due to the encoding model predictions being less noisy than fMRI measurements.

We do see a general agreement in the content/features of the synthetic images with what is expected from how the regions are defined. For FFA1 (face), the top 10 classes varied across individuals but were generally those that produced optimized images with human faces (ImageNet categories are terms like groom, mortarboard, shower cap, ice lolly, bow tie etc.) or dog faces (ImageNet categories were Pembroke, Basenji, Yorkshire terrier, etc.); the wordcloud reflects this as "person" is the dominant term. One interesting thing to note when comparing the natural and synthetic images for FFA1 is that the synthetic images tend to contain more than one person's face. Most of the top ImageNet classes for EBA (body) were sports-related (rugby ball, basket ball, volleyball, etc.) and the resulting images' most prominent features were active human bodies; the wordcloud again prominently features "person" but also sport-related terms. Typical indoor and outdoor place scenes emerged for place area PPA (wordcloud contained "dining table", "chair", "bed", etc.) and images with objects and/or places containing text were generated for word area mfs-words (wordcloud contained "book", "bottle" and "truck"). Synthetic images with high spatial frequency and color variety were produced when maximizing V1v (early visual), which generally responds to low-level image features such high frequency textures. In addition to the good agreement of the content and features of the optimized synthetic images with expectations, they are also predicted by the encoding model to achieve significantly higher activation than the top natural images (see Supplementary Figure S7). In fact, for all single region optimizations (over all 8 subjects and 24 regions), the synthetic images had significantly higher predicted activations than the top natural images. Supplementary Video S1 shows a video of the sets of top natural and synthetic images for all 8 individuals, for all 24 regions; Supplementary Figures S3-6 for wordclouds for all regions. Looking across all single region optimizations, there were some obvious differences in image content/features that emerged across individuals within the same region and within the same individual across regions in the same perception category, some of which we investigate further below. These results provide evidence that the NeuroGen framework can produce im-

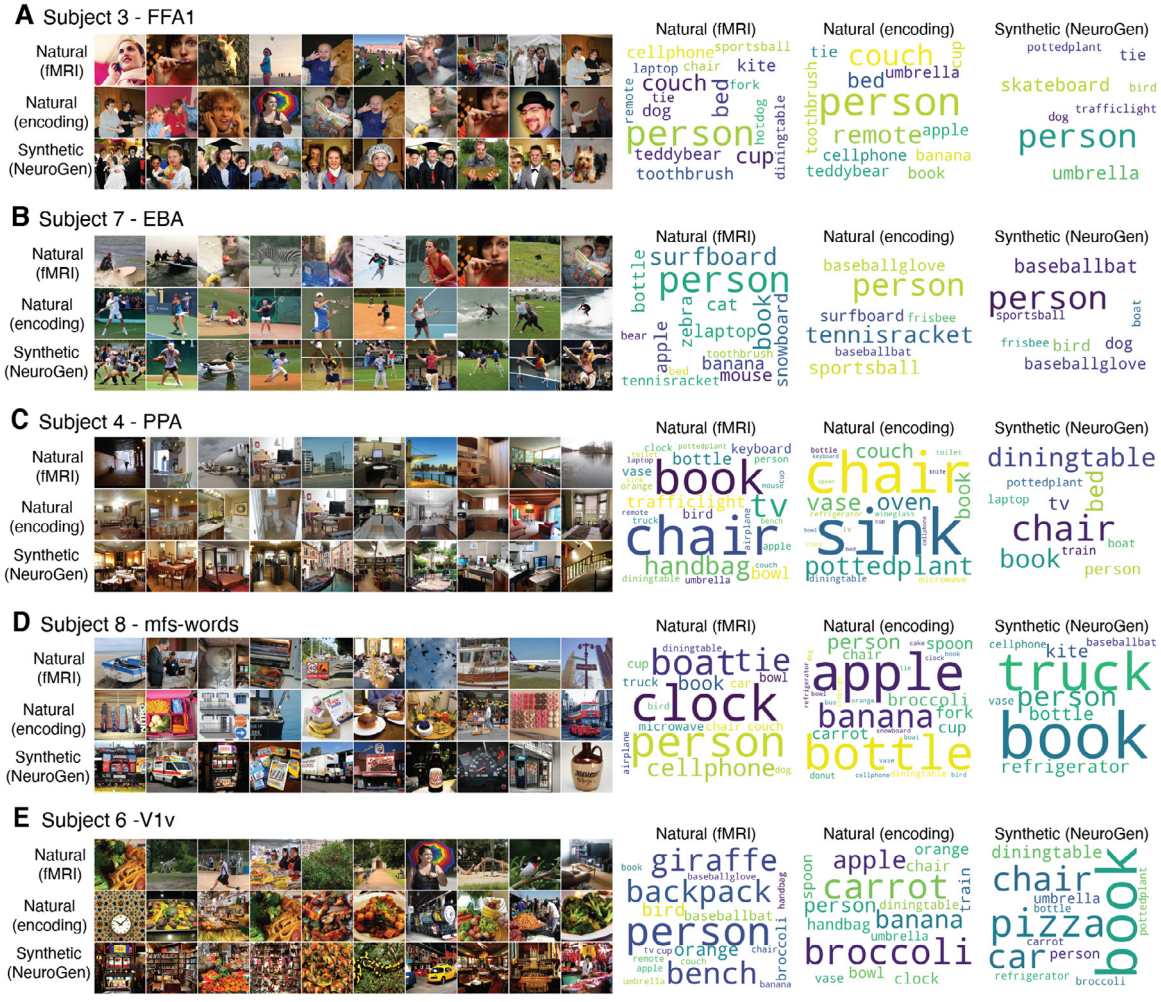


Fig. 4. Synthetic images generated to maximize predicted activation in five example visual regions in different individuals. In each panel, the top row contains the 10 natural images that had highest measured fMRI responses for that region, the second row contains the 10 natural images that had the highest encoding model predictions for that region, and the third row contains the 10 optimized synthetic images from NeuroGen for that region. Five example regions are shown, i.e. FFA1 (face), EBA (body), mfs-words (word), PPA (place) and V1v (early visual), each for a different individual that had median encoding model accuracy. The top encoding model images and NeuroGen images appear more consistent in their reflection of expected features/content compared to the top natural images with highest observed activation (e.g. fMRI top images include a giraffe for FFA1, a zebra for EBA and many lack text in the mfs-words area). The corresponding wordcloud plots show the image labels. **A** Subject 3's FFA1 region. **B** Subject 7's EBA region. **C** Subject 4's PPA region. **D** Subject 8's mfs-words region. **E** Subject 6's V1v region.

ages that generally agree in content and features with *a priori* knowledge of neural representations of visual stimuli, and may be able to amplify differences in response patterns across individuals or brain regions.

