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

Matthias S. Treder¹, Ryan Codrai¹, 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 Wasserstein 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 which consistently failed to fully reproduce human data. While the metrics Structural Similarity Index Measure (SSIM) and Naturalness Image Quality Evaluator (NIQE) showed good overall agreement with human assessment for lower-quality images (i.e. images from early stages of GAN training), only a Deep Quality Assessment (QA) model trained on human ratings was sensitive to the subtle differences

between higher-quality images.

Conclusions: We recommend a combination of group analyses, spatial correlation analyses, and both distortion metrics (SSIM, 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

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

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

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

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

72 Additionally, while human assessment and image quality metrics can
73 quantify the extent to which brain images 'look' realistic, we believe that they
74 provide limited information on their *biological plausibility*. Human observers
75 are sensitive to image artifacts such as checkerboard patterns but other dis-
76 tortions of brain images involve subtle statistical relationships across images
77 that may not be visible to the human eye. For instance, the discovery of
78 structural networks such as the Default Mode Network requires multivariate
79 statistical analyses such as spatial Independent Component Analysis (ICA)
80 performed across dozens or hundreds of MRIs [27]. Furthermore, establish-
81 ing group differences (e.g. young vs old, male vs female) requires group
82 statistical analysis. It is not clear whether the GAN learns to reproduce
83 such group differences and large-scale structural networks because it is not
84 explicitly trained to do so. We therefore believe that biological plausibility
85 should be investigated as an additional dimension of quality assessment using
86 a combination of group analysis and spatial ICA.

87 To summarize, the goal of this study was to provide ingredients for a sys-
88 tematic and efficient evaluation of generated brain images. Our contributions
89 are the following:

90 1. *Behavioral experiment.* We conducted the hitherto largest behavioral
91 study on generated MRIs with 26 neuroimaging experts assessing real and
92 generated grey-matter (GM) density maps from a noise-to-image Wasserstein
93 GAN. Participants performed a detection task (real vs generated images) and
94 a subjective quality rating task using a 5-points Likert scale. To determine
95 the experts' sensitivity to objective changes in image quality, we exported
96 images at five different stages of GAN training, from early stages (where
97 image quality was supposed to be poor) to later stages. We hypothesized
98 that both detection performance and quality ratings increase with training
99 time and ultimately approach the responses participants give for real images.

100 2. *Biological plausibility.* For the first time, we performed a detailed
101 analysis of the structural properties of the generated 3D MRIs by perform-
102 ing spatial Independent Component Analysis (ICA) to investigate structural
103 networks and group comparisons that show that GANs are sensitive to gen-
104 der and age differences. We hypothesized that structural networks and group

105 differences show a high degree of correspondence between real and generated
106 images.

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

119 **2. Method**

120 The Method section is split into four parts. In Section 2.1, we introduce
121 our generative modeling approach. In Section 2.2, we introduce the behavioral
122 experiment. In Section 2.3 we introduce spatial ICA and group analysis
123 performed in order to investigate the images' biological plausibility. In Section
124 2.4, we introduce various image quality metrics and a Deep QA as an
125 objective way of measuring the visual quality of the generated images. For
126 brevity, we refer to any type of brain image derived from an MR sequence
127 (e.g. T1, T2, grey-matter density maps derived from T1) as *MRI* in the rest
128 of the paper.

129 *2.1. Generative modeling of MRIs*

130 *2.1.1. Data*

131 Models were trained using the Cambridge Centre for Ageing and Neu-
132 roscience (Cam-CAN) data [28, 29]. A T1-weighted 3D-structural MRI was
133 acquired with the following parameters: repetition time (TR) = 2,250 ms;
134 echo time (TE) = 2.99 ms; inversion time (TI) = 900 ms; flip angle $\alpha = 9^\circ$;
135 field of view (FOV) = $256 \times 240 \times 192$ mm³; resolution = 1 mm isotropic;
136 accelerated factor = 2; acquisition time, 4 min and 32 s [29]. The T1 image
137 was initially coregistered to the MNI template, and the T2 image was then
138 coregistered to the T1 image using a rigid-body linear transformation. The
139 coregistered T1 and T2 images were used in a multichannel segmentation

140 to extract probabilistic maps of six tissue classes: gray matter (GM), white
141 matter (WM), cerebrospinal fluid (CSF), bone, soft tissue, and residual noise.
142 The native space GM and WM images were submitted to diffeomorphic reg-
143 istration (DARTEL; [30]) to create group template images. Each template
144 was normalized to the MNI template using a 12-parameter affine transfor-
145 mation. After applying the combined normalization parameters (native to
146 group template and group template to MNI template) to each individual
147 participant’s GM images, the normalized images were smoothed using an 8
148 mm Gaussian kernel. In total, GM images from 653 participants, aged 18-88
149 years, were considered as input to noise-to-brain GAN.

150 It is worth noting that GM maps were chosen over more common modal-
151 ities such as T1 since our main focus was on the ability of GANs to generate
152 brain structures. In contrast, non-segmented anatomical MR images also
153 involve signals from non-neural structures such as skull, face, and ears. Con-
154 sequently, it would be impossible to unambiguously attribute human rating
155 and image quality results to brain or non-brain structures. Moreover, spatial
156 brain networks (see Section 2.3) are typically defined using using GM-based
157 morphometry images rather than non-segmented T1 images [31].

