Quality assessment of anatomical MRI images from Generative Adversarial Networks: human assessment and image quality metrics

Matthias S. Treder[1], Ryan Codrai[1], and Kamen A. Tsvetanov[2],[3]

[1] *School of Computer Science and Informatics, Cardiff University, Cardiff CF24 3AA, UK*

[2] *Department of Clinical Neurosciences, University of Cambridge, CB2 0SZ, UK*

[3] *Department of Psychology, University of Cambridge, Cambridge, CB2 3EB, UK*

Abstract

Background: Generative Adversarial Networks (GANs) can synthesize brain images from image or noise input. So far, the gold standard for assessing the quality of the generated images has been human expert ratings. However, due to limitations of human assessment in terms of cost, scalability, and the limited sensitivity of the human eye to more subtle statistical relationships, a more automated approach towards evaluating GANs is required.

New method: We investigated to what extent visual quality can be assessed using image quality metrics and we used group analysis and spatial independent components analysis to verify that the GAN reproduces multivariate statistical relationships found in real data. Reference human data was obtained by recruiting neuroimaging experts to assess real Magnetic Resonance (MR) images and images generated by a GAN. Image quality was manipulated by exporting images at different stages of GAN training.

Results: Experts were sensitive to changes in image quality as evidenced by ratings and reaction times, and the generated images reproduced group effects (age, gender) and spatial correlations moderately well. We also surveyed a number of image quality metrics. Overall, Fréchet Inception Distance (FID), Maximum Mean Discrepancy (MMD) and Naturalness Image Quality Evaluator (NIQE) showed sensitivity to image quality and good correspondence with the human data, especially for lower-quality images (i.e. images from early stages of GAN training). However, only a Deep Quality Assessment (QA) model trained on human ratings was able to reproduce the subtle

differences between higher-quality images.
Conclusions: We recommend a combination of group analyses, spatial correlation analyses, and both distortion metrics (FID, MMD, NIQE) and perceptual models (Deep QA) for a comprehensive evaluation and comparison of brain images produced by GANs.

Keywords: Generative Adversarial Network, GAN, MRI, machine learning, generative models, ageing, quality assessment, deep learning

1. Introduction

Magnetic resonance imaging (MRI) provided a means to understanding the structural and functional heterogeneity of the human brain in health and disease. The recent surge in computing horsepower together with large international collaborative initiatives advanced neuroimaging into a big data science, opening the field to deep learning with a promise for new discoveries [1]. One of the most successful deep learning models has been the Convolutional Neural Network (CNN). Using brain images (e.g. T1-weighted or grey-matter density maps) as input, it has been used in brain age regression [2], brain tumor classification [3], tumor segmentation [4], and a plethora of other applications such as image enhancement, image modality translation, and data augmentation (see [5] for a review).

Many of the latter applications build on generative models which take as input either an image (*image-to-image* approach) or a random vector (*noise-to-image*) and produce an artificially generated brain image as output. The two main architectures for generative models are Generative Adversarial Networks (GANs; [6]) and Variational Autoencoders (VAEs; [7]). GANs have been shown to produce more crisp images than VAEs in diffusion-weighted [8] and T1-weighted images [9] and will be our focus in the rest of the paper. In GANs, the generator creates brain images as outputs using transposed convolution layers, the discriminator is a CNN that tries to classify images as generated or real. Both networks act as adversaries towards each other, with the generator using feedback from the discriminator to create increasingly realistic brain images. In the *image-to-image* approach, the generator receives an input image, typically from a different modality or lower resolution, and produces an output image in a different modality or higher resolution. For instance, GANs have been used to translate T1-weighted images into T2-weighted images [10], Computed Tomography [11], functional

MRI [12], and diffusion weighted images [13]. Furthermore, [14, 15] recovered high-resolution images from images reduced in k-space. A potential limitation of the image-to-image approach is that it requires input images from two imaging domains. In the *noise-to-image* approach, realistic MR images are synthesized *de novo*, starting from a random noise vector [16]. Their success rests on the fact that MRIs form a low-dimensional manifold and the generator acts as a forward model that maps from this manifold to image space. The primary application of noise-to-image GANs has been data augmentation. Several authors successfully used noise-to-image GANs to create realistic T1-weighted 2D image slices [17, 18] or 3D brain images [19, 9, 20]. The noise-to-image approach has also been used in tandem with the image-to-image approach in order to improve tumor detection [21]. *Noise-to-image* problems are arguably harder than *image-to-image* problems, since all anatomical features have to be learned from scratch, including the position and shape of the brain volume and its internal 3D structures.

While much effort has gone into improving generative models, no comprehensive framework for assessing the image quality and biological plausibility of the generated images has emerged yet. The current gold standard for judging visual quality is assessments by neuroimaging experts [8, 22]. In [17] two neuroimaging experts scored images on a 5-points Likert scale with real images scoring higher than generated ones. Several studies used a binary detection task wherein the experts were presented either real vs generated images [22, 8, 18]. Results indicate that generated MRIs are able to mislead experts into believing they are real to a high extent, albeit not perfectly so. Participants reported abnormalities in the shape of landmarks, changes in image contrast, or unusually high symmetry between the two hemispheres as giving away whether or not an image was real.

However, relying on human assessment alone has significant shortcomings. Recruiting human experts is costly, time consuming, and the number of images that can be rated is limited. To avoid this bottleneck and enable fast development cycles, reliable *image quality metrics* are required that can serve as a proxy for human assessment. They can be readily applied to datasets of any size. In image-to-image applications output images can be compared against ground truth reference images using a simple distance metric (see [16] for an overview). A popular metric is L2 loss and quantities derived from it such as mean squared error and peak signal-to-noise ratio (PSNR) have been used for quality assessment [10, 23, 24, 25], as well as Structural Similarity Index Measure (SSIM) [26] and kernel density es-

timates [22]. A different approach is to use metrics based on intermediate layers of deep learning models. A widely adopted metric in the GAN literature is the Inception Score (IS). However, in [22] IS did not agree well with human perception, since generated images yielded a higher score than real ones.

Additionally, while human assessment and image quality metrics can quantify the extent to which brain images 'look' realistic, we believe that they provide limited information on their *biological plausibility*. Human observers are sensitive to image artifacts such as checkerboard patterns but other distortions of brain images involve subtle statistical relationships across images that may not be visible to the human eye. For instance, the discovery of structural networks such as the Default Mode Network requires multivariate statistical analyses such as spatial Independent Component Analysis (ICA) performed across dozens or hundreds of MRIs [27]. Furthermore, establishing group differences (e.g. young vs old, male vs female) requires group statistical analysis. It is not clear whether the GAN learns to reproduce such group differences and large-scale structural networks because it is not explicitly trained to do so. We therefore believe that biological plausibility should be investigated as an additional dimension of quality assessment using a combination of group analysis and spatial ICA. To summarize, the goal of this study was to provide ingredients for a systematic and efficient evaluation of generated brain images. Our contributions are the following:

1. Behavioral experiment. We conducted the hitherto largest behavioral study on generated MRIs with 26 neuroimaging experts assessing real and generated grey-matter (GM) density maps from a noise-to-image Wasserstein GAN. Participants performed a detection task (real vs generated images) and a subjective quality rating task using a 5-points Likert scale. To determine the experts' sensitivity to objective changes in image quality, we exported images at five different stages (iterations) of GAN training, from early stages (where image quality was supposed to be poor) to later stages. We hypothesized that the proportion of images labeled as 'real' increases with training iteration in the detection task. Analogously, we hypothesized that quality ratings increase with iteration in the rating task. Ultimately both quantities were expected to approach the responses participants gave for real images.

2. Biological plausibility. For the first time, we performed an analysis of the structural properties of the generated 3D MRIs by performing spatial Independent Component Analysis (ICA) to investigate structural networks and group comparisons that show that GANs are sensitive to gender and age

differences. We hypothesized that structural networks and group differences show a high degree of correspondence between real and generated images.

3. Image quality metrics. We performed a comparison of image quality metrics, ranging from metrics popular in the GAN literature (e.g. Inception Score, Fréchet Inception Distance) to metrics used in image quality assessment, some of which have not been used in the context of brain images before. Additionally, we trained a Deep Quality Assessment (QA) model on the human data with the goal to provide an automated metric that mimicks human assessment. Since our ultimate goal was to identify a metric that can serve as proxy for human perception, all metrics were tested for their consistency with the behavioral data. We hypothesized that the Deep QA model would outperform standard image quality metrics, especially for high-quality generated images at the end of training.

2. Method

The Method section is split into four parts. In Section 2.1, we introduce our generative modeling approach. In Section 2.2, we introduce the behavioral experiment. In Section 2.3 we introduce spatial ICA and group analysis performed in order to investigate the images' biological plausibility. In Section 2.4, we introduce various image quality metrics and a Deep QA as an objective way of measuring the visual quality of the generated images, as well as control analyses using a StyleGAN [28]. For brevity, we refer to any type of brain image derived from an MR sequence (e.g. T1, T2, grey-matter density maps derived from T1) as *MRI* in the rest of the paper.

2.1. Generative modeling of MRIs

2.1.1. Data

Models were trained using the Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data [29, 30]. A T1-weighted 3D-structural MRI was acquired with the following parameters: repetition time (TR) = 2,250 ms; echo time (TE) = 2.99 ms; inversion time (TI) = 900 ms; flip angle $\alpha = 9°$; field of view (FOV) = $256 \times 240 \times 192$ mm^3; resolution = 1 mm isotropic; accelerated factor = 2; acquisition time, 4 min and 32 s [30]. The T1 image was initially coregistered to the MNI template, and the T2 image was then coregistered to the T1 image using a rigid-body linear transformation. The coregistered T1 and T2 images were used in a multichannel segmentation to extract probabilistic maps of six tissue classes: gray matter (GM), white

matter (WM), cerebrospinal fluid (CSF), bone, soft tissue, and residual noise. The native space GM and WM images were submitted to diffeomorphic registration (DARTEL; [31]) to create group template images. Each template was normalized to the MNI template using a 12-parameter affine transformation. After applying the combined normalization parameters (native to group template and group template to MNI template) to each individual participant's GM images, the normalized images were smoothed using an 8 mm Gaussian kernel. In total, GM images from 653 participants, aged 18-88 years, were considered as input to noise-to-brain GAN.

It is worth noting that GM maps were chosen over more common modalities such as T1 since our main focus was on the ability of GANs to generate brain structures. In contrast, non-segmented anatomical MR images also involve signals from non-neural structures such as skull, face, and ears. Consequently, it would be impossible to unambiguously attribute human rating and image quality results to brain or non-brain structures. Moreover, spatial brain networks (see Section 2.3) are typically defined using using GM-based morphometry images rather than non-segmented T1 images [32].

Before being fed into the GAN, the following additional preprocessing steps were performed. In the first step, we applied a cuboidal crop on the MRIs thereby removing the empty lateral space that surrounds the brain. Due to the four up-scaling layers in the generator that each up-scale by a factor of 2, the generator was restricted to producing images whose length of any given dimension is a multiple of 2^4. To bring the size of the real images in line with the generated ones, we resized the cropped MRIs to a final resolution of $96 \times 112 \times 96$. The resizing was necessary in order to fit the MRIs in GPU memory, since the intermittent tensors in the GAN were 5D (input channels x output channels x spatial dimensions) and could consume several gigabytes of video RAM each.