3.3. Optimized synthetic images reflect and amplify features important in evoking individual-specific and region-specific brain responses

One deviation in the content of the synthetic images from what was generally expected was the prevalence of dogs in all the five of face regions we modeled; 96 of the 350 top images (10 top images × 8 individuals × 5 face areas - 10 × 5 missing subjects' face areas) were of dog faces and 219 were of human faces. We observed some individuals' top 10 synthetic images all contained dogs while others had none. This imbalance also varied by brain region, one individual could have all synthetic images containing dogs for one face region while they would have fewer in another face region (see Supplementary Video S1). This apparent region- or individual-specific dog versus human preference was not apparent from the content of either the top 10 natural images giving highest observed or predicted activations. Figure 5A–C, show all eight subjects' 10 natural images with the highest measured activation, 10 natural images with the highest predicted activation and 10 synthetic images, respectively, for an example face area (FFA1). For subject 3, nine out of ten synthetic images contain human faces, with only one dog. On the other hand, subject 4 had eight dog images, one human and one lion. Frequency based wordcloud plots of the top 10 synthetic images' labels for all individuals for each of the five face areas are shown in the third row of Fig. 6. The wordclouds confirm the regionally-varying prevalence of dogs in NeuroGen's synthetic images, which is not obvious from looking at either of the natural image wordclouds. We hypothesized that this individually and regionally varying dog/human preference in the synthetic images may be reflecting the actual underlying preferences in the data. To test this hypothesis, we calculated 1) the t-statistic of the measured fMRI activation from dog images (images in the NSD dataset that had a single label "dog") and the measured fMRI activation from images of humans (images in the NSD dataset that had a label "person" and any of the following: "accessory", "sports") and 2) the top 10 and top 100 synthetic images' dog vs human image ratios, calculated as ((number of dog images - number of person images)/(number of dog images + number of person images)). We calculated these two measures (dog vs. person t-stat and synthetic image ratio) for all eight individuals and all five face ROIs (see Fig. 5D). We found a significant correlation between the t statistic and image ratios for NeuroGen's top 10

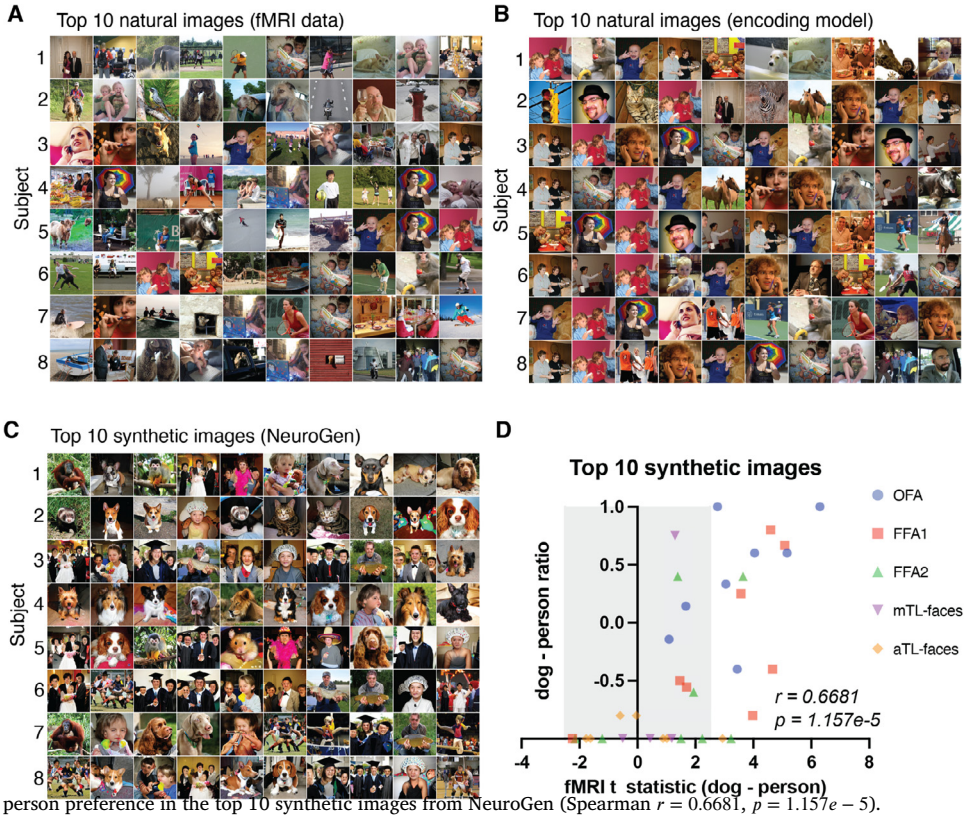


Fig. 5. Individual-specific and region-specific differences in face region responses are reflected in and amplified by the NeuroGen framework. **A, B** and **C** Sets of images that had the highest activation in FFA1 (fusiform face area 1) for all individuals, one per row, derived from three different sources. **A** Natural images that have the highest observed activation measured directly via fMRI. **B** Natural images that have the highest predicted activations from the encoding model. **C** Synthetic images that were created using NeuroGen. **D** The x-axis displays the dog vs. person preference from the observed fMRI data, quantified by the t-statistic of observed fMRI activations from all natural dog images compared to the observed activations from all natural people images. The quantities were calculated for each of the five face areas in each of the eight individuals. The y-axes represent the dog vs person preference present in the top 10 synthetic images, calculated by taking the difference in the count of dog images minus the count of person images, divided by the total count of dog and person images. Values close to -1 indicate strong person preference and values close to 1 indicate strong dog preference. Points outside the grey area have t-statistics that are significant after FDR correction. A significant correlation exists between the observed dog-person preference from the entire fMRI dataset and the dog-person preference in the top 10 synthetic images from NeuroGen (Spearman $r = 0.6681$, $p = 1.157e-5$).