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

169 2.1.2. Generative Adversarial Network

170 In a Generative Adversarial Network (GAN) two neural networks are
171 pitted against each other in a zero-sum game [6]. In this framework, the
172 generator G acts as the adversary of the discriminator D . Whereas G aims
173 to learn the generative distribution of the training data p_{real} by approxim-
174 ating it with a generative distribution p_{gen} , D seeks to accurately predict the
175 probability that an image is generated rather than real. The discriminator
176 is a function $D : \mathbb{R}^{96 \times 112 \times 96} \rightarrow [0, 1]$, $x \mapsto D(x)$ that takes an image as input

177 and returns the probability that this image is real. The generator is a func-
 178 tion $G : \mathbb{R}^{100} \rightarrow \mathbb{R}^{96 \times 112 \times 96}$, $z \mapsto G(z)$ that takes a random vector z sampled
 179 from a probability distribution p_z over $\mathbb{R}^{100}$ and returns an MRI. Training
 180 is designed to maximize the quality of the generated images by learning the
 181 generative distribution p_{real} . The model training reaches optimality when D
 182 is unable to discriminate between real and generated samples. The objective
 183 of the discriminator is governed by the equation

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

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

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

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

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

192 Since G learns a mapping from a low dimensional vector space to a high
 193 dimensional output, the space from which the input vector originates is a rep-
 194 resentation of the low-dimensional manifold of MRIs. This quantity is a lower
 195 bound of the Jensen-Shannon divergence between the real and generated dis-
 196 tributions. GAN training using this objective function can fail to converge
 197 due to near-zero gradients when the true and generated distribution do not
 198 overlap and also. A more stable alternative that provides useful gradients
 199 for non-overlapping distributions is given by Wasserstein GANs (WGANs)
 200 [32, 33]. To this end, the objective function is modified to the Earth Mover
 201 or Wasserstein-1 distance between p_{real} and p_{gen} given by

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.

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

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 [32, 33].

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 batch numbers: 344, 1055, 7954, 24440, and 60000 (final batch).

The experiment was conducted online using PsyToolkit [34, 35] (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 per batch. 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 batch. 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 [36]. For the subjective rating task, Mean Opinion Scores (MOS) were calculated by averaging across all images in a batch. 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 [37]. To investigate age effects, we split the data into elderly (> 70 years; 172 MRIs) and young (< 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 to decompose the synthetic data in a small set of spatially independent maps that correspond sensibly with known neurobiological relationships in the real data [31]. 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 [38] in the Group ICA for fMRI Toolbox (GIFT)¹. By combining the 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>

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

317 We first calculated four metrics widely adopted in the GAN literature,
318 namely Inception Score (IS), Modified Inception Score (MIS), Fréchet Incep-
319 tion Distance (FID), Maximum Mean Discrepancy (MMD) [39]. In addi-
320 tion, we used an array of metrics developed by the image/video processing
321 community that are supposed to approximate human quality assessments
322 [26]. We tested six among the most widely adopted metrics in the image
323 quality assessment field, Peak signal to noise ratio (PSNR), Structural Sim-
324 ilarity Index (SSIM), Structural Similarity Index (MS-SSIM), Video Mul-
325 timethod Assessment Fusion (VMAF), Natural Image Quality Evaluator
326 (NIQE), Blind/referenceless image spatial quality evaluator (BRISQUE); see
327 [40] for a review). In the following, each of these metrics is introduced in
328 more detail. As before, p_{real} and p_{gen} are used to denote the distributions of
329 the real and generated images. Furthermore, $p(y)$ denotes the distribution
330 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 [41]. 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

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

331 where $\mathbb{E}$ is the expectation, $x \sim p_{\text{gen}}$ are the images sampled from the
332 generator, KL refers to the Kullback-Leibler divergence, and $p(y)$ is
333 the marginal distribution of class labels. IS favours generated images
334 that show a clear class membership, characterized by conditional class
335 labels $p(y|x)$ with a 'peaky' (low entropy) distribution, at the same
336 time favouring a broad coverage of multiple classes characterized by a
337 'flat' (high entropy) marginal distribution $p(y)$. A detailed discussion
338 of IS is provided in [42].

- *MIS*, unlike IS, takes into account the desire for diversity of images within a given class. To achieve this, [43] 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

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

- *FID* uses the feature embeddings of a CNN wherein it models the distributions of real and generated images as multivariate Gaussians [44]. 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. [45] 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 $k(x, x') = \exp(-\frac{1}{2}\|x - x'\|^2)$.

- *PSNR* is a simple metrics based on the pixel-wise difference between a target image x and a reference image y ,

$$\text{PSNR}(x, y) = 10 \log_{10} \left(\frac{L^2}{\text{MSE}(x, y)} \right)$$

where MSE is the mean-squared error between target and reference and L^2 is the maximum pixel value (e.g. 1.0 for data normalized in the 0-1 range).

- *SSIM* quantifies similarity using a product of three separate terms for luminance, contrast, and structure similarity [26]. The product can be represented as

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + k_1)(2\sigma_{xy} + k_2)}{(\mu_x^2 + \mu_y^2 + k_1)(\sigma_x^2 + \sigma_y^2 + k_2)}$$

where μ_x, μ_y are the mean pixel values of the target and reference images, σ_x, σ_y are their variances and σ_{xy} is the covariance, and k_1, k_2 are

regularization hyperparameters used for numerical stability. We computed SSIM using TensorFlow’s implementation with default hyperparameters (11x11 Gaussian filter with width 1.5, $k_1 = 0.1$, $k_2 = 0.3$).

- *Multi-scale Structural Similarity Index (MS-SSIM)* is an extension of SSIM that aims towards scale invariance [46]. The image is analyzed at its original scale and at multiple low-pass filtered and downscaled versions of it. The SSIM values from the different scales are then combined into a single estimate. We used the VMAF command line tool to calculate MS-SSIM.
- *VMAF* was originally developed for video quality assessment [47]. The VMAF model has been trained on a Netflix quality assessment dataset wherein viewers scored videos on a 5-points Likert scale summarized by Mean Opinion Scores (MOS) and normalized to the 0-100 range. We used the VMAF command line tool (github.com/Netflix/vmaf). Images were padded to 480 x 360 px and then exported to yuv format using FFmpeg (www.ffmpeg.org).
- *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 `nique` and `fitnique` 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 [48]. 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

378 unbiased scores for the images. Matlab’s `brisque` and `fitbrisque`
379 functions were used.