2.1.2. Generative Adversarial Network

In a Generative Adversarial Network (GAN) two neural networks are pitted against each other in a zero-sum game [6]. In this framework, the generator G acts as the adversary of the discriminator D. Whereas G aims to learn the generative distribution of the training data p_{real} by approximating it with a generative distribution p_{gen}, D seeks to accurately predict the probability that an image is generated rather than real. The discriminator is a function $D : \mathbb{R}^{96\times112\times96} \to [0, 1],\ x \mapsto D(x)$ that takes an image as input and returns the probability that this image is real. The generator is a func-

tion $G : \mathbb{R}^{100} \to \mathbb{R}^{96\times112\times96},\ z \mapsto G(z)$ that takes a random vector z sampled from a probability distribution p_z over $\mathbb{R}^{100}$ and returns an MRI. Training is designed to maximize the quality of the generated images by learning the generative distribution p_{real}. The model training reaches optimality when D is unable to discriminate between real and generated samples. The objective of the discriminator is governed by the equation

$$\max_D V(D) = \underbrace{\mathbb{E}_{x\sim p_{\text{real}}}[\log D(x)]}_{\text{Ability to label real samples}} + \underbrace{\mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))]}_{\text{Ability to label generated samples}}. \quad (1)$$

D seeks to maximize the probability for samples x from the training data and minimize the probability for samples generated from a random vector z. This formulation encourages the discriminator to become sensitive to image features that tell the difference between real and generated images. The generator is governed by the equation

$$\min_G V(G) = \underbrace{\mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))]}_{\text{Ability to generate quality images}}. \quad (2)$$

G seeks to maximize the probability that its generated images are labeled as 'real' by D. Both objectives can be combined into a joint loss function that is used for GAN training and optimized using alternating gradient descent

$$\min_G \max_D V(D, G) = \mathbb{E}_{x\sim p_{\text{data}}}[\log D(x)] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))]. \quad (3)$$

Since G learns a mapping from a low dimensional vector space to a high dimensional output, the space from which the input vector originates is a representation of the low-dimensional manifold of MRIs. This quantity is a lower bound of the Jensen-Shannon divergence between the real and generated distributions. GAN training using this objective function can fail to converge due to near-zero gradients when the true and generated distribution do not overlap and also. A more stable alternative that provides useful gradients for non-overlapping distributions is given by Wasserstein GANs (WGANs) [33, 34].To this end, the objective function is modified to the Earth Mover or Wasserstein-1 distance between p_{real} and p_{gen} given by

$$W(p_{\text{real}}, p_{\text{gen}}) = \inf_{\gamma\in\Pi(p_{\text{real}},p_{\text{gen}})} \mathbb{E}_{(x,y)\sim\gamma}[||x - y||] \quad (4)$$

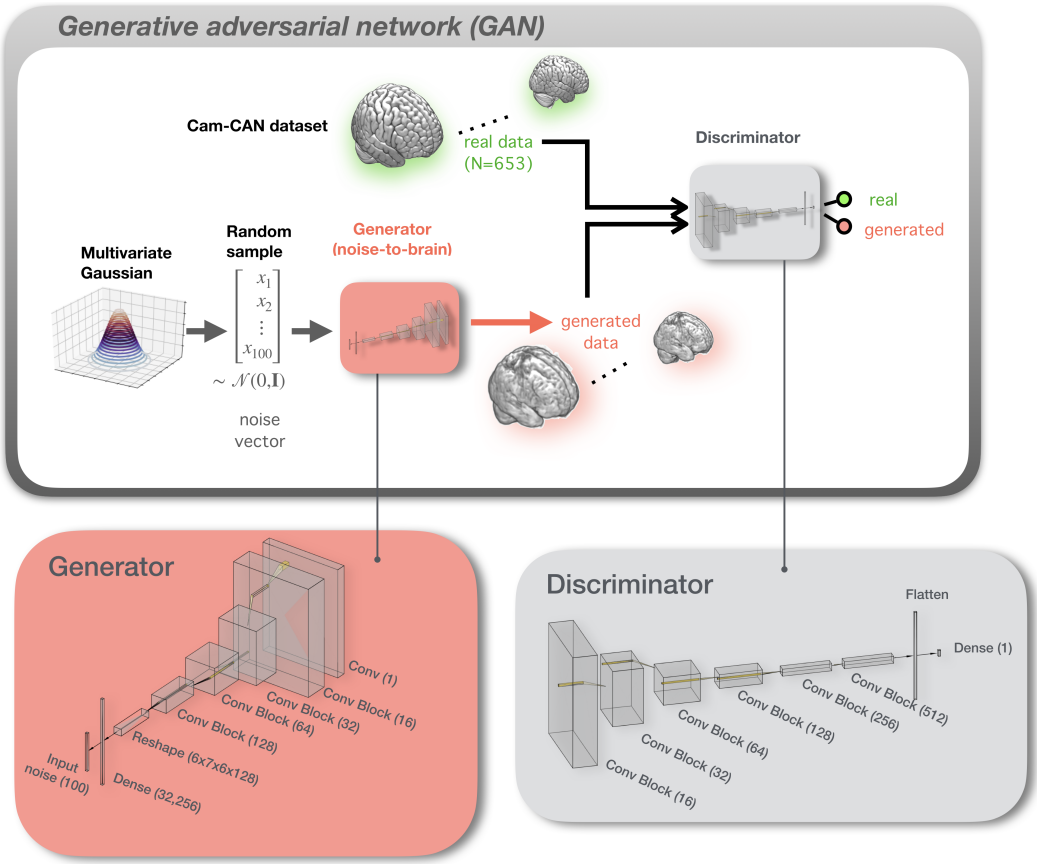

Figure 1: Structure of the noise-to-brain GAN. For the generator and discriminator, the number of filters is given in brackets, whereas for a dense layer, the number of units is shown in brackets.

where $\Pi(p_{\text{real}}, p_{\text{gen}})$ is the set of joint distributions whose marginals are given by p_{real} and p_{gen}. Practical implementations of this loss function are given in [33, 34].

2.1.3. Generating MR images using GANs

The structure of the noise-to-brain is shown in Figure 1. Note that the output of each convolutional layer is 4D (3 spatial dimensions and 1 output channels dimension). For displaying purposes, it is shown as 3D in the figure.

2.2. Behavioral experiment

2.2.1. Participants

Forty-five participants volunteered in the online study. Out of these, 26 participants (10 female, 16 male) completed all parts of the study (questionnaire, detection task, and subjective rating task). Only complete datasets were used for further analysis. Participants were aged 18 to 62 years ($\mu = 33.12, \sigma = 8.48$), with the majority of participants (16) in their 30s. All participants were neuroimaging experts (e.g. neuroscientists, neurologists, radiologists) at various levels of seniority (from undergraduate student to professor). They did not receive remuneration for their participation. Due to COVID-19 related restrictions (most of the data was collected during UK lockdowns) and the difficulty of recruiting a sufficient number of MR experts locally, the study was conducted online. Participants were mainly recruited through the authors' research networks, although no contact details were recorded. Participants gave written informed consent using an online form. Ethical approval was obtained from the COMSC Research Ethics Group at Cardiff University (COMSC/Ethics/2020/034).

2.2.2. Behavioral experiment

As stimuli, we used middle slides in the sagittal, coronal, and transverse plane extracted from the 3D images. These slices were horizontally concatenated to 432 x 288 pixels images. An image was either based on the real MRI data or generated by the GAN. To sample the generated images at different stages of GAN training, 'fake' images were drawn from five different iterations: 344, 1055, 7954, 24440, and 60000 (final iteration).

The experiment was conducted online using PsyToolkit [35, 36] (see Figure S1, Supplementary Material, for a visual depiction). Participants carried it out on their own computer by following a weblink. After filling out a brief questionnaire to indicate their expertise and familiarity with MRI data, the experiment started in fullscreen mode. It took about 15 minutes and was split into three phases: In the training phase, participants performed a practice detection task to familiarize themselves with the stimuli. The next two phases were the detection task and the subjective rating task. The order of these two phases was randomized across participants. Each phase was preceded by a 3 s countdown.

Training consisted of 30 randomized trials. In each trial, a real or generated ('fake') MR image was presented at the center of the screen. Participants were instructed to quickly and accurately indicate their choice with

the arrow keys on their keyboard, using the left index finger for the left arrow key ('real') and the middle finger for the right arrow key ('fake'). The presentation time of the image was unlimited, although for practical reasons the trial ended after 20 s and it was marked as timed out. After the trial, feedback was presented ('correct', 'wrong', 'time out'). Trials were separated by an inter-trial interval of 400-600 ms. Button press and reaction time were recorded. The proportion of stimuli was 1/3 real (10 images) and 2/3 fake (20 images) with equal proportions for each of the five fake levels (4 images each).

The detection task was structured just like the training task but no feedback was presented. There were 240 trials split into three blocks of 80 trials. The proportion of stimuli was 1/3 real (80 images) and 2/3 fake (160 images) with equal proportions for each of the five fake levels (32 images each). Participants were asked to take a break after every block. The experiment resumed upon key press.

The subjective rating task consisted of 30 trials comprising 5 real images plus 5 images from each iteration. Participants were rated their visual quality by selecting an item from a 5-points Likert scale with the options "very real", "relatively real", "neutral", "relatively fake", and "very fake". The options were presented below the image and chosen via a mouse click. Presentation time was unlimited although for practical reasons the trial ended after at most 10 s. Across the experiment, all images were unique. Once it was finished, participants had the option to submit a comment in a text box.

2.2.3. Behavioral data processing

For the detection task, the proportion of 'real' responses, proportion of correct responses, and reaction time (RT) were investigated. Trials with timeouts or <150 ms RT were discarded. One participant was removed due to a large amount of timed out trials (>10%). For every participant, mean 'real' responses and correct responses were calculated by averaging across all images corresponding to a given iteration. For the RT analysis, mean RTs were calculated using a trimmed mean approach wherein the 5% trials with the largest RTs were excluded and mean RT was calculated across the remaining trials [37]. For the subjective rating task, Mean Opinion Scores (MOS) were calculated by averaging across all images in an iteration. RTs were processed in the same way as for the detection task.

2.3. Group analysis and spatial Independent Components Analysis

To test whether the GAN generates biologically plausible data we compared the similarity between generated and real data using data-driven and model-based analyses. For the model-based approach, we sought to reproduce subtle differences between groups wherein we trained separate GANs on different subsets of the data. Training separate models for different groups is an approach that has been explored before [38]. To investigate age effects, we split the data into elderly ($>$ 70 years; 172 MRIs) and young 287 ($<$ 40 years, 185 MRIs). To test for effects of sex, we contrasted male (323 MRIs) and female (330 MRIs) MRIs. Once trained, we randomly generated as many MRIs as were in each of the subsets. The hypothesis was that if the GAN is sensitive to group characteristics, the generated MRIs should display similar age and gender effects as the real MRIs. For the data-driven approach, we used Independent Component Analysis (ICA) to decompose the synthetic data in a small set of spatially independent maps that correspond sensibly with known neurobiological relationships in the real data [32]. In particular, spatial ICA was implemented on real and generated GM maps separately. For each dataset, data were decomposed to a small number of spatially independent sources using the Source-Based Morphometry toolbox [39] in the Group ICA for fMRI Toolbox (GIFT)[1]. By combining PCA and ICA, one can decompose a participants-by-voxels into a source matrix that maps independent components (ICs) to voxels (here referred to as "IC maps"), and a mixing matrix that maps ICs to participants. The voxels that carried similar information across participants would have higher values and group to a set of regions. The spatial IC maps were then converted to z-scores. The correspondence between real and generated MRIs was based on the spatial correlation between thresholded IC maps, where the threshold was set to z-value $>$ 3. To further explore that the results were independent of the selected number of components, we repeated the procedure for a range of different components (5, 10, 15, 20, 25, 30).