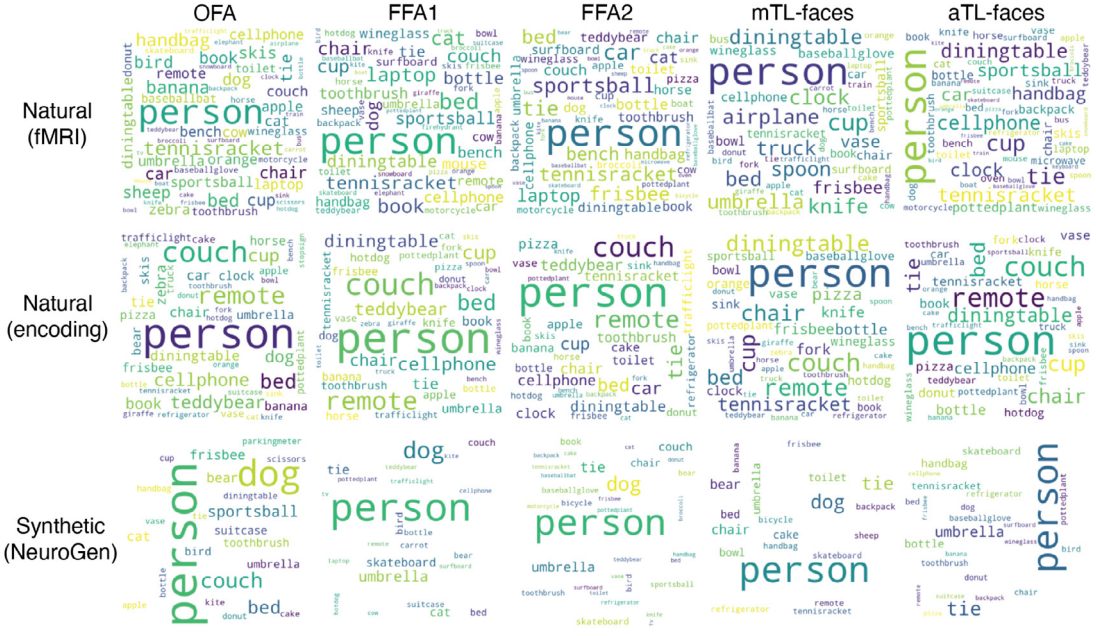


Fig. 6. Face area preferences in semantic image content are reflected in the wordclouds illustrating image content for the top 10 natural and synthetic images. Each row represents the source of the top 10 images: natural images that have the highest observed activation measured directly via fMRI, natural images that have the highest predicted activations from the encoding model, and synthetic images that were created using NeuroGen. Each column represents one of the five face regions. The presence of the "dog" label (as well as, unsurprisingly, "person" label) can be appreciated most prominently in NeuroGen's synthetic images.

synthetic images (Spearman rank $r = 0.6681$, $p = 1.157e-5$); this correlation was even higher when considering NeuroGen's top 100 synthetic images (Spearman rank $r = 0.7513$, $p = 1.987e-7$), see Supplementary Figure S8. Correlations with the top 10 natural images from the encoding model and fMRI measurements were also significant, but not as strong as the top 10 synthetic images correlation (Spearman rank $r = 0.4468$, $p = 7.126e-3$ and $r = 0.2948$, $p = 0.086$), respectively. These results show that previously identified "face" regions in the human visual cortex also respond robustly to dog faces, and, furthermore, that the dog/human balance in response patterns varies across individuals and brain regions. These results highlight NeuroGen's potential as a discovery architecture, which can be used to amplify and concisely summarize (even with only 10 images) region-specific and individual-specific differences in neural representations of visual stimuli.

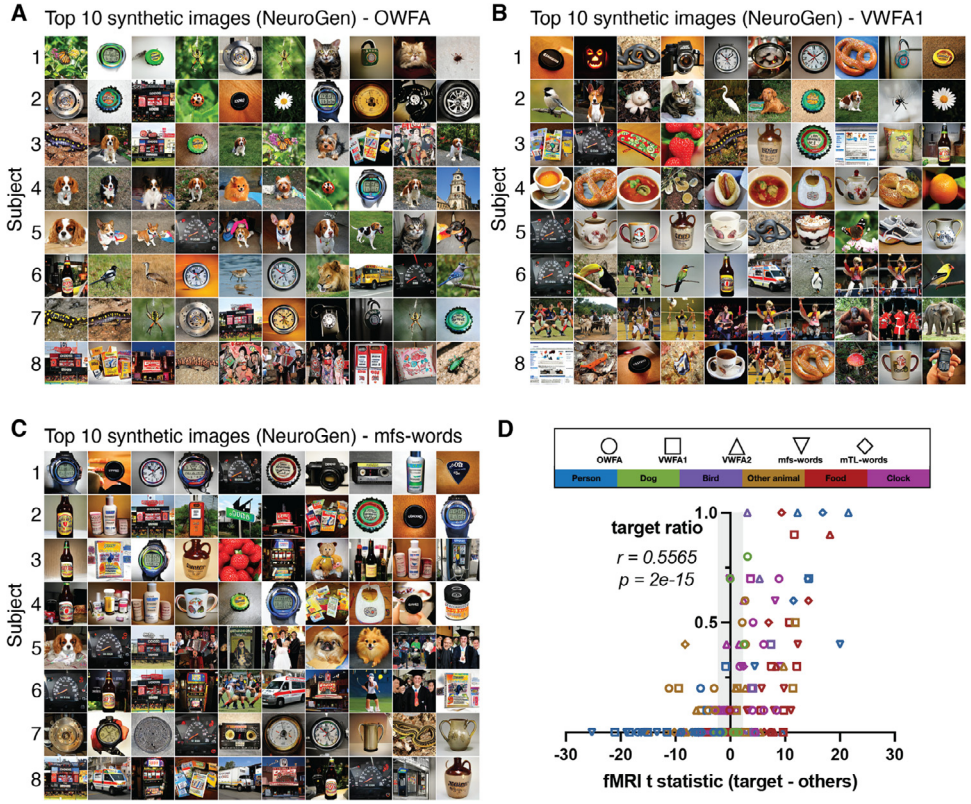


Fig. 7. Individual-specific and region-specific differences in word area responses are reflected in and amplified by only 10 images from the NeuroGen framework. Sets of images that had the highest activation in **A** OWFA (occipital word form area), **B** ventral word form area 1 (VWFA1) and **C** mfs-words, for all eight individuals (one per row). **D** The x-axis displays the activation contrast for six categories of natural images (human, dog, bird, other animal, food and clock) from the observed fMRI data, quantified by the t-statistic of observed fMRI activations from natural images of the category in question compared to the observed activations from all other natural images (not containing that item). The y-axis represents the proportion of top 10 synthetic images that contain that item. Points outside the grey area have t-statistics that are significant after FDR correction. A significant correlation exists between the observed image category contrasts from the entire fMRI dataset and the proportions of that image category in the top 10 synthetic images from NeuroGen (Spearman rank $r = 0.557$, $p = 2e-15$).