380 Note that IS, MIS, NIQE and BRISQUE are reference free metrics. There-
381 fore, we could calculate scores for both generated and real images. For the
382 other metrics, we used generated images from each of the batches as target
383 images and the real MRIs as reference images. FID and MMD only require
384 a distribution of generated images. However, PSNR, SSIM, MS-SSIM, and
385 VMAF are full reference metrics, that is, for each generated image they re-
386 quire a matched real image. Since there was no matching between generated
387 and real images, we considered all possible pairs of real/generated images.
388 Another distinction between the metrics is that IS, MIS, PSNR, SSIM, MS-
389 SSIM, and VMAF are quality metrics (better quality = higher value) whereas
390 FID, MMD, NIQE and BRISQUE are distance metrics (better quality =
391 lower value).

392 2.4.2. Deep QA model

393 An alternative to image-based metrics is to train a model that tries to
394 mimick human perceptual assessments. Such perceptual models have been
395 used both for MRIs and in the wider image/video quality literature [49, 50].
396 We implemented two CNNs that essentially performed the same task as the
397 human experts, namely classifying images as ‘real’ or ‘fake’ (detection task)
398 and assigning a subjective rating to an image on a 5-points Likert scale
399 (rating task). The model architecture is depicted in Figure 6a.

400 *Detection task model.* The first part of the model consisted of a Incep-
401 tionV3 model pretrained on ImageNet with the classification layer removed.
402 The InceptionV3 output was fed into two dense layers (32 and 16 units,
403 dropout rate 0.1, leaky ReLU activation with $\alpha = 0.1$), denoted as MRI
404 features. The final layer was a single sigmoid unit. The model was trained
405 on a set of 3611 real and generated images not used in the behavioral ex-
406 periment. Images from all batches were pooled and labeled as ‘fake’. The
407 model was initially trained for 5 epochs with the InceptionV3 model frozen.
408 Subsequently, the last two convolutional layers of InceptionV3 were unfrozen
409 and training continued for another 20 epochs. Binary cross-entropy was used
410 as loss function with an Adam optimizer, a learning rate of 10^{-4} (reduced to
411 10^{-5} after 5 epochs), and a batch size of 16. Finally, the model was tested
412 on the 240 images used in the detection task and the predicted logits were
413 recorded for each image. No behavioral data was used.

414 *Rating task model.* The challenge of training a model on the rating task
 415 was the small number of only 30 images. Therefore, we explored a transfer
 416 learning approach wherein we first trained the detection task model and then
 417 fine-tuned it on the rating task data. To this end, we took the detection task
 418 model and replaced the sigmoid by a softmax layer with 5 units represent-
 419 ing the Likert scale ratings. The 30 rating task images were split into train
 420 and test sets using 5-fold cross-validation. For the training images, class la-
 421 bels (representing ratings) could not be assigned unequivocally since there
 422 were disagreements between the human raters. To encode this uncertainty in
 423 the model we used probabilistic class labels: For each image, the empirical
 424 distribution of ratings was determined across raters and used to initialize a
 425 probability distribution. In each batch, class labels were randomly sampled
 426 from these distributions. The model was initially trained for 20 epochs with
 427 the InceptionV3 model and the MRI features frozen. After this, the MRI
 428 features were unfrozen and model was trained for another 200 epochs. Cat-
 429 egorical cross-entropy was used as loss function with an Adam optimizer,
 430 a learning rate of 10^{-4} (reduced to 10^{-5} after 20 epochs), and batch size
 431 6. Both models were trained 100 times with randomly initialized weights.
 432 Results were averaged across runs.

433 2.4.3. Hardware and software

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

439 3. Results

440 3.1. Qualitative analysis of generated MRIs

441 Figure 2 shows a 3D rendering of the generated images using MRICroGL
 442 [51]. To indicate how the images evolve across the different stages of training,
 443 Figure 2a shows images for six logarithmically spaced batch numbers ranging
 444 from an early batch (batch 11) to the final batch (batch 60000). Each of
 445 the six renderings is based on the same random vector. The GAN initially
 446 outputs a block of noise (batches 11 and 90) from which a brain is eventually
 447 carved out (batches 344 and 843). After this, GAN training seems to focus
 448 on improving details and removing noise voxels outside the brain volume

Figure 2: Qualitative results showing a MRicroGL rendering of the generated 3D images. (a) Generated images at six different times (batch numbers) during GAN training. Whereas only a volume of noise is generated in the initial stages (batches 11 and 90), the basic brain structure is evident from batch 843. Much of the subsequent training appears devoted to refining details and removing the noise outside the brain. Batch 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.

(compare batches 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 S2 (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 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 batches. 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

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 batch number. 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-batch similarity, average correlation between pairs of images within a batch. 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. (f) 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.

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 batch number, suggesting that the GAN was sensitive to these group characteristics. In the final batch, 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 batch number, 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 batch, 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 batch number, approaching the RT for real images

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

(Figure 4a and d). For the detection task, the percentage of 'real' responses increased with the batch number, 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 batch number 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 batch number) 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 batch 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 batch ($\beta = 6.184, z = 9.343, p < 0.001$) 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 batch 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 batch ($\beta = 279.875, z = 11.377, p < 0.001$). For the analysis of the ratings, a linear mixed model approach would not be 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 batch as fixed effect and participant as a random effect with random intercept. We found a significant positive slope for log batch ($\beta = 1.234, z = 17.94, p < 0.001$) signifying higher ratings for later batches. The same statistical approach was applied to compare ratings for batch 60000 vs real MRIs, showing significantly higher ratings for

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

561 real MRIs as compared to generated ones ($\beta = 2.133, z = 7.748, p < 0.0001$).

562 3.4. Image quality metrics and Deep QA model

563 Figure 5 shows the metrics calculated for the detection task images. An
 564 analogous analysis applied to the rating task images is depicted in Figure S3
 565 (Supplementary Material). Figure 6b shows the corresponding result for the
 566 Deep QA model. Linear regression analyses of log batch on the metric were
 567 conducted to investigate whether the metric increases/decreases with batch
 568 number.