2.4. Image Quality Metrics and Models

2.4.1. Image quality metrics

In pursuit of an automated metric that can serve as a proxy for human assessment, we surveyed a range of image metrics. For the sake of compara-

[1] http://mialab.mrn.org/software/gift

bility we applied these metrics to the same 2D images that were used in the behavioral experiment.

We first calculated four metrics widely adopted in the GAN literature, namely Inception Score (IS), Modified Inception Score (MIS), Fréchet Inception Distance (FID), Maximum Mean Discrepancy (MMD) [40]. Additionally, we tested two reference-free image quality metrics namely Natural Image Quality Evaluator (NIQE) and Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [41]. Other popular image quality metrics such as Peak signal to noise ratio (PSNR), Structural Similarity Index (SSIM), and Video Multimethod Assessment Fusion (VMAF), require reference images and thus do not strictly apply in this setting. However, since they are relevant to image-to-image GANs, they are investigated in the Supplementary Material. In the following, each of these metrics is introduced in more detail. As before, p_{real} and p_{gen} are used to denote the distributions of the real and generated images. Furthermore, $p(y)$ denotes the distribution of the class labels and $p(x|y)$ the distribution of images in class y.

- *IS* is a popular GAN metric using a CNN pre-trained on ImageNet [42]. We used the InceptionV3 model trained with 1000 classes. Its final layer is a softmax layer that represents the conditional distribution $p(y|x)$ of class labels y given an input image x. IS is then given by the formula

$$\text{IS} = \exp \mathbb{E}_{x \sim p_{\text{gen}}}(\mathbb{KL}(p(y|x) \,||\, p(y)))$$

  where $\mathbb{E}$ is the expectation, $x \sim p_{\text{gen}}$ are the images sampled from the generator, $\mathbb{KL}$ refers to the Kullback-Leibler divergence, and $p(y)$ is the marginal distribution of class labels. IS favours generated images that show a clear class membership, characterized by conditional class labels $p(y|x)$ with a 'peaky' (low entropy) distribution, at the same time favouring a broad coverage of multiple classes characterized by a 'flat' (high entropy) marginal distribution $p(y)$. A detailed discussion of IS is provided in [43].

- *MIS*, unlike IS, takes into account the desire for diversity of images within a given class. To achieve this, [44] incorporated a cross-entropy term $-p(y|x_i) \log p(y|x_j)$ into IS that represents within-class diversity yielding the modified Inception Score

$$\text{MIS} = \exp \mathbb{E}_{x_i}(\mathbb{E}_{x_j}(\mathbb{KL}(p(y|x_i) \,||\, p(y|x_j))).$$

- *FID* uses the feature embeddings of a CNN wherein it models the distributions of real and generated images as multivariate Gaussians [45]. FID is then given as the Fréchet distance between these two Gaussians,

$$\text{FID} = ||\mu_{\text{real}} - \mu_{\text{gen}}||_2^2 + \text{trace}(\Sigma_{\text{real}} + \Sigma_{\text{gen}} - 2(\Sigma_{\text{real}}\Sigma_{\text{gen}})^{\frac{1}{2}})$$

  where μ_{real} and μ_{gen} are the means for real and generated images, and Σ_{real} and Σ_{gen} are their respective covariance matrices. We calculated FID using InceptionV3's penultimate layer. The activation maps were condensed into 2,048 features by applying global average pooling.

- *MMD* measures the distance between two distributions, operationalized as the distance between its mean embeddings in feature space. Gretton et al. [46] implemented MMD using kernels, yielding

$$\text{MMD} = \mathbb{E}_{x,x' \sim p_r}[k(x, x')] + \mathbb{E}_{y,y' \sim p_g}[k(y, y')] - 2\mathbb{E}_{x \sim p_r, y \sim p_g}[k(x, y)]$$

  where k is a kernel function. We calculated MMD by again using the penultimate layer of InceptionV3 and the Radial Basis Function (RBF) kernel $k(x, x') = \exp(-\frac{1}{2}||x - x'||^2)$.

- *NIQE* computes image statistics based on normalized luminance values. A multivariate Gaussian distribution is fit to the image statistics and its Mahalanobis distance to a fit obtained on a corpus of natural images is calculated

$$\text{NIQE}(\nu_{\text{real}}, \nu_{\text{gen}}) = \sqrt{(\nu_{\text{real}} - \nu_{\text{gen}})^\top \left(\frac{\Sigma_{\text{real}} + \Sigma_{\text{gen}}}{2}\right)^{-1} (\nu_{\text{real}} - \nu_{\text{gen}})}$$

  where ν_{real} and Σ_{real} are the mean feature vector and covariance matrix obtained on the corpus and ν_{gen} and Σ_{gen} are the corresponding quantities computed on the image that is being assessed. We also used a second version of NIQE referred to as NIQE-MRI wherein the model was pretrained on an independent set of MRIs. Matlab's `niqe` and `fitniqe` functions were used.

- *BRISQUE* is a metric that, in contrast to NIQE, requires reference images with distortions as well as MOS to train a SVR model on the image statistics [47]. Just as NIQE, we also pretrained a second version

of the model on MRIs. Since BRISQUE requires MOS it was trained on the subjective rating data and cross-validation was used to obtain unbiased scores for the images. Matlab's `brisque` and `fitbrisque` functions were used.

IS, MIS, NIQE and BRISQUE are reference free metrics that provide scores for both generated and real images. For the other metrics, we used generated images from each of the iterations as target images and the real MRIs as reference images. FID and MMD only require a distribution of generated images. Another distinction between the metrics is that IS and MIS are quality metrics (better quality = higher value) whereas FID, MMD, NIQE and BRISQUE are distance metrics (better quality = lower value).

2.4.2. Deep QA model

An alternative to image-based metrics is to train a model that tries to mimick human perceptual assessments. Such perceptual models have been used both for MRIs and in the wider image/video quality literature [48, 49]. We implemented two CNNs that essentially performed the same task as the human experts, namely classifying images as 'real' or 'fake' (detection task) and assigning a subjective rating to an image on a 5-points Likert scale (rating task). The model architecture is depicted in Figure 6a.

Detection task model. The first part of the model consisted of a InceptionV3 model pretrained on ImageNet with the classification layer removed. The InceptionV3 output was fed into two dense layers (32 and 16 units, dropout rate 0.1, leaky ReLU activation with $\alpha = 0.1$), denoted as MRI features. The final layer was a single sigmoid unit. The model was trained on a set of 3611 real and generated images not used in the behavioral experiment. Images from all iterations were pooled and labeled as 'fake'. The model was initially trained for 5 epochs with the InceptionV3 model frozen. Subsequently, the last two convolutional layers of InceptionV3 were unfrozen and training continued for another 20 epochs. Binary cross-entropy was used as loss function with an Adam optimizer, a learning rate of 10^{-4} (reduced to 10^{-5} after 5 epochs), and a batch size of 16. Finally, the model was tested on the 240 images used in the detection task and the predicted logits were recorded for each image. No behavioral data was used.

Rating task model. The challenge of training a model on the rating task was the small number of only 30 images. Therefore, we explored a transfer learning approach wherein we first trained the detection task model and then

fine-tuned it on the rating task data. To this end, we took the detection task model and replaced the sigmoid by a softmax layer with 5 units representing the Likert scale ratings. The 30 rating task images were split into train and test sets using 5-fold cross-validation. For the training images, class labels (representing ratings) could not be assigned unequivocally since there were disagreements between the human raters. To encode this uncertainty in the model we used probabilistic class labels: For each image, the empirical distribution of ratings was determined across raters and used to initialize a probability distribution. In each batch, class labels were randomly sampled from these distributions. The model was initially trained for 20 epochs with the InceptionV3 model and the MRI features frozen. After this, the MRI features were unfrozen and model was trained for another 200 epochs. Categorical cross-entropy was used as loss function with an Adam optimizer, a learning rate of 10^{-4} (reduced to 10^{-5} after 20 epochs), and batch size 6. Both models were trained 100 times with randomly initialized weights. Results were averaged across runs.

2.4.3. Hardware and software

GANs were trained on NVIDIA Tesla P100 GPUs with 16 GB VRAM using Supercomputing Wales with TensorFlow 1. Statistical analyses were performed on a standard Desktop computer using Python 3.6 with the packages Statsmodels 0.13, Scipy 1.5.2, and TensorFlow 2.4, as well as Matlab R2018b.

2.4.4. Control analyses using a 3D StyleGAN

To verify that our results are not contingent on a specific choice of a generative model or imaging contrast we performed two control analyses. We used a different generative model called StyleGAN which uses a latent space combined with dense layers to modify activation maps at different stages of the generator [50]. StyleGANs have been successfully applied to 3D structural neuroimaging data [28]. First, to verify that our results generalize to a different model using the same data, we trained the StyleGAN on the GM maps used in our noise-to-image GAN. To this end, we used code from a publicly available repository[2]. The arrangement of 2D slices was very similar but not identical to the arrangement in the main experiment, due

[2]https://github.com/sh4174/3DStyleGAN

to differences in the processing pipeline. Second, to verify that our results also hold for a different imaging modality, we trained the model again using T1-weighted images from the same cohort. No behavioral data has been collected for these additional analyses. Therefore our analysis focused on applying the image quality metrics and investigating their consistency across models and imaging contrasts.

3. Results

3.1. Qualitative analysis of generated MRIs

Figure 2 shows a 3D rendering of the generated images using MRIcroGL [51]. To indicate how the images evolve across the different stages of training, Figure 2a shows images for six logarithmically spaced iterations ranging from an early iteration (iteration 11) to the final iteration (iteration 60000). Each of the six renderings is based on the same random vector. The GAN initially outputs a block of noise (iterations 11 and 90) from which a brain is eventually carved out (iterations 344 and 843). After this, GAN training seems to focus on improving details and removing noise voxels outside the brain volume (compare iterations 3,240 and 60000). More incidental evidence for the GAN devoting a lot of computational resources to cancelling noise outside the brain is provided in Figure S3 (Supplementary Material). Figure 2b shows an 'exploded' version of a generated MRI followed by two different random examples with a 3D view and various planar views. The model is able to convincingly mimic the overall shape of the brain and prominent structures such as lateral fissure and the cerebellum. The 'exploded' view indicates that the internal grey-matter structure is reproduced as well. The shortcomings of the GAN become more evident when viewed side by side with a real MRI example in Figure 2c. Here, large-scale structures such as the lateral fissure are more regular and the image looks less noisy overall.

3.2. Correlation analysis and spatial ICA

To ground these observations in a more quantitative analysis, we performed correlation analyses, an analysis of age and sex group effects, and a spatial ICA analysis, depicted in Figure 3. Correlations were calculated across all brain voxels between pairs of MRIs. To account for possible effects of small spatial misalignments and high-frequency noise, correlation analyses were repeated for images post-processed with a 2D Gaussian smoothing filter ($\sigma = 3$ voxels).