Another unexpected observation of individual and regional variability was found in the word areas. Because the images in NSD are natural, they do not solely contain text; therefore many of the top natural and synthetic images for the word-preferred regions contained objects or scenes with integrated text (scoreboard, packet, cinema, bottle cap, sign, etc.), see Fig. 7. However, images with integrated text were only a portion of the word form area top images; many also contained humans, dogs, cats, birds, other animals, food and clocks. We aimed to test if the content of the synthetic images from NeuroGen could accurately reflect underlying patterns in the measured activation data, even when considering several categories of images. Therefore, we calculated the proportion of each individual's top 10 images that contained objects in each of the six categories of interest (humans, dogs, birds, other animals, food and clocks) for each of the top 5 word areas. These six categories of interest were selected based on the most common categories in the top 10 synthetic images across all 5 word form areas. We then correlated each individual's synthetic image proportions with their t-statistic of measured activation for images containing that target object (and not any objects in the other 5 categories) contrasted against the measured activation for all images not containing that object. Categories were included for a given word form area's analysis if there existed at least 10 images of that category out of the 240 total top natural or synthetic images for that region (8 individuals x 10 images x 3 sources). Figure 7D shows the scatter plot of the image proportions versus the t-statistic giving the activation contrast for synthetic images of a given category (indicated by color) for the various word form regions (indicated by shape), which were significantly correlated (Spearman rank $r = 0.557$, $p = 2e-15$). The proportion of the 10 natural images with highest encoding model predicted activation and highest measured activation via fMRI had a somewhat weaker but still significant correlation (Spearman rank $r = 0.4545$, $p = 3.781e-10$ and Spearman rank $r = 0.1897$, $p = 0.0127$, respectively). These results again highlight the utility of NeuroGen as a discovery architecture that can concisely uncover, even across several categories of images at once, neural representation variability across individuals and brain regions within an individual.

3.4. *Optimized synthetic images have more extreme predicted co-activation of region-pairs than natural images*

The NeuroGen framework is flexible and can be used to synthesize images predicted to achieve an arbitrary target activation level for any or all of the 24 regions for which we have encoding models. We now provide examples of two-region and three-region optimization; specifically, we aimed to create synthetic images that jointly maximize and/or minimize predicted activation in the target regions together. For example, two region optimization could either be joint maximization (+ROI1 + ROI2), or joint maximization of one region and minimization of the other, or maximizing (+ROI1-ROI2 or -ROI1 + ROI2). Figure 8 shows the results of three example two region optimizations for subject 8's A) FFA1 (face) and V1v (early visual) regions, B) PPA (place) and V1v (early visual) regions, C) FFA1 (face) and PPA (place) regions, and one example of three region optimization (FFA1, PPA and V1v). When jointly maximizing FFA1 or PPA with V1v, NeuroGen's synthetic images are of places or faces with an abundance of texture (e.g. comic book, steel drum and kimono labels for FFA1 and toy shop, bookstore, slot labels for PPA). Alternatively, when jointly maximizing FFA1 or PPA and minimizing V1v, we see human/dog faces or places/indoor scenes with flat, non-textured colors in both the foreground and background (e.g. neck brace, ice lolly and bikini for FFA1 and beacon, pier, container ship labels for PPA). Natural images show a similar pattern, although they are not as obvious and appear less consistent (see Supplementary Figure S9). In addition, we see that the top 10 optimized synthetic images generally had significantly more extreme predicted activation values in the desired direction than the top 10 natural images for most cases (see Fig. 8 and Supplementary Figure S9). Both the two- and three- region optimizations demonstrate the synthetic images push the predicted activations past the boundaries of the natural image activations; there is a

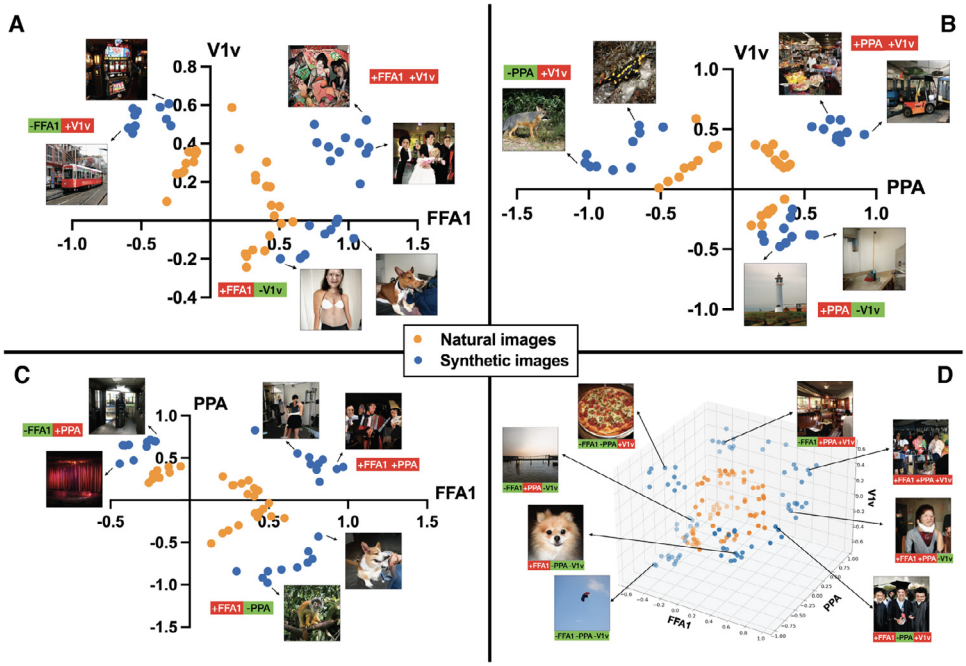


Fig. 8. Multiple-region optimization creates synthetic images with generally more extreme predicted activation when compared to natural images. Scatter plots show the predicted response from top 10 natural and synthetic images that either jointly maximize the sum of the two regions (+ROI1 + ROI2) or jointly maximize one and minimize the other (+ROI1-ROI2) and (-ROI1 + ROI2), for the following region-pairs **A** FFA1 and V1v, **B** PPA and V1v and **C** FFA1 and PPA. **D** shows all sets of three-region optimization combinations for FFA1, PPA and V1v together. Typical synthetic image examples are shown for each dual and triple optimization; the optimization specifics are listed below each synthetic image (green = that region's minimization, red = that region's maximization).

clear separation of the orange and blue points representing these values. These results demonstrate that the NeuroGen framework can be applied to create synthetic images predicted to achieve optimized activation in more than one region at the same time; the example synthetic images provided align with expectations from prior knowledge. Furthermore, they suggest that the benefit of NeuroGen's synthetic images over natural ones in terms of pushing the predicted activation levels to those not achievable by natural images.