569 For the detection data, all regression slopes were significant (all p-values
 570 < 0.0001 except for MIS $p = 0.0009$) except for BRISQUE-MRI which gave
 571 a constant output and was discarded from further analysis. IS and MIS both
 572 decreased with batch number. Since they are quality metrics, the opposite
 573 pattern was expected. FID and MMD both decreased with batch number
 574 but images at batch 24400 were predicted to have a better quality than

images at batch 60000 (Wilcoxon Rank Sum, FID $z = -12.217, p < 0.001$;
MMD $z = -12.217, p < 0.001$), conflicting with the behavioral data. PSNR,
SSIM, MS-SSIM, and VMAF increased with batch number. Of these, only
VMAF increased monotonically, in line with the behavioral data, and SSIM
increased monotonically except for the last batch. However, in neither case
was the difference between batches 24400 and 60000 significant (Wilcoxon
Rank Sum test, SSIM $p = 0.493$; VMAF $p = 0.282$). Although NIQE(-
MRI) and BRISQUE(-MRI) showed overall regression trends in line with
behavioral data, the pattern for batches 7954, 24400 and 60000 conflicted
with human data. NIQE predicted a lower quality for the 60000 batch than
for 24400 ($z = -2.611, p = 0.009$). Except for VMAF, the Deep QA model
was the only model for which predictions on the detection data changed
monotonically with batch number, although the difference between batches
24400 and 60000 was not significant ($p = 0.51$).

We repeated the same analysis on the rating task data (Figure S3 and
Figure 6b). Except for IS ($p = 0.4424$), MIS ($p = 0.1943$), and BRISQUE-
MRI ($p = 1.0$), all regression slopes were significant ($p < 0.001$ except PSNR
 $p = 0.0023$). Again, VMAF increased monotonically and SSIM increased
monotonically except for the last batch. However, in neither case was the
difference between batches 24400 and 60000 significant (Wilcoxon Rank Sum
test, SSIM $p = 0.4647$, VMAF $p = 0.347$). Deep QA was the only metric that
showed a significant difference between these two batches ($z = -2.193, p =$
 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 could not be subjected to correlation analysis because they esti-
mate the distance between distributions rather than individual images. Since
some metrics did not provide separate predictions for real images, we con-
ducted two correlation analyses, one analysis that covered only generated
images and another one that included real images (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 S3).

Except for IS, MIS, BRISQUE, and BRISQUE-MRI, all metrics showed
moderate to high correlation with the behavioral data. For the detection

task, SSIM ($\rho = 0.754$), NIQE ($\rho = -0.706$), and NIQE-MRI ($\rho = -0.536$) were the three best performing metrics when the real data was excluded. When real images were included, the best performing metrics were Deep QA ($\rho = 0.831$), NIQE-MRI ($\rho = -0.761$) and NIQE ($\rho = -0.519$). Despite the fact that the Deep QA model performed the classification task better than humans (100% accuracy), it assigned higher probabilities to later batches 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, SSIM ($\rho = 0.925$), Deep QA ($\rho = -0.879$), and NIQE ($\rho = -0.767$) were the best performing metrics when the real data was excluded. When real images were included, the best performing metrics were Deep QA ($\rho = 0.912$), NIQE-MRI ($\rho = -0.722$) and NIQE ($\rho = -0.614$).

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 three metrics with the largest absolute correlation are highlighted in bold.

Task		IS	MIS	PSNR	SSIM	MS-SSIM	VMAF	NIQE	NIQE-MRI	BRISQUE	BRISQUE-MRI	DeepQA
Detection	ρ	-0.250	-0.016	0.421	0.754	0.459	0.404	-0.706	-0.536	0.465	NaN	0.516
	p-value	0.001	0.845	0.000	0.000	0.000	0.000	0.000	0.000	0.000	NaN	0.000
Detection (include real)	ρ	0.019	0.241	-	-	-	-	-0.519	-0.761	-0.013	NaN	0.831
	p-value	0.767	0.000	-	-	-	-	0.000	0.000	0.840	NaN	0.000
Rating	ρ	-0.206	0.112	0.705	0.925	0.657	0.504	-0.767	-0.575	0.521	-0.067	0.879
	p-value	0.323	0.593	0.000	0.000	0.000	0.010	0.000	0.003	0.008	0.751	0.000
Rating (include real)	ρ	-0.065	0.104	-	-	-	-	-0.614	-0.722	0.294	-0.052	0.912
	p-value	0.733	0.585	-	-	-	-	0.000	0.000	0.114	0.786	0.000

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

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.

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 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, 31]. This correlation

671 between the spatial maps was high irrespective of the exact choice of the
672 number of ICs.

673 *4.2. Image quality metrics as a proxy for human assessment*

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

687 To relate behavioral data to the image quality metrics, we calculated
688 Spearman's rank correlation between quality metrics and the proportion of
689 'real' responses in the behavioral data. The standard GAN metrics IS and
690 MIS showed poor correspondence with human assessment, in line with earlier
691 research [22], and only few metrics agreed well with behavior. Both NIQE
692 (detection task $\rho = -0.706$; rating task $\rho = -0.767$) and SSIM (detection
693 $\rho = 0.754$; rating $\rho = 0.925$) showed a good agreement. However, SSIM
694 did not show a significant difference for batches 24400 and 60000, whereas
695 NIQE even predicted the 60000 batch to be of lower quality. Our Deep QA
696 models showed a moderate correlation with human data in the detection
697 task ($\rho = 0.516$, rising to $\rho = 0.831$ when including real images) and a
698 high correlation in the rating task ($\rho = 0.879$ and $\rho = 0.912$ when including
699 real images). It is worth noting that only the rating model was explicitly
700 trained on human responses, whereas the detection model was only trained
701 to discriminate between real and generated images. Crucially, the Deep QA
702 rating model was the only metric that showed a significant increase in quality
703 ratings from the 24400 to the 60000 batch.