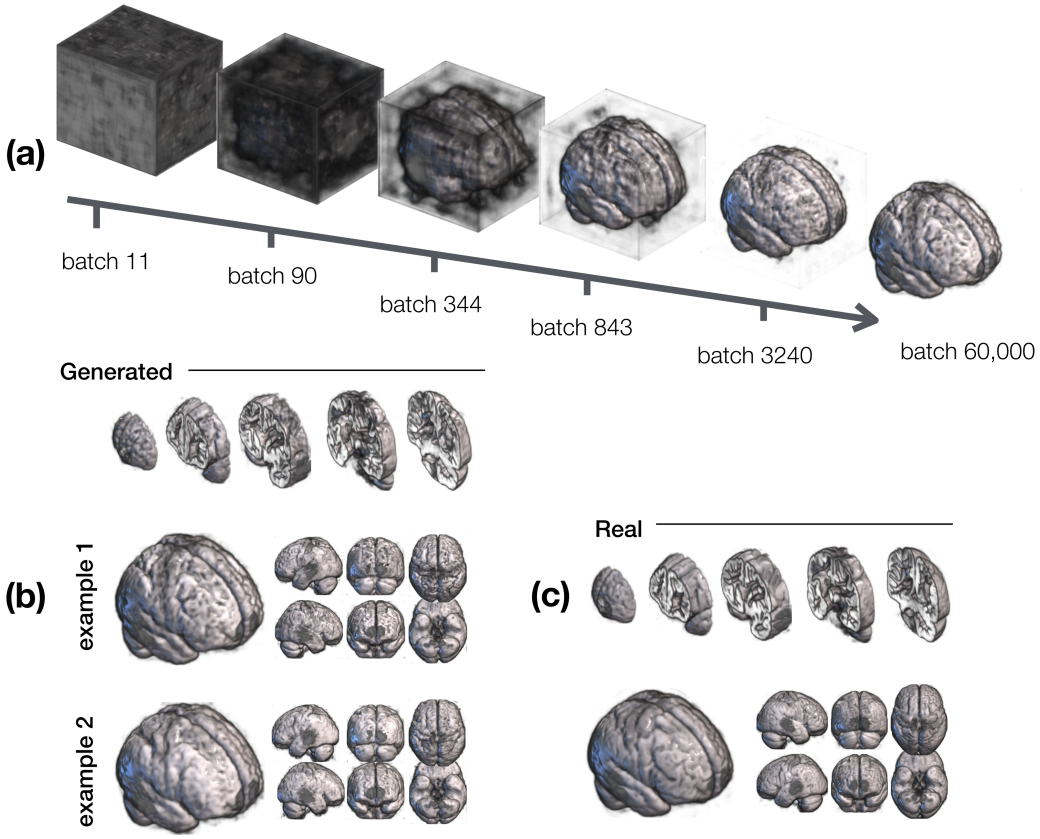

Figure 2: Qualitative results showing a MRIcroGL rendering of the generated 3D images. (a) Generated images at six different times (iterations) during GAN training. Whereas only a volume of noise is generated in the initial stages (iterations 11 and 90), the basic brain structure is evident from iteration 843. Much of the subsequent training appears devoted to refining details and removing the noise outside the brain. Iteration 60000 corresponds to the final model. (b) 'Exploded' view of a generated MRI in the first row indicates that the GAN correctly reproduces the internal structure of the brain, followed by two examples for generated MRIs. (c) For comparative purposes, the same rendering for a real MRI is shown. Comparing real and generated MRI, we observe that the generated image looks slightly more noisy and irregular, especially for large elongated structures such as the lateral fissure.

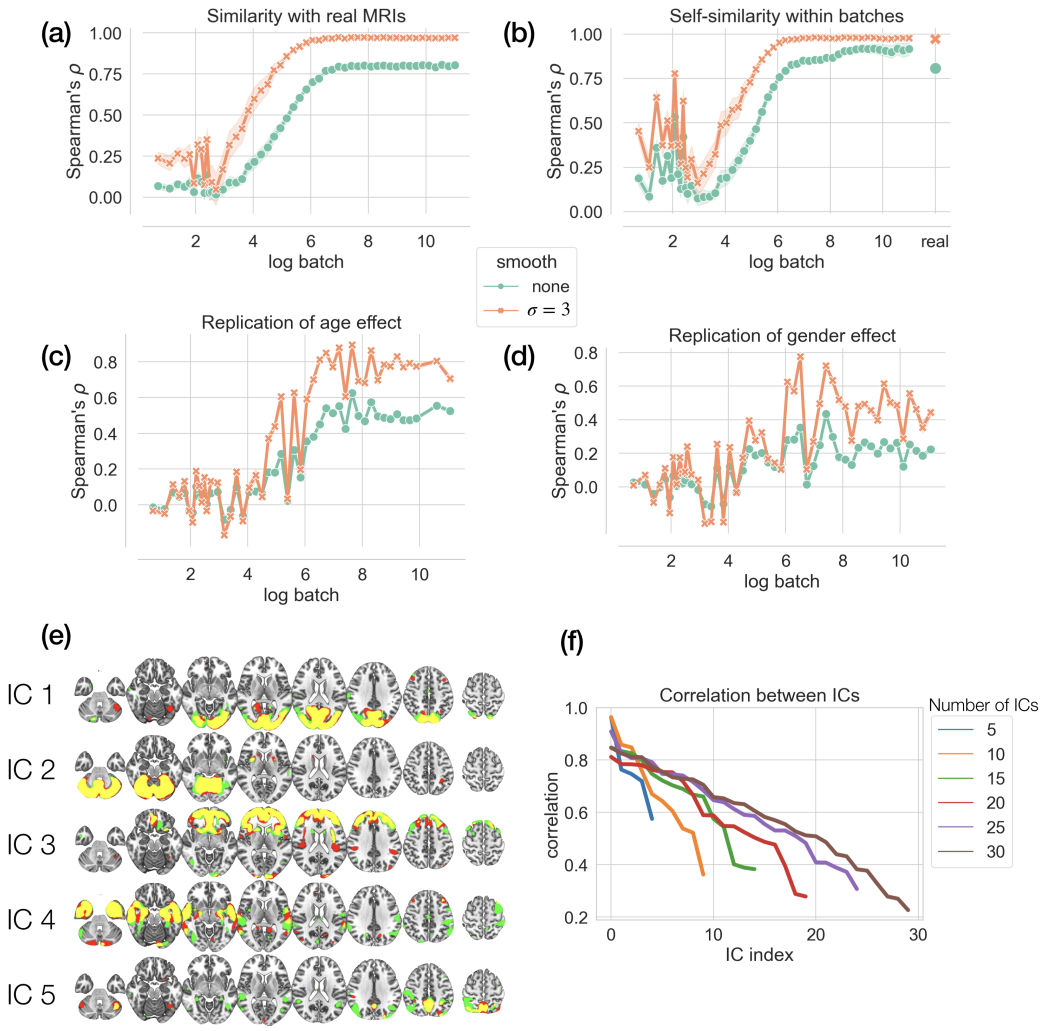

Figure 3: Correlation analyses and spatial ICA. (a) Similarity between generated and real 3D MRIs measured as their average correlation across voxels as a function of log iteration. The shaded area corresponds to 1 standard deviation. Results shown for original data without smoothing (green) and smoothed data (Gaussian kernel, $\sigma = 3$, orange). (b) Within-iteration similarity, average correlation between pairs of images within an iteration. Correlation within real MRIs is shown as an additional data point on the right. (c) Replication of the age effect ('old minus young'). (d) Replication of sex effect ('male minus female'). Correlation starts around zero and increases significantly with training, settling at a significant but moderately low value of 0.17. Smoothing significantly boosts correlation (0.42). (e) Spatial ICA maps for ICA with 5 components. Components for both real and generated MRIs have been matched and overlayed. Color coding indicates whether a voxel belongs the real (red) or generated (green) MRIs, with the intersection shown in yellow. The large overlap suggests significant spatial correspondence between real and generated MRIs. (d) Correlations between real and generated ICs for ICAs with different numbers of components (between 5 and 30). High degree of correspondence for the first few components is consistently found.

Figure 3a depicts Spearman's rank correlation ρ between generated and real MRIs. Every generated image was paired and correlated with every real image, then mean and standard deviation were calculated across all pairs. Correlation increased significantly then saturated relatively early in training. Since we masked out non-brain voxels, this early saturation nicely dovetails with our earlier observation that the late stages of GAN training seem largely devoted to removing noise outside the brain. The correlation for the final model was 0.8. Smoothing with a Gaussian kernel with a standard deviation of $\sigma = 3$ voxels yielded a correlation of 0.97, a statistically significant increase (Wilcoxon Rank Sum test, $z = 799.76, p < 0.0001$). The significant effect of smoothing suggests a high degree of correspondence between the large-scale structure of generated and real MRIs.

Figure 3b depicts the correlation among MRIs within iterations. Correlation among real MRIs has been added as additional data points (separate rightmost data points). Correlation started low and quickly saturated at a high level. For the final model the correlation amounted to 0.92, whereas it was 0.98 with smoothing, a significant increase ($z = 563.47, p < 0.0001$). Corresponding correlations for real MRIs were 0.81 without and 0.97 with smoothing, again a significant difference ($z = 565.08, p < 0.0001$). Directly comparing generated and real MRIs, correlation was significantly higher for generated MRIs both for unsmoothed ($z = 549.74, p < 0.0001$) and smoothed ($z = 350.98, p < 0.0001$) data.

Figure 3c depicts how well the GAN replicates the age effect found in real data. In real data, the age effect was operationalized as the average MRI in the elderly group (> 70 years) minus the average MRI in the young group (< 40 years). To generate MRIs specifically from these groups, two different GANs were trained, one on the elderly MRIs and one on the young ones. Since the age effect was a group effect, correlations for individual MRIs were not available. Therefore, to calculate errorbars, 100 different bootstrap samples of the GAN data were created. Correlation between real age effects and the generated age effects increased with iteration, suggesting that the GAN was sensitive to these group characteristics. In the final iteration, correlation was 0.52 for the unsmoothed and 0.71 for the smoothed data, a significant difference ($z = 12.22, p < 0.0001$).

Figure 3d depicts the replication of a gender effect, operationalized as the average 'male MRI' minus the average 'female MRI'. Again, two different GANs were trained on male and female MRIs separately and bootstrapping was used to calculate errorbars. Correlation increased with iteration, again

suggesting that the GAN is sensitive to these group characteristics. However, it settled at a lower value than for the age effect. In the final iteration, correlation was 0.22 for the unsmoothed and 0.44 for the smoothed data, a significant difference ($z = 12.22, p < 0.0001$).

Figure 3e depicts 5 ICs for a spatial ICA with 5 components. ICs in real and generated MRIs were matched by their correlation. To binarize the maps, one-sample t-values were calculated for all participants and thresholded at a t-value of 3. A large amount of overlap (shown in yellow) is evident especially for the first few ICs. Non-overlapping voxels are shown in red for real and green for generated data. Figure 3f depicts the correlations between IC maps for different spatial ICAs using a different number of components (5, 10, 15, 20, 25, 30). For all analyses, we found correlation values > 0.57 for the first five components, suggesting a moderate to high degree of consistency of components for real and generated MRIs.

3.3. Behavioral results

Figure 4 depicts the behavioral results. For both tasks, reaction times (RTs) increased with iteration, approaching the RT for real images (Figure 4a and d). For the detection task, the percentage of 'real' responses increased with the iteration, although it stayed short of the corresponding number for real MRIs (Figure 4b). For completeness, the correct score is also depicted in Figure 4c. For the subjective rating task, Mean Opinion Scores (MOS) are shown, that is, averages when ratings are encoded as integer numbers 1 through 5. MOS increased with iteration but did not reach the MOS obtained for real MRIs.

To substantiate these observations statistically, we used linear mixed-effects models (LMMs) with a fixed effect (log iteration) and a random effect (participant). LMMs are useful in the case of correlated responses arising from multiple measurements per participant and are more flexible than repeated-measures Analysis of Variance [52]. LMMs deal with pseudo-replication and account for the fact that participants display individual differences in overall reaction time and accuracy.