4. Discussion

Here, we introduce NeuroGen, a novel architecture designed to synthesize realistic images predicted to maximize or minimize activation in pre-selected regions of the human visual cortex. NeuroGen leverages three recent scientific advances: 1) the development of encoding models that can accurately predict brain responses to visual stimuli (St-Yves and Naselaris, 2018a), 2) in deep generative networks' abilities to synthesize high-fidelity and variety images (Brock et al., 2018) and 3) in the recent curation of the Natural Scenes Dataset (NSD), which consists of tens of thousands of paired images and human brain responses (Allen, 2021). We showed that encoding models trained on the NSD data could accurately maps images to their neural representations in individual subjects, and, importantly, that the encoding model smooths noisy measured fMRI response maps. Once the encoding model was validated, we used the NeuroGen framework to concisely amplify and reveal individual and regional preferences for certain image types. We began by using NeuroGen to create synthetic images that were predicted to maximize a single region's activation response. The resulting synthetic images agreed with expectations from previous knowledge of regional neural representations of visual stimuli and, furthermore, the predicted activations from synthetic images were significantly higher than from natural images. Once NeuroGen was validated, it was used as a discovery architecture to uncover region-specific and individual-specific visual cortex response patterns. Our main discovery was a remarkable, previously not well-described balance of dog-human preferences in face areas that both varied across face regions and individuals. The synthetic image human/dog preference ratios were validated by showing strong, significant correlations with dog/human preference ratios calculated by contrasting measured fMRI activations in response to thousands of dog and human images. Secondly, we used the NeuroGen framework to show that the content of images preferentially activating word form areas were of a wide variety, including humans, dogs, birds, other animals, clocks, food and more. Despite the fact that several categories of images were represented in the word form areas, we again validated the top 10 synthetic image ratios by showing significant correlations with underlying preferences extracted by contrasting measured fMRI activations in response to hundreds of images. Finally, we extended the single region analysis to demonstrate the capacity of the NeuroGen framework in optimizing activation for two or three regions at a time. We found that these two- and three-region optimizations not only produced images that agreed with expectations, but also provided significantly more extreme predicted activations than natural images, above and beyond the activation levels observed in response to the best-matching natural images. Taken together, these results validate and demonstrate that the NeuroGen framework can create new hypotheses for neuroscience and thus facilitate a tight loop between modeling and experiments, and thus is a robust and flexible discovery architecture for vision neuroscience.

The visual system provides an excellent model with which to understand how organisms experience the environment. Mapping the visual system's neural representations of external stimuli has often centered around identifying features that maximally activate various neurons or populations of neurons (Hubel and Wiesel, 1962; 1968). This "activation maximization" approach, more commonly called the tuning curve approach, has lead to discoveries of visual regions that selectively respond to specific patterns (Kobatake and Tanaka, 1994; Wandell et al., 2007) or images with a certain content, most prominently, faces (Kanwisher et al., 1997; Tsao et al., 2006), places (Epstein and Kanwisher, 1998), bodies (Downing et al., 2001; Popivanov et al., 2014) and visual words (Baker, 2007). This "activation maximization" approach using *in vivo* measurements is by nature limited to the stimuli presented while observing responses, which is in turn biased by *a priori* hypotheses. There may be more complex, obscure stimuli-response maps that exist (for example, perhaps, to images of a dog riding a bicycle) but are not tested due to our limited imaginations or the lack of representation of that type of image in natural image sets. In addition, fMRI can be very noisy, and response maps to a handful of images (or even hundreds of them) are quite noisy even within an individual, let alone across the population. Encoding models that can perform "offline" mapping of stimuli to brain responses can provide a computational stand-in for a human brain that also smooths measurement noise in the stimuli-response maps. Neuro-

Gen's framework that couples the benefits of encoding models with recent advances in generative networks in creating naturalistic-looking images, may prove to be an advance in discovery neuroscience that is more than the sum of its parts.

Only a few previous studies have used generative networks to create synthetic images made to achieve activation maximization of single neurons or populations of neurons in non-human primates. Both used closed-loop physiological experimental designs to record and optimize neuronal responses, e.g. maximize firing rates, to synthetic images (Bashivan et al., 2019; Ponce, 2019). One synthesized images by directly optimizing in image space using an ANN model for the brain's ventral visual stream (Bashivan et al., 2019), and the other synthesized images in code space via a genetic algorithm to maximize neuronal firing in real time (Ponce, 2019). Both of these studies successfully demonstrated that single neurons or neuronal populations in monkeys can be controlled via optimization of synthetic images using generative networks. One difference in our framework, other than the species in question, is the use of a conditional generator network that requires the identification of an image class before synthesis. We wanted our framework to synthesize images that were as natural-looking as possible for two reasons: because our encoding model was trained on natural images and because future work will include presentation of these synthetic images to humans while they are undergoing fMRI to test if they achieve activation above and beyond the best natural images. One recently published work in humans used a similar approach to NeuroGen to synthesize images designed to achieve maximal activation in one of three late visual regions (FFA, EBA and PPA) to validate category selectivity of these regions (Ratan Murty et al., 2021). Our approach is different from theirs in many ways, most importantly that our focus is on individual-level and region-level differences in category preference (with which we make some interesting observations that reflect underlying data) and, further, our interest in creating synthetic images designed to achieve optimal activity (either maximized or minimized) in multiple regions at once.