704 An interesting conclusion of this analysis is that most image quality met-
705 rics correlate with human behavior reasonably well for lower batch numbers
706 (batches 344, 1055, 7954) but are less sensitive to image differences later in
707 training (batches 24400 vs 60000). We believe this discrepancy might be due

708 to most metrics being sensitive to *distortion* rather than *perceptual quality*,
 709 two different dimensions of image quality that have been explored theoreti-
 710 cally in image restoration tasks [56]. *Distortion* refers to the deviation of a
 711 distorted image from a noise-free reference whereas *perceptual quality* mea-
 712 sures the more elusive ‘naturalness’ of an image regardless of a reference. In
 713 our experiment, informal evidence in line with this is a participant stating²
 714 that they focused on image artifacts such as checkerboard patterns for low-
 715 quality images (i.e., distortion), but that it was rather subtle differences in
 716 luminance across the image would that made higher-quality generated images
 717 look fake (i.e., perceptual quality).

718 If human assessment of image quality was indeed governed by distortion
 719 and artifacts for earlier batches and a more intuitive understanding of what
 720 an MRI should look like for later batches, it is reasonable to believe that
 721 most of the metrics were sensitive to artifacts, hence explaining their good
 722 performance for early batches and bad performance for the last batch. The
 723 only model that consistently agreed with human data across all quality levels
 724 was the Deep QA model. The model was directly trained on the images and
 725 human responses which gave it the opportunity to learn the more subtle
 726 features that humans use for more high-quality generated images.

727 4.3. The limits of GANs

728 The promise of GANs is somewhat uncanny: After being exposed to a few
 729 hundred images it supposedly learns the brain image manifold in image space.
 730 A key question therefore is whether the GAN indeed learns (a) the generative
 731 distribution of MRIs, whether it learns (b) a different distribution, or (c)
 732 simply memorizes and reproduces the training data. Regarding (c), it turns
 733 out that memorization of data requires a large number of different samples
 734 [57]. Other indicators contra memorization are the fact that interpolation
 735 between noise vectors can produce novel and meaningful image variations
 736 [58] and the clear disparity seen between real and generated images [59].

737 In accordance with (a), GANs can closely approximate the underlying
 738 generative distribution for sufficiently large datasets [6]. Generalization be-
 739 yond the training data, indicated by the emergence of novel combinations of
 740 features, has been shown in simple datasets [60]. However, there is doubt

²the participant disclosed their participation in the experiment to the author in a personal email

that the same holds for limited sample sizes [59]. 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. [61] 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 [61] 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 [62]. 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. Furthermore, PSNR can directly operate on 3D images and for VMAF the third spatial dimension can be encoded as a temporal dimension, although this constitutes a ‘hack’ of the metric. 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

The usage of a specific model (noise-to-image GAN) with a specific type of data (GM maps) leads to the question whether our results are relevant to other model architectures (e.g. VAEs), hyperparameter choices (e.g. number of convolutional layers), or imaging modalities. We believe they are, since all CNN-based models use the same computational building blocks such as upscaling operations and convolutional layers that produce well-known artifacts known as checkerboard patterns [63, 49]. With regard to imaging modalities, we believe that our results for GM maps are relevant to other types of images such as T1 and T2-weighted. For instance, [64] developed a multi-modal CNN that performs liver segmentation in both MRI and CT images, and [65] reported successful transfer learning from structural MRIs to Diffusion Tensor Imaging (DTI) data. Both studies suggest that there different imaging modalities share visual commonalities.

Furthermore, 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, [66] 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 such as SSIM.

4.6. Conclusion

As a practical recommendation for evaluating brain images produced by GANs, we propose that researchers assess both local properties (visual quality of individual images) and global properties (statistical regularities across the 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 [67] 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 [56]. As a distortion metric, SSIM stood out as a versatile, widely used image quality metric that does not require tuning to the MRI data [26]. It was sensitive to a variety of image

distortions and can be therefore be expected to perform similarly in other types of imaging datasets. It is also widely available on different platforms (e.g. `ssim` in Matlab and TensorFlow). A similar performance was found for NIQE (`niqe` in Matlab). Either of them or both metrics could be reported as distortion measures. 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 pretrain a Deep QA model on a large, diverse dataset that integrates multiple data modalities [65, 64]. 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, collected human data and performed the statistical analyses. RC trained the GAN. MT and KT wrote 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.

- 881 [2] H. Peng, W. Gong, C. F. Beckmann, A. Vedaldi, S. M. Smith, Accurate
882 brain age prediction with lightweight deep neural networks, *Medical*
883 *Image Analysis* 68 (2021) 101871.
- 884 [3] M. Sajjad, S. Khan, K. Muhammad, W. Wu, A. Ullah, S. W. Baik,
885 Multi-grade brain tumor classification using deep CNN with extensive
886 data augmentation, *Journal of Computational Science* 30 (2019) 174–
887 182.
- 888 [4] G. Mohan, M. M. Subashini, MRI based medical image analysis: Survey
889 on brain tumor grade classification, *Biomedical Signal Processing and*
890 *Control* 39 (2018) 139–161.
- 891 [5] V. Sorin, Y. Barash, E. Konen, E. Klang, Creating Artificial Images for
892 Radiology Applications Using Generative Adversarial Networks (GANs)
893 – A Systematic Review, *Academic Radiology* 27 (2020) 1175–1185.
- 894 [6] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley,
895 S. Ozair, A. Courville, Y. Bengio, Generative Adversarial Nets, in: Z. G.
896 Weinberger, M. Welling, C. Cortes, N. D. Lawrence, K. Q. (Eds.), *Ad-*
897 *vances in Neural Information Processing Systems* 27, Curran Associates,
898 Inc., 2014, pp. 2672–2680.
- 899 [7] D. P. Kingma, M. Welling, Auto-encoding variational bayes, in: 2nd In-
900 ternational Conference on Learning Representations, ICLR 2014 - Con-
901 ference Track Proceedings, International Conference on Learning Rep-
902 resentations, ICLR, 2014.
- 903 [8] A. U. Hirte, M. Platscher, T. Joyce, J. J. Heit, E. Tranvinh, C. Fed-
904 erau, Diffusion-Weighted Magnetic Resonance Brain Images Generation
905 with Generative Adversarial Networks and Variational Autoencoders: A
906 Comparison Study, *arXiv* (2020) 2006.13944.
- 907 [9] C. K. Chong, E. T. W. Ho, Synthesis of 3D MRI Brain Images with
908 Shape and Texture Generative Adversarial Deep Neural Networks, *IEEE*
909 *Access* 9 (2021) 64747–64760.
- 910 [10] S. U. H. Dar, M. Yurt, L. Karacan, A. Erdem, E. Erdem, T. Çukur,
911 Image Synthesis in Multi-Contrast MRI with Conditional Generative
912 Adversarial Networks, *IEEE Transactions on Medical Imaging* 38 (2019)
913 2375–2388.