In the detection task, RT increased significantly with log iteration as evidenced by the positive regression slope (LMM, $\beta = 224.705, z = 15.837, p < 0.001$). Analogously, there was significant increase in 'real' responses with log iteration ($\beta = 6.184, z = 9.343, p < 0.001$) and a significant decrease in correct responses ($\beta = -6.195, z = -9.405, p < 0.001$). We then used Wilcoxon Signed-Rank tests to compare RTs and performance for iteration

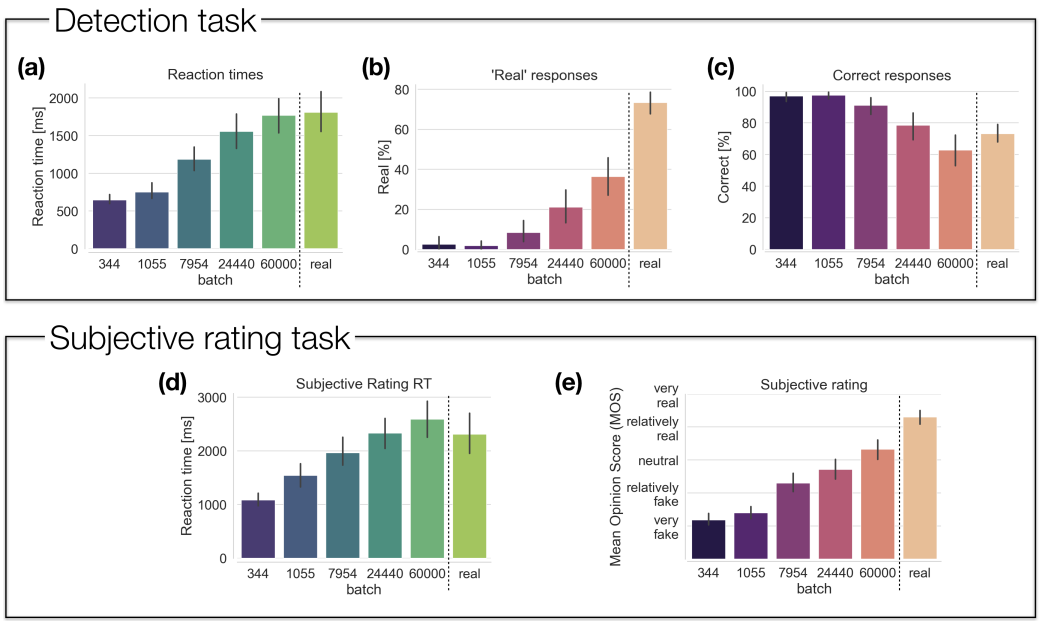

Figure 4: Behavioral experiment results for the detection task and subjective rating task. Results are shown for generated MRIs from different training iterations (indicated by the iterations) and for real MRI data. (a) In the detection task, reaction times (RTs) increased with iteration, approaching the RT obtained for real MRIs. (b) Percentage 'real' responses increased with iteration but stayed short of the proportion obtained for real data. (c) Percentage correct responses decreased with iteration, indicating that it becomes harder to distinguish real and generated MRIs. (d) In the subjective rating task, RTs increased with iteration. For iteration 60000 RT slightly overshot the RT found for real MRI. (e) Subjective ratings, here summarized as Mean Opinion Scores, increased with iteration albeit staying below the rating for real MRIs for iteration 60000.

60000 vs real MRIs across participants. No difference was found for RTs ($w = 146, p = 0.672$) and correct responses ($w = 119, p = 0.252$) but the proportion 'real' was significantly higher for real MRIs than for generated ones ($w = 0, p < 0.0001$).

In the subjective rating task, RT increased significantly with log iteration ($\beta = 279.875, z = 11.377, p < 0.001$). For the analysis of the ratings, a linear mixed model approach would not be not sufficient since the target variable was ordinal and the 'distances' between neighbouring categories were unknown [53]. The Mean Opinion Scores displayed in Figure 4e were used for illustrative purposes. In line with the recommendations in [53] we used a multi-level proportional odds model with a logit link function which simultaneously deals with mixed effects and ordinal responses. To this end, we fit a Cumulative Link Mixed Model with Laplace approximation [54], wherein rate was used as the ordinal target variable, log iteration as fixed effect and participant as a random effect with random intercept. We found a significant positive slope for log iteration ($\beta = 1.234, z = 17.94, p < 0.001$) signifying higher ratings for later iterations. The same statistical approach was applied to compare ratings for iteration 60000 vs real MRIs, showing significantly higher ratings for real MRIs as compared to generated ones ($\beta = 2.133, z = 7.748, p < 0.0001$).

3.4. Image quality metrics and Deep QA model

Figure 5 shows the metrics calculated for the detection task images. An analogous analysis applied to the rating task images and additional full reference metrics are depicted in Figure S2 (Supplementary Material). Figure 6b shows the corresponding result for the Deep QA model. Linear regression analyses of log iteration on the metric were conducted to investigate whether the metric increases/decreases with iteration.

For the detection data, all regression slopes were significant (all p-values < 0.0001). IS and MIS both decreased with iteration. Since they are quality metrics, the opposite pattern was expected. FID and MMD both decreased with iteration but images at iteration 24400 had a lower value than images at iteration 60000 (Wilcoxon Rank Sum, FID $z = 12.217, p < 0.001$; MMD $z = 12.217, p < 0.001$), conflicting with the behavioral data. Although NIQE, NIQE-MRI and BRISQUE-MRI showed overall regression trends in line with behavioral data, the metrics did not change monotonically with iteration. For instance, NIQE decreased from iteration 344 to 24440 but then increased again for iteration 60000. BRISQUE showed an inverted U-shaped

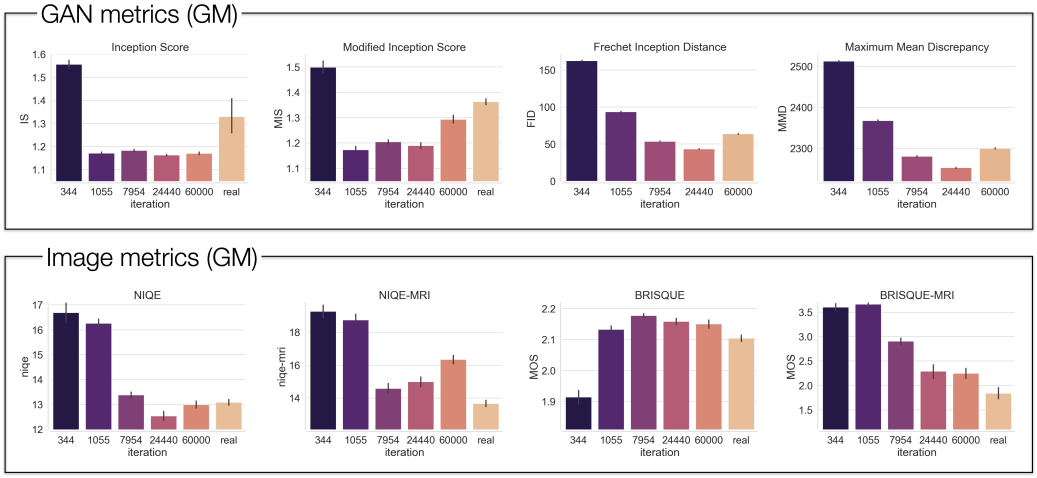

Figure 5: Image quality metrics applied to the 240 images used in the detection experiment. Different GAN and image quality metrics are shown, applied to generated images from different iterations. If the metric does not require a reference, the value for real images is shown as well.

pattern. Contrary to expectations, iteration 344 had the lowest score making it the best in terms of image quality. The Deep QA model was the only model for which predictions on the detection data changed monotonically with iteration, although the difference between iterations 24400 and 60000 was not significant ($p = 0.51$).

We repeated the same analysis on the rating task data (Figure S2 and Figure 6b). All regression slopes were significant ($p < 0.001$) except for IS ($p = 0.44$). Deep QA was the only metric that showed a significant difference between iterations 24400 and 60000 ($z = -2.193, p = 0.0282$).

To statistically compare the image quality metrics to the behavioral data, we performed Spearman's rank correlation analyses for both tasks, shown in Table 1. For the detection task, we used the proportion 'real' responses for each of the 240 images, averaged across participants, and correlated them with the metrics obtained on the same data (Figure 5 and Figure 6b). FID and MMD estimate the distance between distributions and hence do not provide values for individual images. To nevertheless estimate correlation, we used the value obtained for each iteration for all images belonging to this iteration. Furthermore, since some metrics did not provide separate predictions for real images, we conducted two correlation analyses, one analysis that covered only generated images and another one that included real im-

ages (denoted as 'include real' in the table). For the rating task, we used the Mean Opinion Scores for each of the 30 images, averaged across participants, and correlated them with the metrics obtained on the rating task data (Figure 6c and Figure S2).

Except for IS and MIS all metrics showed moderate to high correlation with the behavioral data. For the detection task, FID ($\rho = -0.602$), MMD ($\rho = -0.602$), NIQE ($\rho = -0.706$), and BRISQUE-MRI ($\rho = -0.859$) were the best performing metrics when the real data was excluded. When real images were included, the best performing metrics were NIQE ($\rho = -0.519$), NIQE-MRI ($\rho = -0.761$), BRISQUE-MRI ($\rho = -0.858$), and Deep QA ($\rho = 0.831$). Despite the fact that the Deep QA model performed the classification task better than humans (100% accuracy), it assigned higher probabilities to later iterations which led to a ranking consistent with humans (Figure 6b). However, when real images were removed from the correlation analysis, the correlation for the Deep QA model plummeted to $\rho = 0.516$, indicating that the effect was largely driven by the difference between real and generated images.

For the rating task, NIQE ($\rho = -0.767$), NIQE-MRI ($\rho = -0.575$), BRISQUE-MRI ($\rho = -0.840$), and Deep QA ($\rho = -0.879$) were the best performing metrics when the real data was excluded. The same qualitative pattern was found when real images were included (NIQE $\rho = -0.614$, NIQE-MRI $\rho = -0.722$, BRISQUE-MRI $\rho = -0.879$, and Deep QA $\rho = -0.912$).

Table 1: Correlation between behavioral data and image quality metrics reported as Spearman's ρ and corresponding p-value, for the two behavioral tasks separately. Each analysis was performed twice, one time using only the data on the generated images, and a second time including the real data, too. For each analysis, the four metrics with the largest absolute correlation are highlighted in bold.