Our main discovery using the NeuroGen architecture was a previously not well-described balance of dog-person preference in face areas, which varied over individuals within the population and regions within an individual. After inspecting the content of the top 10 images from NeuroGen, we noted an abundance of dog faces in addition to human faces that was not obvious in the top 10 natural images with the highest measured or predicted activation; this was also apparent from looking at the face regions' wordclouds. We showed that the dog-human preference ratio observed in NeuroGen's synthetic images was reflected in the underlying data by observing a strong, significant correlation with the t-statistic of the measured activation (via fMRI response) from dog images versus human images. One idiosyncrasy of the ImageNet data used to train our generative network is its prevalence of dogs; 120 out of the approximately 1000 ImageNet classes are dog breeds and, furthermore, dog images in ImageNet generally feature close-ups of dog's faces. This over-representation of dogs in the ImageNet database could have biased NeuroGen to more easily identify and amplify any existing dog-human preferences in the underlying data. In addition, measured contrasts for some regions in some individuals showed a clear preference for dog faces over human faces (t-statistic > 4), which we conjecture could be due to either differences in visual attention between the two categories or the fact that the NSD "person" images used to calculate the contrast are not all close-ups of human faces while the dog images do tend to be close up dog faces. As the face ROIs were derived based on the t values of activation contrast for face>non-face images, it may alter our dog v human findings when the threshold for t varies and/or when a more restrictive contrast is applied (e.g. face>objects). However, one can imagine if the ROI definition for face selectivity became less restrictive, you would see a drop in the number of human faces and an increase of non-face images. The fact that we are specifically observing dog images indicates that it is not an effect of reduced selectivity of the face ROI, unless that reduction happened to encompass regions that were dog-face selective. There have been a few previous works investigating humans' face-processing areas' responses to images of human versus animal faces (Downing et al., 2006; Whyte et al., 2016). One of the first studies showed that human face areas respond to mammals, although at a population level, the activation in response to mammals was not stronger than responses to humans (Downing et al., 2006). Another study found that face areas in adolescents with high functioning autism had a weaker response to unfamiliar human, but not animal, faces and greater activation in affective face regions in response to animal, but not human, faces compared to typically developing adolescents (Whyte et al., 2016). One of the few studies comparing humans' neural representations of dog faces and human faces showed very similar response maps to both species, with lingual/medial fusiform gyri being the only region having higher activations for dog over human faces (Blonder, 2004). We conjecture that differences in our findings may be due to their population-level approach to identifying differences in neural representations, as they used coregistered contrast maps to identify group-level, voxel-wise significance. We see that NeuroGen's dog-human balance in response patterns varies widely over individuals and brain regions, indicating that population-level approaches may not be adequate for creating stimulus-response maps.

While humans' neural representations of faces, places and bodies are generally robust across the population, it has been shown that word form responses can vary based on an individual's experience (Baker, 2007; Kanwisher, 2010). Our findings generally revealed more divergence in the word form area preferred content across individuals than other categories of visual regions. This large individual-level variability in preferred image content, including images of several very different categories (humans, dogs, cats, birds, food, clocks), could be due to the effect of individual experience in forming the neural representations in these word form areas. On the other hand, these areas do tend to be quite small and more susceptible to noise in the measured activation patterns which were used to define the regions leading to more population-level divergence (Brett et al., 2002). The natural images in the NSD dataset used to create the encoding model also did not contain isolated text, which could further contribute to noise in applying the NeuroGen framework to word form areas. However, many of the synthetic images were derived from categories that contain items with text, including "odometer", "comic book", "book jacket", "street sign", "scoreboard", "packet" and "pill bottle". The word form regions did also overlap regions in other categories (see Supplementary Figure S10-18), including face and place areas. This overlap could explain the presence of dogs and humans (and possibly other mammals) but it does not explain, for example, the strong presence of food images in many of the individuals' top images (see Fig. 5B). Despite these potential shortcomings, using only 10 images, NeuroGen was able to reflect measured, underlying preferences across several categories for these complex and widely varying word form regions.

One of the advantages of the NeuroGen framework is its flexibility and capacity - one can provide an arbitrary target response map containing desired activation levels for any (or all) of the 24 brain regions that have encoding models and produce synthetic images that achieve that vector as closely as possible. As a simple example, we performed joint optimization of two or three regions, where we maximized and/or minimized their activations together. We chose to use V1v as one of the regions as this is known to activate in response to high-frequency patterns and results could be readily validated visually. Indeed we do see that when maximizing V1v and face/place areas, we get faces/places with an abundance of texture and when minimizing V1v we get places/faces with flat features. From looking at the scatter plots representing the synthetic and natural images' predicted activations in Fig. 8, there is a clear separation of the two, where the synthetic images clearly push the predicted brain activations to levels not achievable by the best-matching natural images. This example application of NeuroGen highlights another advantage of this framework in that one could synthesize stimuli predicted to evoke response patterns not generally observed in response to natural images.

There are a few limitations in this work. First, the range of the synthetic images is constrained by the images on which the generative network is trained, in this case ImageNet. Any preferences that exist for image content or features in the encoding model that do not exist in the ImageNet database may remain obscured in NeuroGen. Second, the deep generative network has a parameter that controls the balance between fidelity and variety of the synthetic images produced. It could be that varying this parameter would provide more realistic images, but it may also result in images that do not have as extreme predicted activation and/or have less variety and thus contain less information about the underlying stimuli-response landscape. Third, the optimization of the synthetic images is done in two steps, by first selecting the top 10 image classes and then optimizing the noise vector in that image class space. The classes identified in the first step could be constraining the synthesizer so that it is not identifying a global optimum; however, this trade-off was deemed an acceptable sacrifice for the more natural-looking images provided by a conditional generator. Lastly, this study employed AlexNet but more recent studies have found that other recent state-of-the-art methods like ResNet (He et al., 2016) and VGG19 (Simonyan and Zisserman, 2014) can perform better in terms of neural predictivity. Exploring these architectures can also be useful in subsequent studies.

The NSD data on which the encoding model was trained is unsurpassed in its quality and quantity, consisting of densely-sampled fMRI in 8 individuals with several thousand image-response pairs per subject. Still, the natural images sourced from the COCO dataset used in the NSD experiments are inevitably limited in their content and features, which can mean possibly inaccurate brain-response mappings for images not used to train the encoding model. Additionally, when calculating preference ratios in the measured NSD data, it was at times difficult to choose the combination of image labels that produced the desired image content or features (e.g. only a person's face). Relatedly, it is not always straightforward to classify the natural or synthetic images into the appropriate category; the word form areas were particularly challenging. In addition, fMRI has many known sources of noise/confounds such as system-related instabilities, subject motion and possibly non-neuronal physiological effects from breathing and blood oxygenation patterns (Liu, 2016). Careful design of the acquisition and post-processing pipeline for the NSD data mitigated these effects. Finally, the localizer task, while previously validated, may have some variability due to the contrast threshold applied. Using a more or less liberal threshold for the region boundary definition may result in different results than what is presented here. Different visual regions used in this work did have some overlap within certain individuals, which could have contributed to similarities in synthetic image content for regions of different categories. Supplementary Figure S10-18 show each individual's regional definitions and a heatmap of the Dice overlap of regions from different categories for each individual.