- 914 [11] D. Nie, R. Trullo, J. Lian, C. Petitjean, S. Ruan, Q. Wang, D. Shen,
915 Medical Image Synthesis with Context-Aware Generative Adversarial
916 Networks, in: Medical Image Computing and Computer Assisted In-
917 tervention - MICCAI 2017. MICCAI 2017. Lecture Notes in Computer
918 Science, Springer, Cham, 2017, pp. 417–425.
- 919 [12] D. Abramian, A. Eklund, Generating fMRI volumes from T1-weighted
920 volumes using 3D CycleGAN, arXiv abs/1907.0 (2019).
- 921 [13] X. Gu, H. Knutsson, M. Nilsson, A. Eklund, Generating Diffusion MRI
922 Scalar Maps from T1 Weighted Images Using Generative Adversarial
923 Networks, Lecture Notes in Computer Science 11482 (2019) 489–498.
- 924 [14] Y. Chen, F. Shi, A. G. Christodoulou, Y. Xie, Z. Zhou, D. Li, Efficient
925 and Accurate MRI Super-Resolution Using a Generative Adversarial
926 Network and 3D Multi-level Densely Connected Network, in: Medical
927 Image Computing and Computer Assisted Intervention – MICCAI 2018,
928 Springer, Cham, 2018, pp. 91–99.
- 929 [15] T. M. Quan, T. Nguyen-Duc, W.-K. Jeong, Compressed Sensing MRI
930 Reconstruction Using a Generative Adversarial Network With a Cyclic
931 Loss, IEEE Transactions on Medical Imaging 37 (2018) 1488–1497.
- 932 [16] X. Yi, E. Walia, P. Babyn, Generative adversarial network in medical
933 imaging: A review, arXiv e-prints 1809.07294 (2018).
- 934 [17] C. Bermudez, A. J. Plassard, T. L. Davis, A. T. Newton, S. M. Resnick,
935 B. A. Landman, Learning Implicit Brain MRI Manifolds with Deep
936 Learning, Proceedings of SPIE—the International Society for Optical
937 Engineering 10574 (2018).
- 938 [18] K. Kazuhiro, R. A. Werner, F. Toriumi, M. S. Javadi, M. G. Pomper,
939 L. B. Solnes, F. Verde, T. Higuchi, S. P. Rowe, Generative Adversarial
940 Networks for the Creation of Realistic Artificial Brain Magnetic Reso-
941 nance Images, Tomography (Ann Arbor, Mich.) 4 (2018) 159–163.
- 942 [19] A. Volokitin, E. Erdil, N. Karani, K. C. Tezcan, X. Chen, L. Van Gool,
943 E. Konukoglu, Modelling the Distribution of 3D Brain MRI Using a
944 2D Slice VAE, in: Lecture Notes in Computer Science, volume 12267
945 LNCS, Springer Science and Business Media Deutschland GmbH, 2020,
946 pp. 657–666.

- 947 [20] G. Kwon, C. Han, D.-s. Kim, Generation of 3D Brain MRI Using Auto-
948 Encoding Generative Adversarial Networks, in: MICCAI 2019: Medical
949 Image Computing and Computer Assisted Intervention, pp. 118–126.
- 950 [21] C. Han, L. Rundo, K. Murao, T. Noguchi, Y. Shimahara, Z. A. Milacski,
951 S. Koshino, E. Sala, H. Nakayama, S. Satoh, MADGAN: unsupervised
952 medical anomaly detection GAN using multiple adjacent brain MRI slice
953 reconstruction, BMC Bioinformatics 2021 22:2 22 (2021) 1–20.
- 954 [22] F. Calimeri, A. Marzullo, C. Stamile, G. Terracina, Biomedical Data
955 Augmentation Using Generative Adversarial Neural Networks, in:
956 P. Verschure, A. Villa, A. Lintas, S. Rovetta (Eds.), Artificial Neural
957 Networks and Machine Learning - ICANN 2017. ICANN 2017. Lecture
958 Notes in Computer Science, Springer, Cham, 2017, pp. 626–634.
- 959 [23] C. Han, L. Rundo, K. Murao, Z. A. Milacski, K. Umemoto,
960 H. Nakayama, S. Satoh, GAN-based Multiple Adjacent Brain MRI
961 Slice Reconstruction for Unsupervised Alzheimer’s Disease Diagnosis,
962 in: In International Meeting on Computational Intelligence Methods for
963 Bioinformatics and Biostatistics. Lecture Notes in Computer Science,
964 Springer, Cham, 2019, pp. 44–54.
- 965 [24] K. H. Kim, W. Do, S. Park, Improving resolution of MR images with an
966 adversarial network incorporating images with different contrast, Med-
967 ical Physics 45 (2018) 3120–3131.
- 968 [25] D. Yang, B. Liu, L. Axel, D. Metaxas, 3D LV Probabilistic Segmentation
969 in Cardiac MRI Using Generative Adversarial Network, in: Statistical
970 Atlases and Computational Models of the Heart. Atrial Segmentation
971 and LV Quantification Challenges, Springer, Cham, 2019, pp. 181–190.
- 972 [26] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality
973 assessment: From error visibility to structural similarity, IEEE Trans-
974 actions on Image Processing 13 (2004) 600–612.
- 975 [27] N. Luo, J. Sui, A. Abrol, J. Chen, J. A. Turner, E. Damaraju, Z. Fu,
976 L. Fan, D. Lin, C. Zhuo, Y. Xu, D. C. Glahn, A. L. Rodrigue, M. T.
977 Banich, G. D. Pearlson, V. D. Calhoun, Structural brain networks match
978 intrinsic functional networks and vary across domains: a study from
979 15000+ individuals, Cerebral Cortex 30 (2020) 5460–5470.