Task		IS	MIS	FID	MMD	NIQE	NIQE-MRI	BRISQUE	BRISQUE-MRI	DeepQA
Detection	ρ	-0.255	0.11	**-0.602**	**-0.602**	**-0.706**	-0.536	0.465	**-0.859**	0.516
	p-value	0.001	0.11	**0.000**	**0.000**	**0.000**	0.000	0.000	**0.000**	0.000
Detection (include real)	ρ	0.008	0.344	-	-	**-0.519**	**-0.761**	-0.013	**-0.858**	**0.831**
	p-value	0.9	0.000	-	-	**0.000**	**0.000**	0.840	**0.000**	**0.000**
Rating	ρ	-0.092	0.367	-0.422	-0.422	**-0.767**	**-0.575**	0.521	**-0.840**	**0.879**
	p-value	0.66	0.071	0.035	0.035	**0.000**	**0.003**	0.008	**0.000**	**0.000**
Rating (include real)	ρ	0.003	0.232	-	-	**-0.614**	**-0.722**	0.294	**-0.879**	**0.912**
	p-value	0.98	0.213	-	-	**0.000**	**0.000**	0.114	**0.000**	**0.000**

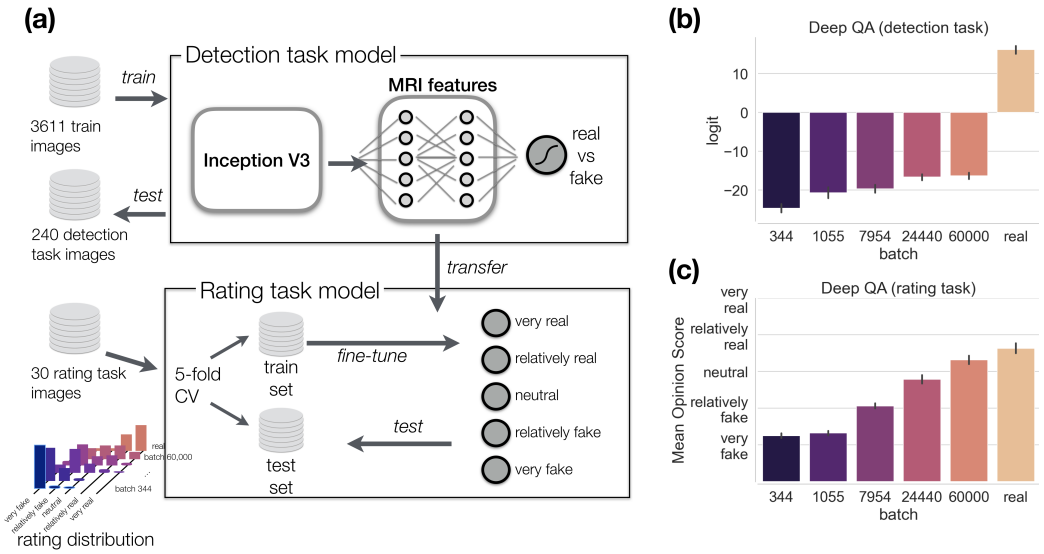

Figure 6: Deep QA model. (a) Model architecture. A pretrained InceptionV3 model was used followed by two dense layers (MRI features) trained on independent data. It was then fine-tuned on the rating data using cross-validation. (b) Logits of the predicted probabilities on the detection task images. (c) Mean Opinion Scores on the rating task images. Means and confidence intervals were calculated across 100 runs.

3.5. Control analyses

Figure 7 shows the metrics calculated for the GM and T1 images using a StyleGAN. Linear regression analyses of log iteration on the metric were conducted on both GM and T1 to investigate whether the metric increases/decreases with iteration. For IS, the regression slope was not significantly different from zero (GM $p = 0.12$; T1 $p = 0.13$). For MIS, results were inconsistent, with a small negative slope for GM data ($\beta = -0.0685, p < 0.0001$) but a positive slope for T1 ($\beta = 0.0066, p = 0.03$). For both FID and MMD, regression slopes were significantly negative for both GM and T1 (all $p < 0.0001$). Furthermore, these two metrics were the only to have a strictly monotonic relationship with iteration. NIQE, NIQE-MRI and BRISQUE decreased significantly with iteration for both GM and T1 (all $p < 0.0001$), although the only metric that did so strictly monotonically was BRISQUE applied on GM data. The BRISQUE-MRI transfer used the BRISQUE model trained on the GM data used in the behavioral experiment. The model did not transfer well to T1. It did not even yield consistent results on the GM data used in the StyleGAN when compared to the GM data used in the noise-to-image GAN. Possibly, this was due to differences in the way the slices were placed next to each other to produce a 2D image. The Deep QA model was tested in an out-of-distribution setting on the StyleGAN images. For GM, it yielded higher scores for real images than for early iterations, but iterations 128 and 200 were ranked as higher than real images. For T1 data, it yielded a constant output.

3.6. Data sharing

All scripts, real and generated images, experimental data, behavioral results and metrics, and the Deep QA model are available at `github.com/treder/MRI-GAN-QA`.

4. Discussion

Human assessment is the current gold standard for judging image quality of images generated by GANs [8, 22]. However, the utility of human assessment is limited by economical factors, limited scalability of human raters to large datasets, and the limited sensitivity of the human eye to subtle statistical relationships across dozens or hundreds of images. By collecting both behavioral data, performing group analysis and spatial ICA, and surveying a whole range of image quality metrics, our goal was to identify a more automated approach for the assessment of generated images. These points will

Figure 7: Image quality metrics applied to the StyleGAN images for (a) GM and (b) T1-weighted images, and separately for (c) Deep QA model applied to the GM and T1 images. Iterations given in units of thousands.

be expanded on in the next two subsections, followed by a broader discussion of our findings, the limits of GANs for image generation, and the ideal image quality metric.

4.1. Biological plausibility: group analysis and spatial ICA

Prior to performing spatial ICA, we investigated the degree of correspondence between real and generated MRIs at a voxel level. The correlation between generated and real images was 0.80 (Figure 3a), which was very similar to the correlation of 0.81 of real images with themselves (Figure 3b). However, the correlation of generated images with each other was 0.92. This suggests that while generated MRIs adequately approximate real MRIs, the latter have a larger amount of diversity, in line with a previous research [8]. One explanation for this is that the GAN lacks *capacity*, that is, it is simply unable to replicate the diversity found in real data because the data does not lie within the range of possible outputs of the generator. Alternatively, the GAN might simply have failed to learn the generative distribution of the MRIs, a possibility discussed in Section 4.3.

We then performed group analysis to investigate whether the GAN reproduces group differences. This is not a trivial question since the GAN is only trained to produce images that 'look like brains'. It is not trained to reproduce statistical regularities and the discriminator that guides the training process does not have access to all images simultaneously to facilitate such a learning. Analysis of group differences considered training separate GANs to characterize chronological age (young vs old) and sex (male vs female). The groups of generated images were then subjected to group analysis. We found that the age effect is fairly well reproduced (Figure 3c) while the replication of the gender effect was rather low (Figure 3d). The low reproduction of the gender effect might be due to the fact that gender differences contribute less variance to the grey matter signal than aging which involves global atrophy.

Furthermore, we performed spatial ICA to investigate whether the GAN reproduces structural brain networks known to support cognitive function. Again, the model is not explicitly trained to perform this. Performing spatial ICA on real and generated images separately and comparing the results we found a large degree of correspondence between the spatial maps (Figure 3e). We found spatial patterns reminiscent of the networks identified in other studies of ageing with source-based morphometry [55, 32]. This correlation between the spatial maps was high irrespective of the exact choice of the number of ICs.

4.2. Image quality metrics as a proxy for human assessment

To identify whether image quality metrics correspond well with human assessment, we collected human data in two different tasks (detection and rating) and calculated a series of image quality metrics for the same images. For the human data, we found a significant increase in 'real' responses with iteration. Since better performance was associated with lower RT, this could not be explained in terms of a speed-accuracy trade-off. A similar pattern of results was found in the subjective rating task, with Mean Opinion Scores (MOS) increasing significantly with iteration and approaching the ratings for real images. As hypothesized, these results confirmed that changes in objective image quality operationalized by iteration are perceptually relevant and measurable. Behavioral measures converged towards the responses obtained for real images, but there remained a gap, indicating that our model was not consistently able to fool human experts.

To relate behavioral data to the image quality metrics, we calculated Spearman's rank correlation between quality metrics and the proportion of 'real' responses in the behavioral data. The standard GAN metrics IS and MIS showed poor correspondence with human assessment, in line with earlier research [22]. Both FID and MMD showed a good correspondence with behavior in the detection task (both $\rho = -0.602$). In the rating task, the correlation decreased (both $\rho = -0.422$) but this can possibly explained by the fact that there were only 5 images in each iteration. Since FID and MMD are distribution-level metrics, the estimation of distribution parameters could be noisy or unreliable with so few examples. The control analyses using StyleGAN gave further credence to FID and MMD as useful image quality metrics for MRIs. Both metrics were the only to decrease strictly monotonically with iteration.

Unlike FID and MMD, NIQE and BRISQUE do not require a reference and provide ratings for individual images. NIQE showed a moderate correlation with behavior in the detection task ($\rho = -0.706$, dropping to $\rho = -0.519$ when including real images) and the rating task ($\rho = -0.767$, $\rho = -0.614$ when including real images). Upon closer inspection, the metric decreased monotonically until iteration 24440, yielding larger values for iteration 60000 and real images. A roughly similar pattern for was found for the StyleGAN T1 images. A pattern contradicting this was found for the StyleGAN GM images, with the real images receiving a higher score than all generated images. NIQE can be adapted to the target domain in an unsupervised way. The resultant NIQE-MRI also showed a moderate correlation with behavior

(detection task $\rho = -0.536$, $\rho = -0.761$ including real images; rating task $\rho = -0.575$, $\rho = -0.722$ including real images). NIQE-MRI also showed a slightly more consistent pattern for StyleGAN GM and T1 images, with the lowest score given to real images for all but one iteration and generally higher scores given to earlier iterations. Overall, NIQE-MRI provided an accurate tripartite pattern, which high values for early iterations, medium values for later iterations, and lower values for real images. However, both NIQE and NIQE-MRI fail to consistently reproduce monotonically decreasing values for later iterations.

BRISQUE showed an inconsistent, inverted U-shaped pattern in the main experiment. Correlation with behavior was moderate (detection task $\rho = 0.465$, rating task $\rho = 0.521$; correlations not significant when including real images) but it was positive whereas a negative correlation was expected. In the StyleGAN, it provided a nearly monotonic pattern for GM images but for not for T1 images. BRISQUE can be adapted to the target domain in a supervised way by training on target MOS. The resultant BRISQUE-MRI showed an adequate monotonic pattern which however did not transfer well to the StyleGAN images, suggesting possible overfitting to the images from the main experiment. We conclude that neigher BRISQUE nor BRISQUE-MRI are reliable image quality metrics for the MRI domain.

The two Deep QA models showed a moderate correlation with human data in the detection task ($\rho = 0.516$, rising to $\rho = 0.831$ when including real images) and a high correlation in the rating task ($\rho = 0.879$ and $\rho = 0.912$ when including real images). The Deep QA models where the only metrics that changed monotonically with iteration. It is worth noting that only the rating model was explicitly trained on human responses, whereas the detection model was trained to discriminate between real and generated images. Crucially, the Deep QA rating model was the only metric that showed a significant increase in quality ratings from the 24400 to the 60000 iteration.

However, the Deep QA rating model did not readily transfer to the StyleGAN images, with a near constant output for most iterations in both GM and T1 images (Figure 7c). This suggests that the model overfit to the training dataset. To develop a more versatile model that we envision several possible avenues. Firstly, training the model using data augmentation (e.g. skewing and rotations of the images) might increase generalizability. Secondly, our data could be combined with ratings obtained in other studies. A single model could then be trained to predict ratings on different datasets simulta-

neously. To achieve this, the model would need to learn some generalization capabilities and hence be more likely to transfer to new datasets. To facilitate this, our human ratings, corresponding images and training scripts have been published in a GitHub repository [3]. Thirdly, a semi-supervised learning approach could be considered wherein a labeled dataset is combined with an additional unlabeled dataset (e.g., [56]). For instance, the Deep QA could be trained to simultaneously predict MOS on our data and predict whether images in the unlabeled dataset are real or fake by adding a separate real/fake output unit to the model. At inference time, this output unit is dropped and ratings are obtained via the other units. Exploring these avenues in more detail is left for future work.