To validate and demonstrate the capability of our novel NeuroGen framework, we present here as a proof-of-concept optimization of predicted responses in one, two or three regions. However, this optimization can be performed on an arbitrary desired activation pattern over any regions (or voxels) that have existing encoding models. Generative networks for creating synthetic images are an highly active area of research; specialized generators for faces or natural scenes could be integrated into the NeuroGen framework to further improve the range and fidelity of the synthetic images. Furthermore, the work presented here does not investigate the measured responses in humans to NeuroGen's synthetic images. However, we believe that the current paper introducing and validating the NeuroGen framework and demonstrating its use in a discovery neuroscience context represents an important technical and conceptual contribution to the field. The idea that this type of synthetic generator coupled with an encoding model can be used to make discoveries about regional or individual-level selectivity to stimuli is itself a novel conceptual contribution, and the NeuroGen framework is a novel technical contribution whose utility is demonstrated here with examples. Future work will involve presentation of these synthetic images to individuals while undergoing fMRI to test if their responses are indeed more extreme than the best natural images. One hypothesis is that the synthetic images may command more attention, as it is clear they are not perfectly natural and thus may produce a more extreme response than natural images, as found in studies of single or neuronal population responses (Bashivan et al., 2019; Ponce, 2019). The other is that there may be some confusion about what the image contains or additional processing that an individual will undergo when interpreting the image that will result in an unpredictable pattern of response. If it can be demonstrated that synthetic images indeed produce activations matching a pre-selected target pattern, the NeuroGen framework could be used to perform macro-scale neuronal population control in humans. Such a novel, noninvasive neuromodulatory tool would not only be powerful in the hands of neuroscientists, but could also open up possible avenues for therapeutic applications.

5. Conclusions

The NeuroGen framework presented here represents a robust and flexible framework that can synthesize images predicted to achieve a target pattern of regional activation responses in the human visual cortex that exceeds that of predicted responses to natural images. We posit that NeuroGen can be used for discovery neuroscience to uncover novel stimuli-response relationships. If it can be shown with future work that the synthetic images actually produce the desired target responses, this approach could be used to perform macro-scale, non-invasive neuronal population control in humans.

Data availability

The NSD dataset is publicly available at http://naturalscenesdataset.org.

Code availability

Code is available at https://github.com/zijin-gu/NeuroGen.

Ethics statement

All studies were approved by an ethical standards committee on human experimentation, and written informed consent was obtained from all patients.

Citation gender diversity statement

Recent work in several fields of science has identified a bias in citation practices such that papers from women and other minorities are under-cited relative to the number of such papers in the field (Dworkin, 2020). Here we sought to proactively consider choosing references that reflect the diversity of the field in thought, form of contribution, gender, and other factors. We obtained predicted gender of the first and last author of each reference by using databases that store the probability of a name being carried by a woman (Dworkin, 2020). By this measure (and excluding self-citations to the first and last authors of our current paper), our references contain 4% woman(first)/woman(last), 25% man/woman, 10% woman/man, and 61% man/man. This method is limited in that a) names, pronouns, and social media profiles used to construct the databases may not, in every case, be indicative of gender identity and b) it cannot account for intersex, non-binary, or transgender people. We look forward to future work that could help us to better understand how to support equitable practices in science.

Declaration of Competing Interest

The authors declare no competing interests.

Credit authorship contribution statement

Zijin Gu: Methodology, Software, Validation, Formal analysis, Writing – original draft, Writing – review & editing. **Keith Wakefield Jamison:** Conceptualization, Methodology, Writing – review & editing. **Meenakshi Khosla:** Formal analysis, Writing – review & editing. **Emily J. Allen:** Resources, Data curation, Writing – review & editing. **Yihan Wu:** Resources, Data curation, Writing – review & editing. **Thomas Naselaris:** Software, Writing – review & editing. **Kendrick Kay:** Resources, Data curation, Writing – review & editing. **Mert R. Sabuncu:** Conceptualization, Methodology, Supervision, Writing – review & editing. **Amy Kuceyeski:** Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing.

Acknowledgements

This work was funded by the following grants: R01 NS102646 (AK), RF1 MH123232 (AK), R21 NS104634 (AK), R01 LM012719 (MS), R01 AG053949 (MS), NSF CAREER 1748377 (MS), and NSF NeuroNex Grant 1707312 (MS). The NSD data were collected under the NSF CRCNS grants IIS-1822683 (KK) and IIS-1822929 (TN).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.neuroimage.2021.118812.

References

Allen, E.J., 2021. A massive 7T fMRI dataset to bridge cognitive and computational neuroscience. bioRxiv doi:10.1101/2021.02.22.432340. https://www.biorxiv.org/content/early/2021/02/22/2021.02.22.432340.full.pdf

Baker, C.I., 2007. Visual word processing and experiential origins of functional selectivity in human extrastriate cortex. Proc. Natl. Acad. Sci. U.S.A. 104, 9087–9092. doi:10.1073/pnas.0703300104.

Bashivan, P., Kar, K., DiCarlo, J.J., 2019. Neural population control via deep image synthesis. Science 364. doi:10.1126/science.aav9436. https://science.sciencemag.org/content/364/6439/eaav9436.full.pdf

Belagiannis, V., Rupprecht, C., Carneiro, G., Navab, N., 2015. Robust optimization for deep regression. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2830–2838.

Benson, N.C., Winawer, J., 2018. Bayesian analysis of retinotopic maps. Elife 7, e40224.

Blonder, L.X., 2004. Regional brain response to faces of humans and dogs. Cognit. Brain Res. 20, 384–394. doi:10.1016/j.cogbrainres.2004.03.020.

Brett, M., Johnsrude, I.S., Owen, A.M., 2002. The problem of functional localization in the human brain. Nat. Rev. Neurosci. 3, 243–249.

Brock, A., Donahue, J., Simonyan, K. et al., 2018. Large scale GAN training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096.

Cadena, S.A., 2019. Deep convolutional models improve predictions of macaque V1 responses to natural images. PLoS Comput. Biol. 15, e1006897.

Charest, I., Kriegeskorte, N., Kay, K.N., 2018. GLMdenoise improves multivariate pattern analysis of fMRI data. Neuroimage 183, 606–616.

Cichy, R.M., Khosla, A., Pantazis, D., Torralba, A., Oliva, A., 2016. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Sci. Rep. 6, 1–13. doi:10.1038/srep27755.

De Valois, R.L., De Valois, K.K., 1980. Spatial vision. Annu. Rev. Psychol. 31, 309–341.

DeAngelis, G.C., Ohzawa, I., Freeman, R.D., 1995. Receptive-field dynamics in the central visual pathways. Trends Neurosci. 18, 451–458.

Downing, P.E., Chan, A.-Y., Peelen, M., Dodds, C., Kanwisher, N., 2006. Domain specificity in visual cortex. Cereb. Cortex 16, 1453–1461.

Downing, P.E., Jiang, Y., Shuman, M., Kanwisher, N., 2001. A cortical area selective for visual processing of the human body. Science 293, 2470–2473.

Dworkin, J.D., 2020. The extent and drivers of gender imbalance in neuroscience reference lists. Nat. Neurosci. 23, 918–926.