- 980 [28] M. Shafto, L. K. Tyler, M. Dixon, J. R. Taylor, J. B. Rowe, R. Cusack,
981 A. J. Calder, W. D. Marslen-Wilson, J. Duncan, T. Dalgleish, R. N.
982 Henson, C. Brayne, F. E. Matthews, The Cambridge Centre for Ageing
983 and Neuroscience (Cam-CAN) study protocol: a cross-sectional, lifes-
984 pan, multidisciplinary examination of healthy cognitive ageing, *BMC*
985 *neurology* 14 (2014) 204.
- 986 [29] J. R. Taylor, N. Williams, R. Cusack, T. Auer, M. A. Shafto, M. Dixon,
987 L. K. Tyler, Cam-CAN, R. N. Henson, The Cambridge Centre for Age-
988 ing and Neuroscience (Cam-CAN) data repository: Structural and func-
989 tional MRI, MEG, and cognitive data from a cross-sectional adult lifes-
990 pan sample, *NeuroImage* 144 (2017) 262–269.
- 991 [30] J. Ashburner, A fast diffeomorphic image registration algorithm, *Neu-*
992 *roImage* 38 (2007) 95–113.
- 993 [31] K. A. Tsvetanov, R. N. A. Henson, P. S. Jones, H. Mutsaerts,
994 D. Fuhrmann, L. K. Tyler, J. B. Rowe, The effects of age on resting-state
995 BOLD signal variability is explained by cardiovascular and cerebrovas-
996 cular factors, *Psychophysiology* 58 (2021) e13714.
- 997 [32] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein Generative Adversar-
998 ial Networks, in: *Proceedings of the 34th International Conference on*
999 *Machine Learning*, pp. 214–223.
- 1000 [33] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. C. Courville, Im-
1001 proved Training of Wasserstein GANs, in: *Proceedings of the 31st Inter-*
1002 *national Conference on Neural Information Processing Systems*, Curran
1003 Associates Inc., Long Beach, California, USA, 2017, p. 5769–5779.
- 1004 [34] G. Stoet, PsyToolkit: A software package for programming psycholog-
1005 ical experiments using Linux, *Behavior Research Methods* 42 (2010)
1006 1096–1104.
- 1007 [35] G. Stoet, PsyToolkit: A Novel Web-Based Method for Running Online
1008 Questionnaires and Reaction-Time Experiments, *Teaching of Psychol-*
1009 *ogy* 44 (2017) 24–31.
- 1010 [36] R. Whelan, Effective analysis of reaction time data, *Psychological*
1011 *Record* 58 (2008) 475–482.

- 1012 [37] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger,
1013 H. Greenspan, GAN-based synthetic medical image augmentation for
1014 increased CNN performance in liver lesion classification, *Neurocomputing* 321 (2018) 321–331.
1015
- 1016 [38] L. Xu, K. M. Groth, G. Pearlson, D. J. Schretlen, V. D. Calhoun, Source-
1017 based morphometry: The use of independent component analysis to
1018 identify gray matter differences with application to schizophrenia, *Human Brain Mapping* 30 (2009) 711–724.
1019
- 1020 [39] A. Borji, Pros and cons of GAN evaluation measures, *Computer Vision*
1021 *and Image Understanding* 179 (2019) 41–65.
- 1022 [40] C. Lee, S. Woo, S. Baek, J. Han, J. Chae, J. Rim, Comparison of
1023 objective quality models for adaptive bit-streaming services, in: *2017*
1024 *8th International Conference on Information, Intelligence, Systems &*
1025 *Applications (IISA)*, volume 2017, Institute of Electrical and Electronics
1026 Engineers Inc., 2017, pp. 1–4.
- 1027 [41] J. Jia Deng, W. Wei Dong, R. Socher, L.-J. Li-Jia Li, K. Kai Li, L. Li Fei-
1028 Fei, ImageNet: A large-scale hierarchical image database, in: *2009 IEEE*
1029 *Conference on Computer Vision and Pattern Recognition, IEEE*, 2009,
1030 pp. 248–255.
- 1031 [42] S. Barratt, R. Sharma, A Note on the Inception Score, *arXiv* (2018)
1032 1801.01973.
- 1033 [43] S. Gurumurthy, R. K. Sarvadevabhatla, V. B. Radhakrishnan, DeLi-
1034 GAN : Generative Adversarial Networks for Diverse and Limited Data,
1035 *Proceedings - 30th IEEE Conference on Computer Vision and Pattern*
1036 *Recognition, CVPR 2017 2017-Janua* (2017) 4941–4949.
- 1037 [44] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter,
1038 GANs Trained by a Two Time-Scale Update Rule Converge to a Local
1039 Nash Equilibrium, *Advances in Neural Information Processing Systems*
1040 *2017-Decem* (2017) 6627–6638.
- 1041 [45] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, A. Smola,
1042 A Kernel Two-Sample Test, *Journal of Machine Learning Research* 13
1043 (2012) 723–773.