Across many of the surveyed metrics, we found that they correlate with human behavior reasonably well for lower iterations (iterations 344, 1055, 7954) but are less sensitive to image differences later in training (iterations 24400 vs 60000). We believe this discrepancy might be due to most metrics being sensitive to *distortion* rather than *perceptual quality*, two different dimensions of image quality that have been explored theoretically in image restoration tasks [57]. *Distortion* refers to the deviation of a distorted image from a noise-free reference whereas *perceptual quality* measures the more elusive 'naturalness' of an image regardless of a reference. In our experiment, informal evidence in line with this is a participant stating[4] that they focused on image artifacts such as checkerboard patterns for low-quality images (i.e., distortion), but that it was rather subtle differences in luminance across the image would that made higher-quality generated images look fake (i.e., perceptual quality).

If human assessment of image quality was indeed governed by distortion and artifacts for earlier iterations and a more intuitive understanding of what an MRI should look like for later iterations, it is reasonable to believe that most of the metrics were sensitive to artifacts, hence explaining their good performance for early iterations and bad performance for the last iteration. The only model that consistently agreed with human data across all quality levels was the Deep QA model. The model was directly trained on the images and human responses which gave it the opportunity to learn the more subtle

[3] https://github.com/treder/MRI-GAN-QA

[4] the participant disclosed their participation in the experiment to the author in a personal email

features that humans use for more high-quality generated images.

4.3. The limits of GANs

The promise of GANs is somewhat uncanny: After being exposed to a few hundred images it supposedly learns the brain image manifold in image space. A key question therefore is whether the GAN indeed learns (a) the generative distribution of MRIs, whether it learns (b) a different distribution, or (c) simply memorizes and reproduces the training data. Regarding (c), it turns out that memorization of data requires a large number of different samples [58]. Other indicators contra memorization are the fact that interpolation between noise vectors can produce novel and meaningful image variations [59] and the clear disparity seen between real and generated images [60].

In accordance with (a), GANs can closely approximate the underlying generative distribution for sufficiently large datasets [6]. Generalization beyond the training data, indicated by the emergence of novel combinations of features, has been shown in simple datasets [61]. However, there is doubt that the same holds for limited sample sizes [60]. For instance, the *generative support* (crudely speaking, the number of individually different images the GAN can produce) found for faces strongly depends on the model architecture, with estimates ranging from 160,000 to over a million faces. This casts doubt on option (a). Further support for (b) is given by the phenomenon of mode collapse, i.e. the tendency of GANs to concentrate its probability mass to a few modes. [62] showed that the relative frequency of attributes such as hair style for faces or room type for indoor scenes is distorted from the real proportion in the data. Furthermore, classifier performance deteriorates when trained on GAN images rather than real data, and [62] estimated that the effective dataset size of the GAN is 100x smaller than the training data.

The debate on the generative support of GANs is ongoing, but results so far favour alternative (b), that is, GANs probably approximate the true distribution rather than just memorizing the data, but the support of the GAN distribution is limited. This limitation can probably be partially counteracted by using a large and diverse training set and a sufficiently capable CNN architecture. For limited size datasets, data augmentation could be a viable tool for increasing set size and diversity.

4.4. The 'ideal' quality metric

Although the focus of this paper was to assess how well existing metrics agree with human data, it seems expedient to list other favourable properties

that an ideal MRI quality metric should possess:

1. *Model agnostic*. The metric should be independent of the specific model architecture in order to enable quantitative comparison across models. This precludes e.g. the discriminator's loss function to act as a metric.
2. *Reference free*. It should determine the quality of an image per se, without resorting to a reference image. This would make it applicable to both noise-to-image and image-to-image applications and also in situations wherein the ground truth is not available.
3. *Individual image ratings*. A measure of image quality should be provided for individual images rather than the whole dataset.
4. *Multi-dimensional*. Although MRIs are often depicted as a series of 2D slices, the slices together actually form a 3D image and an accurate assessment should take into account all spatial dimensions.
5. *Human performance*. The metric has to reproduce the perceptual assessment of expert viewers. Furthermore, there is evidence that image manipulations that are imperceptible to humans can have a significant effect on CNN predictions [63]. Therefore, we believe that human assessment should serve as a *lower sensitivity bound* when evaluating image metrics. In other words, if humans are sensitive to the difference between two types of images, so should be the metric – but the metric could be sensitive to differences that humans are not.

None of the surveyed metrics meets all of these criteria. All of the surveyed metrics are model agnostic but only IS, MIS, NIQE, BRISQUE, and the Deep QA model are reference free. Individual image ratings are not available for FID and MMD because they operate on the level of distributions rather than individual images. Although all metrics were applied to 2D image slices, IS, MIS, MMD and MMD can be seamlessly extended to 3D by deriving a feature embedding from a 3D CNN. The Deep QA model can be trained on 3D images using 3D convolutions. A difficulty in extending metrics to 3D images is the collection of experimental data. A bespoke experimental protocol for assessing 3D images wherein participants can e.g. rotate, zoom into and slice 3D MRI images would need to be developed.

4.5. Limitations

Behavioral results were collected using an online experiment rather than a controlled study in a lab environment. Participants performed the experiments in different environments using different display devices (e.g. laptop

screen vs monitor). This potentially reduces the internal validity of our results, although it may increase their external validity. We believe that this limitation mostly affects group analyses. For instance, the seniority of a participant may be correlated with the display device they use, thereby confounding group differences between junior and senior experts. However, we performed only within-subject analyses. If anything, the fact that our findings are consistent across a variety of environments and display devices strengthens rather than weakens their validity. For instance, using Amazon Mechanical Turk, [64] replicated various visual effects such as Stroop effect and attentional blink that have been studied extensively in laboratory settings. Another limitation is that we only considered a noise-to-image model, no image-to-image model. In the latter, reference images are available which might improve the quality of metrics that rely on reference images.

4.6. Conclusion

As a practical recommendation for evaluating brain images produced by GANs, we propose that researchers assess both local properties (image quality of individual images or groups of images) and global properties (statistical regularities across the whole dataset) of the generated data.

With regard to global properties, the generated dataset should reproduce relevant group differences such as young vs elderly or male vs female, with the specific choice of relevant groups depending on the nature of the dataset and the research question. Sensitivity of the GAN to group differences can be investigated by either re-training it on separate groups, as done in this paper, or using a conditional GAN with the demographics as separate input variables. Furthermore, a realistic GAN should be able to reproduce the large-scale structural networks exposed by multivariate analyses across the dataset. We performed spatial ICA to verify that the GAN reproduces components such as the Default Mode Network. Other approaches such as Principal Component Analysis (PCA) or non-negative matrix factorization [65] could be used for the same purpose.

With regard to local properties, the generated brain image should 'look like' a real one. This can be verified using both distortion and perceptual metrics, two dimensions of image quality [57]. As distortion metrics, FID, MMD, and NIQE stood out as versatile, widely used image quality metrics that do not require tuning to the MRI data. NIQE has complementary properties compared to FID and MMD in that it yields ratings for individual images without requiring a reference, whereas FID and MMD provide

a distribution-level metric for the difference between the set of generated images vs the set of real images. Therefore, we recommend the use of all three metrics. As a perceptual metric, our Deep QA rating model was the only metric that reproduced human data for higher quality images. Unfortunately, the model required human data to be trained. This defeats its purpose to some extent since we set out to escape the need to collect human data. Although speculative, a possible way out of this predicament is to pre-train a Deep QA model on a large, diverse dataset that integrates multiple data modalities [66, 67]. Ideally, this model could be used out-of-the-box for any new dataset. Optionally, it could be fine-tuned to the statistics of a new dataset without requiring human data.

Concluding, we propose a combination of local and global analyses for assessing the quality of generated images. For the time being, human assessment remains the gold standard for assessing individual images. In the future, we believe that a Deep QA model that has been trained to mimick human perceptual assessments can pave the way for quick and cost-effective image quality assessments that will accelerate GAN research and improve its validation.

Acknowledgements

We acknowledge the support of the Supercomputing Wales project, which is part-funded by the European Regional Development Fund (ERDF) via Welsh Government. KAT was supported by the Guarantors of Brain (G101149).

Author contributions

MT conceptualized the study and collected human data. RC trained the GAN. MT and KT performed the statistical analyses and drafted the manuscript. All authors reviewed and approved the manuscript.

References

[1] R. A. Poldrack, K. J. Gorgolewski, Making big data open: data sharing in neuroimaging, Nature Neuroscience 17 (2014) 1510–1517.

[2] H. Peng, W. Gong, C. F. Beckmann, A. Vedaldi, S. M. Smith, Accurate brain age prediction with lightweight deep neural networks, Medical Image Analysis 68 (2021) 101871.

[3] M. Sajjad, S. Khan, K. Muhammad, W. Wu, A. Ullah, S. W. Baik, Multi-grade brain tumor classification using deep CNN with extensive data augmentation, Journal of Computational Science 30 (2019) 174–182.

[4] G. Mohan, M. M. Subashini, MRI based medical image analysis: Survey on brain tumor grade classification, Biomedical Signal Processing and Control 39 (2018) 139–161.

[5] V. Sorin, Y. Barash, E. Konen, E. Klang, Creating Artificial Images for Radiology Applications Using Generative Adversarial Networks (GANs) – A Systematic Review, Academic Radiology 27 (2020) 1175–1185.

[6] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative Adversarial Nets, in: Z. G. Weinberger, M. Welling, C. Cortes, N. D. Lawrence, K. Q. (Eds.), Advances in Neural Information Processing Systems 27, Curran Associates, Inc., 2014, pp. 2672–2680.

[7] D. P. Kingma, M. Welling, Auto-encoding variational bayes, in: 2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings, International Conference on Learning Representations, ICLR, 2014.

[8] A. U. Hirte, M. Platscher, T. Joyce, J. J. Heit, E. Tranvinh, C. Federau, Diffusion-Weighted Magnetic Resonance Brain Images Generation with Generative Adversarial Networks and Variational Autoencoders: A Comparison Study, arXiv (2020) 2006.13944.

[9] C. K. Chong, E. T. W. Ho, Synthesis of 3D MRI Brain Images with Shape and Texture Generative Adversarial Deep Neural Networks, IEEE Access 9 (2021) 64747–64760.

[10] S. U. H. Dar, M. Yurt, L. Karacan, A. Erdem, E. Erdem, T. Çukur, Image Synthesis in Multi-Contrast MRI with Conditional Generative Adversarial Networks, IEEE Transactions on Medical Imaging 38 (2019) 2375–2388.

[11] D. Nie, R. Trullo, J. Lian, C. Petitjean, S. Ruan, Q. Wang, D. Shen, Medical Image Synthesis with Context-Aware Generative Adversarial

Networks, in: Medical Image Computing and Computer Assisted Intervention - MICCAI 2017. MICCAI 2017. Lecture Notes in Computer Science, Springer, Cham, 2017, pp. 417–425.

[12] D. Abramian, A. Eklund, Generating fMRI volumes from T1-weighted volumes using 3D CycleGAN, arXiv abs/1907.0 (2019).

[13] X. Gu, H. Knutsson, M. Nilsson, A. Eklund, Generating Diffusion MRI Scalar Maps from T1 Weighted Images Using Generative Adversarial Networks, Lecture Notes in Computer Science 11482 (2019) 489–498.

[14] Y. Chen, F. Shi, A. G. Christodoulou, Y. Xie, Z. Zhou, D. Li, Efficient and Accurate MRI Super-Resolution Using a Generative Adversarial Network and 3D Multi-level Densely Connected Network, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, Springer, Cham, 2018, pp. 91–99.

[15] T. M. Quan, T. Nguyen-Duc, W.-K. Jeong, Compressed Sensing MRI Reconstruction Using a Generative Adversarial Network With a Cyclic Loss, IEEE Transactions on Medical Imaging 37 (2018) 1488–1497.