Epstein, R., Kanwisher, N., 1998. A cortical representation of the local visual environment. Nature 392, 598–601.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Hubel, D.H., Wiesel, T.N., 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J. Physiol. 160, 106–154.

Hubel, D.H., Wiesel, T.N., 1968. Receptive fields and functional architecture of monkey striate cortex. J. Physiol. 195, 215–243.

Hubel, D.H., Wiesel, T.N., 2020. Receptive fields of single neurones in the cat's striate cortex. In: Brain Physiology and Psychology. University of California Press, pp. 129–150.

Huth, A.G., 2016. Decoding the semantic content of natural movies from human brain activity. Front. Syst. Neurosci. 10, 81. doi:10.3389/fnsys.2016.00081.

Kanwisher, N., 2010. Functional specificity in the human brain: a window into the functional architecture of the mind. Proc. Natl. Acad. Sci. U.S.A. 107, 11163–11170. doi:10.1073/pnas.1005062107.

Kanwisher, N., McDermott, J., Chun, M.M., 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception. J. Neurosci. 17, 4302–4311.

Kay, K., Rokem, A., Winawer, J., Dougherty, R., Wandell, B., 2013. GLMdenoise: a fast, automated technique for denoising task-based fMRI data. Front. Neurosci. 7, 247.

Khaligh-Razavi, S.-M., Kriegeskorte, N., 2014. Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput. Biol. 10, e1003915. doi:10.1371/journal.pcbi.1003915.

Khosla, M., Ngo, G.H., Jamison, K., Kuceyeski, A., Sabuncu, M.R., 2020. Cortical response to naturalistic stimuli is largely predictable with deep neural networks. bioRxiv.

Kobatake, E., Tanaka, K., 1994. Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. J. Neurophysiol. 71, 856–867.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 25, 1097–1105.

Lin, T. Y. et al., 2014. Microsoft COCO: common objects in context. Springer. European Conference on Computer Vision, 740–755

Liu, T.T., 2016. Noise contributions to the fMRI signal: an overview. Neuroimage 143, 141–151. doi:10.1016/j.neuroimage.2016.09.008.

Movshon, J.A., Thompson, I., Tolhurst, D.J., 1978. Spatial and temporal contrast sensitivity of neurones in areas 17 and 18 of the cat's visual cortex. J. Physiol. 283, 101–120.

Mozafari, M., Reddy, L., VanRullen, R., 2020. Reconstructing natural scenes from fMRI patterns using BigBIGAN. IEEE. 2020 International Joint Conference on Neural Networks (IJCNN), 1–8

Nandy, A.S., Sharpee, T.O., Reynolds, J.H., Mitchell, J.F., 2013. The fine structure of shape tuning in area V4. Neuron 78, 1102–1115.

Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., Clune, J., 2016. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. arXiv preprint arXiv:1605.09304.

Ponce, C.R., 2019. Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. Cell 177, 999–1009.e10. doi:10.1016/j.cell.2019.04.005.

Popivanov, I.D., Jastorff, J., Vanduffel, W., Vogels, R., 2014. Heterogeneous single-unit selectivity in an fMRI-defined body-selective patch. J. Neurosci. 34, 95–111. doi:10.1523/JNEUROSCI.2748-13.2014.

Ratan Murty, N.A., Bashivan, P., Abate, A., DiCarlo, J.J., Kanwisher, N., 2021. Computational models of category-selective brain regions enable high-throughput tests of selectivity. Nat. Commun. 12, 1–14.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: towards real-time object detection with region proposal networks. arXiv preprint arXiv:1506.01497.

Rokem, A., Kay, K., 2020. Fractional ridge regression: a fast, interpretable reparameterization of ridge regression. Gigascience 9, giaa133.

Rosenblatt, F., 1958. The perceptron: a probabilistic model for information storage and organization in the brain. Psychol. Rev. 65, 386.

Russakovsky, O., 2015. ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. 115, 211–252.

Sermanet, P. et al., 2013. OverFeat: integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229.

Seymour, K.J., Stein, T., Clifford, C.W., Sterzer, P., 2018. Cortical suppression in human primary visual cortex predicts individual differences in illusory tilt perception. J. Vis. 18, 1–10. doi:10.1167/18.11.3.

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

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

St-Yves, G., Naselaris, T., 2018. The feature-weighted receptive field: an interpretable encoding model for complex feature spaces. Neuroimage 180, 188–202.

St-Yves, G., Naselaris, T., 2018b. Generative adversarial networks conditioned on brain activity reconstruct seen images. IEEE. 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 1054–1061

Stigliani, A., Weiner, K.S., Grill-Spector, K., 2015. Temporal processing capacity in high-level visual cortex is domain specific. J. Neurosci. 35, 12412–12424. doi:10.1523/JNEUROSCI.4822-14.2015, https://www.jneurosci.org/content/35/36/12412.full.pdf.

Thorpe, S., Fize, D., Marlot, C., 1996. Speed of processing in the human visual system. Nature 381, 520–522.

Toshev, A., Szegedy, C., 2014. DeepPose: human pose estimation via deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1653–1660.

Tsao, D.Y., Freiwald, W.A., Tootell, R.B., Livingstone, M.S., 2006. A cortical region consisting entirely of face-selective cells. Science 311, 670–674. doi:10.1126/science.1119983.

Van Essen, D.C., 2013. The WU-Minn human connectome project: an overview. Neuroimage 80, 62–79.

Van Essen, D.C., Anderson, C.H., Felleman, D.J., 1992. Information processing in the primate visual system: an integrated systems perspective. Science 255, 419–423.

VanRullen, R., Reddy, L., 2019. Reconstructing faces from fMRI patterns using deep generative neural networks. Commun. Biol. 2, 1–10.

Wandell, B.A., Dumoulin, S.O., Brewer, A.A., 2007. Visual field maps in human cortex. Neuron 56, 366–383. doi:10.1016/j.neuron.2007.10.012.

Whyte, E.M., Behrmann, M., Minshew, N.J., Garcia, N.V., Scherf, K.S., 2016. Animal, but not human, faces engage the distributed face network in adolescents with autism. Dev. Sci. 19, 306–317.

Yamins, D.L., 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc. Natl. Acad. Sci. U.S.A. 111, 8619–8624. doi:10.1073/pnas.1403112111.

Zhang, H., Goodfellow, I., Metaxas, D., Odena, A., 2019. Self-attention generative adversarial networks. PMLR. International Conference on Machine Learning, 7354–7363

Ziemba, C.M., Freeman, J., Movshon, J.A., Simoncelli, E.P., 2016. Selectivity and tolerance for visual texture in macaque V2. Proc. Natl. Acad. Sci. 113, E3140–E3149.