- 1044 [46] Z. Wang, E. P. Simoncelli, A. C. Bovik, Multi-scale structural similarity
1045 for image quality assessment, in: Conference Record of the Asilomar
1046 Conference on Signals, Systems and Computers, volume 2, pp. 1398–
1047 1402.
- 1048 [47] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, M. Manohara, Toward
1049 A Practical Perceptual Video Quality Metric, Netflix Technology Blog
1050 (2016) Jun 6.
- 1051 [48] A. Mittal, A. Krishna Moorthy, A. Conrad Bovik, No-Reference Image
1052 Quality Assessment in the Spatial Domain, IEEE Transactions on Image
1053 Processing 21 (2012) 4695–4708.
- 1054 [49] G. Yang, Y. Cao, X. Xing, M. Wei, Perceptual Loss Based Super-
1055 Resolution Reconstruction from Single Magnetic Resonance Imaging,
1056 in: Lecture Notes in Computer Science, volume 11632, Springer Verlag,
1057 2019, pp. 411–424.
- 1058 [50] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, O. Wang, The Un-
1059 reasonable Effectiveness of Deep Features as a Perceptual Metric, in:
1060 Proceedings of the IEEE Conference on Computer Vision and Pattern
1061 Recognition (CVPR), pp. 586–595.
- 1062 [51] C. Rorden, M. Brett, Stereotaxic display of brain lesions, Behavioural
1063 Neurology 12 (2000) 191–200.
- 1064 [52] D. A. Magezi, Linear mixed-effects models for within-participant psy-
1065 chology experiments: an introductory tutorial and free, graphical user
1066 interface (LMMgui), Frontiers in Psychology 6 (2015) 2.
- 1067 [53] C. Keeble, P. D. Baxter, A. J. Gislason-Lee, L. A. Treadgold, A. G.
1068 Davies, Methods for the analysis of ordinal response data in medical
1069 image quality assessment, British Journal of Radiology 89 (2016).
- 1070 [54] R. H. B. Christensen, ordinal—Regression Models for Ordinal Data,
1071 2019.
- 1072 [55] K. Liu, S. Yao, K. Chen, J. Zhang, L. Yao, K. Li, Z. Jin, X. Guo,
1073 Structural Brain Network Changes across the Adult Lifespan, Frontiers
1074 in Aging Neuroscience 9 (2017) 275.

- 1075 [56] Y. Blau, T. Michaeli, The Perception-Distortion Tradeoff, in: Proceed-
1076 ings of the IEEE Conference on Computer Vision and Pattern Recogni-
1077 tion (CVPR), pp. 6228–6237.
- 1078 [57] V. Nagarajan, C. Raffel, G. Brain, I. J. Goodfellow Google Brain, The-
1079 oretical Insights into Memorization in GANs, in: 32nd Conference on
1080 Neural Information Processing Systems.
- 1081 [58] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning
1082 with deep convolutional generative adversarial networks, in: 4th Inter-
1083 national Conference on Learning Representations, ICLR 2016 - Confer-
1084 ence Track Proceedings, International Conference on Learning Repre-
1085 sentations, ICLR, 2016.
- 1086 [59] S. Arora, Y. Zhang, Do GANs actually learn the distribution? An
1087 empirical study, arXiv (2017) 1706.08224.
- 1088 [60] S. Zhao, H. Ren, A. Yuan, J. Song, N. Goodman, S. Ermon, Bias
1089 and Generalization in Deep Generative Models: An Empirical Study,
1090 Advances in Neural Information Processing Systems 2018-Decem (2018)
1091 10792–10801.
- 1092 [61] S. Santurkar, L. Schmidt, A. Madry, A Classification-Based Perspec-
1093 tive on GAN Distributions, in: International Conference on Learning
1094 Representations (ICLR).
- 1095 [62] J. Zhang, C. Li, Adversarial Examples: Opportunities and Challenges,
1096 IEEE Transactions on Neural Networks and Learning Systems 31 (2020)
1097 2578–2593.
- 1098 [63] Y. Sugawara, S. Shiota, H. Kiya, Checkerboard artifacts free convolu-
1099 tional neural networks, APSIPA Transactions on Signal and Information
1100 Processing 8 (2019) 1–9.
- 1101 [64] K. Wang, A. Mamidipalli, T. Retson, N. Bahrami, K. Hasenstab,
1102 K. Blansit, E. Bass, T. Delgado, G. Cunha, M. S. Middleton, R. Loomba,
1103 B. A. Neuschwander-Tetri, C. B. Sirlin, A. Hsiao, Automated CT and
1104 MRI Liver Segmentation and Biometry Using a Generalized Convo-
1105 lutional Neural Network, Radiology: Artificial Intelligence 1 (2019)
1106 180022.

- 1107 [65] K. Aderghal, A. Khvostikov, A. Krylov, J. Benois-Pineau, K. Afdel,
1108 G. Catheline, Classification of Alzheimer Disease on Imaging Modalities
1109 with Deep CNNs Using Cross-Modal Transfer Learning, in: Proceedings
1110 - IEEE Symposium on Computer-Based Medical Systems, volume 2018-
1111 June, Institute of Electrical and Electronics Engineers Inc., 2018, pp.
1112 345–350.
- 1113 [66] M. J. C. Crump, J. V. McDonnell, T. M. Gureckis, Evaluating Amazon’s
1114 Mechanical Turk as a Tool for Experimental Behavioral Research, PLoS
1115 ONE 8 (2013) e57410.
- 1116 [67] A. Anderson, P. K. Douglas, W. T. Kerr, V. S. Haynes, A. L. Yuille,
1117 J. Xie, Y. N. Wu, J. A. Brown, M. S. Cohen, Non-negative matrix
1118 factorization of multimodal MRI, fMRI and phenotypic data reveals
1119 differential changes in default mode subnetworks in ADHD, NeuroImage
1120 102 (2014) 207–219.