[16] X. Yi, E. Walia, P. Babyn, Generative adversarial network in medical imaging: A review, arXiv e-prints 1809.07294 (2018).

[17] C. Bermudez, A. J. Plassard, T. L. Davis, A. T. Newton, S. M. Resnick, B. A. Landman, Learning Implicit Brain MRI Manifolds with Deep Learning, Proceedings of SPIE–the International Society for Optical Engineering 10574 (2018).

[18] K. Kazuhiro, R. A. Werner, F. Toriumi, M. S. Javadi, M. G. Pomper, L. B. Solnes, F. Verde, T. Higuchi, S. P. Rowe, Generative Adversarial Networks for the Creation of Realistic Artificial Brain Magnetic Resonance Images, Tomography (Ann Arbor, Mich.) 4 (2018) 159–163.

[19] A. Volokitin, E. Erdil, N. Karani, K. C. Tezcan, X. Chen, L. Van Gool, E. Konukoglu, Modelling the Distribution of 3D Brain MRI Using a 2D Slice VAE, in: Lecture Notes in Computer Science, volume 12267 LNCS, Springer Science and Business Media Deutschland GmbH, 2020, pp. 657–666.

[20] G. Kwon, C. Han, D.-s. Kim, Generation of 3D Brain MRI Using Auto-Encoding Generative Adversarial Networks, in: MICCAI 2019: Medical Image Computing and Computer Assisted Intervention, pp. 118–126.

[21] C. Han, L. Rundo, K. Murao, T. Noguchi, Y. Shimahara, Z. A. Milacski, S. Koshino, E. Sala, H. Nakayama, S. Satoh, MADGAN: unsupervised medical anomaly detection GAN using multiple adjacent brain MRI slice reconstruction, BMC Bioinformatics 2021 22:2 22 (2021) 1–20.

[22] F. Calimeri, A. Marzullo, C. Stamile, G. Terracina, Biomedical Data Augmentation Using Generative Adversarial Neural Networks, in: P. Verschure, A. Villa, A. Lintas, S. Rovetta (Eds.), Artificial Neural Networks and Machine Learning - ICANN 2017. ICANN 2017. Lecture Notes in Computer Science, Springer, Cham, 2017, pp. 626–634.

[23] C. Han, L. Rundo, K. Murao, Z. A. Milacski, K. Umemoto, H. Nakayama, S. Satoh, GAN-based Multiple Adjacent Brain MRI Slice Reconstruction for Unsupervised Alzheimer's Disease Diagnosis, in: In International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics. Lecture Notes in Computer Science, Springer, Cham, 2019, pp. 44–54.

[24] K. H. Kim, W. Do, S. Park, Improving resolution of MR images with an adversarial network incorporating images with different contrast, Medical Physics 45 (2018) 3120–3131.

[25] D. Yang, B. Liu, L. Axel, D. Metaxas, 3D LV Probabilistic Segmentation in Cardiac MRI Using Generative Adversarial Network, in: Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges, Springer, Cham, 2019, pp. 181–190.

[26] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: From error visibility to structural similarity, IEEE Transactions on Image Processing 13 (2004) 600–612.

[27] N. Luo, J. Sui, A. Abrol, J. Chen, J. A. Turner, E. Damaraju, Z. Fu, L. Fan, D. Lin, C. Zhuo, Y. Xu, D. C. Glahn, A. L. Rodrigue, M. T. Banich, G. D. Pearlson, V. D. Calhoun, Structural brain networks match intrinsic functional networks and vary across domains: a study from 15000+ individuals, Cerebral Cortex 30 (2020) 5460–5470.

[28] S. Hong, R. Marinescu, A. V. Dalca, A. K. Bonkhoff, M. Bretzner, N. S. Rost, P. Golland, 3D-StyleGAN: A Style-Based Generative Adversarial Network for Generative Modeling of Three-Dimensional Medical Images, in: 24th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI).

[29] M. Shafto, L. K. Tyler, M. Dixon, J. R. Taylor, J. B. Rowe, R. Cusack, A. J. Calder, W. D. Marslen-Wilson, J. Duncan, T. Dalgleish, R. N. Henson, C. Brayne, F. E. Matthews, The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing, BMC neurology 14 (2014) 204.

[30] J. R. Taylor, N. Williams, R. Cusack, T. Auer, M. A. Shafto, M. Dixon, L. K. Tyler, Cam-CAN, R. N. Henson, The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: Structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample, NeuroImage 144 (2017) 262–269.

[31] J. Ashburner, A fast diffeomorphic image registration algorithm, NeuroImage 38 (2007) 95–113.

[32] K. A. Tsvetanov, R. N. A. Henson, P. S. Jones, H. Mutsaerts, D. Fuhrmann, L. K. Tyler, J. B. Rowe, The effects of age on resting-state BOLD signal variability is explained by cardiovascular and cerebrovascular factors, Psychophysiology 58 (2021) e13714.

[33] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein Generative Adversarial Networks, in: Proceedings of the 34th International Conference on Machine Learning, pp. 214–223.

[34] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. C. Courville, Improved Training of Wasserstein GANs, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., Long Beach, California, USA, 2017, p. 5769–5779.

[35] G. Stoet, PsyToolkit: A software package for programming psychological experiments using Linux, Behavior Research Methods 42 (2010) 1096–1104.

[36] G. Stoet, PsyToolkit: A Novel Web-Based Method for Running Online Questionnaires and Reaction-Time Experiments, Teaching of Psychology 44 (2017) 24–31.

[37] R. Whelan, Effective analysis of reaction time data, Psychological Record 58 (2008) 475–482.

[38] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, H. Greenspan, GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification, Neurocomputing 321 (2018) 321–331.

[39] L. Xu, K. M. Groth, G. Pearlson, D. J. Schretlen, V. D. Calhoun, Source-based morphometry: The use of independent component analysis to identify gray matter differences with application to schizophrenia, Human Brain Mapping 30 (2009) 711–724.

[40] A. Borji, Pros and cons of GAN evaluation measures, Computer Vision and Image Understanding 179 (2019) 41–65.

[41] C. Lee, S. Woo, S. Baek, J. Han, J. Chae, J. Rim, Comparison of objective quality models for adaptive bit-streaming services, in: 2017 8th International Conference on Information, Intelligence, Systems & Applications (IISA), volume 2017, Institute of Electrical and Electronics Engineers Inc., 2017, pp. 1–4.

[42] J. Jia Deng, W. Wei Dong, R. Socher, L.-J. Li-Jia Li, K. Kai Li, L. Li Fei-Fei, ImageNet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 248–255.

[43] S. Barratt, R. Sharma, A Note on the Inception Score, arXiv (2018) 1801.01973.

[44] S. Gurumurthy, R. K. Sarvadevabhatla, V. B. Radhakrishnan, DeLiGAN : Generative Adversarial Networks for Diverse and Limited Data, Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 2017-Janua (2017) 4941–4949.

[45] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, GANs Trained by a Two Time-Scale Update Rule Converge to a Local

Nash Equilibrium, Advances in Neural Information Processing Systems 2017-Decem (2017) 6627–6638.

[46] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, A. Smola, A Kernel Two-Sample Test, Journal of Machine Learning Research 13 (2012) 723–773.

[47] A. Mittal, A. Krishna Moorthy, A. Conrad Bovik, No-Reference Image Quality Assessment in the Spatial Domain, IEEE Transactions on Image Processing 21 (2012) 4695–4708.

[48] G. Yang, Y. Cao, X. Xing, M. Wei, Perceptual Loss Based Super-Resolution Reconstruction from Single Magnetic Resonance Imaging, in: Lecture Notes in Computer Science, volume 11632, Springer Verlag, 2019, pp. 411–424.

[49] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, O. Wang, The Unreasonable Effectiveness of Deep Features as a Perceptual Metric, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 586–595.

[50] T. Karras, S. Laine, T. Aila, A Style-Based Generator Architecture for Generative Adversarial Networks, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2019-June (2018) 4396–4405.

[51] C. Rorden, M. Brett, Stereotaxic display of brain lesions, Behavioural Neurology 12 (2000) 191–200.

[52] D. A. Magezi, Linear mixed-effects models for within-participant psychology experiments: an introductory tutorial and free, graphical user interface (LMMgui), Frontiers in Psychology 6 (2015) 2.

[53] C. Keeble, P. D. Baxter, A. J. Gislason-Lee, L. A. Treadgold, A. G. Davies, Methods for the analysis of ordinal response data inmedical image quality assessment, British Journal of Radiology 89 (2016).

[54] R. H. B. Christensen, ordinal—Regression Models for Ordinal Data, 2019.

[55] K. Liu, S. Yao, K. Chen, J. Zhang, L. Yao, K. Li, Z. Jin, X. Guo, Structural Brain Network Changes across the Adult Lifespan, Frontiers in Aging Neuroscience 9 (2017) 275.

[56] C. Ge, I. Y. H. Gu, A. S. Jakola, J. Yang, Deep semi-supervised learning for brain tumor classification, BMC Medical Imaging 20 (2020).

[57] Y. Blau, T. Michaeli, The Perception-Distortion Tradeoff, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6228–6237.

[58] V. Nagarajan, C. Raffel, G. Brain, I. J. Goodfellow Google Brain, Theoretical Insights into Memorization in GANs, in: 32nd Conference on Neural Information Processing Systems.

[59] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, in: 4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings, International Conference on Learning Representations, ICLR, 2016.

[60] S. Arora, Y. Zhang, Do GANs actually learn the distribution? An empirical study, arXiv (2017) 1706.08224.

[61] S. Zhao, H. Ren, A. Yuan, J. Song, N. Goodman, S. Ermon, Bias and Generalization in Deep Generative Models: An Empirical Study, Advances in Neural Information Processing Systems 2018-Decem (2018) 10792–10801.

[62] S. Santurkar, L. Schmidt, A. Madry, A Classification-Based Perspective on GAN Distributions, in: International Conference on Learning Representations (ICLR).

[63] J. Zhang, C. Li, Adversarial Examples: Opportunities and Challenges, IEEE Transactions on Neural Networks and Learning Systems 31 (2020) 2578–2593.

[64] M. J. C. Crump, J. V. McDonnell, T. M. Gureckis, Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research, PLoS ONE 8 (2013) e57410.

[65] A. Anderson, P. K. Douglas, W. T. Kerr, V. S. Haynes, A. L. Yuille, J. Xie, Y. N. Wu, J. A. Brown, M. S. Cohen, Non-negative matrix factorization of multimodal MRI, fMRI and phenotypic data reveals differential changes in default mode subnetworks in ADHD, NeuroImage 102 (2014) 207–219.

[66] K. Aderghal, A. Khvostikov, A. Krylov, J. Benois-Pineau, K. Afdel, G. Catheline, Classification of Alzheimer Disease on Imaging Modalities with Deep CNNs Using Cross-Modal Transfer Learning, in: Proceedings - IEEE Symposium on Computer-Based Medical Systems, volume 2018-June, Institute of Electrical and Electronics Engineers Inc., 2018, pp. 345–350.

[67] K. Wang, A. Mamidipalli, T. Retson, N. Bahrami, K. Hasenstab, K. Blansit, E. Bass, T. Delgado, G. Cunha, M. S. Middleton, R. Loomba, B. A. Neuschwander-Tetri, C. B. Sirlin, A. Hsiao, Automated CT and MRI Liver Segmentation and Biometry Using a Generalized Convolutional Neural Network, Radiology: Artificial Intelligence 1 (2019) 180022.