USE OF NEURAL SIGNALS TO EVALUATE THE QUALITY OF GENERATIVE ADVERSARIAL NETWORK PERFORMANCE IN FACIAL IMAGE GENERATION

A PREPRINT

Zhengwei Wang[*], **Graham Healy**[*], **Alan F. Smeaton, Tomas E. Ward**
Insight Centre for Data Analytics
Dublin City University
Dublin 9, Ireland
zhengwei.wang22@mail.dcu.ie, {graham.healy, alan.smeaton, tomas.ward}@dcu.ie

November 13, 2018

ABSTRACT

There is a growing interest in using Generative Adversarial Networks (GANs) to produce image content that is indistinguishable from a real image as judged by a typical person. A number of GAN variants for this purpose have been proposed, however, evaluating GANs is inherently difficult because current methods of measuring the quality of the output do not always mirror what is actually perceived by a human. We propose a novel approach that deploys a brain-computer interface to generate a **neural score** that closely mirrors the behavioral ground truth measured from participants discerning real from synthetic images. In this paper, we first compare the three most widely used metrics in the literature for evaluating GANs in terms of visual quality compared to human judgments. Second, we propose and demonstrate a novel approach using neural signals and rapid serial visual presentation (RSVP) that directly measures a human perceptual response to facial production quality independent of a behavioral response measurement. Finally we show that our neural score is more consistent with human judgment compared to the conventional metrics we evaluated. We conclude that neural signals have potential application for high quality, rapid evaluation of GANs in the context of visual image synthesis.

***K*eywords** Generative adversarial networks · Neural score · Brain signals · Brain-computer interface · Electroencephalography

1 Introduction

Generative adversarial networks (GANs) [Goodfellow et al., 2014] are attracting increasing interest with many different applications, for example the generation of plausible synthetic images [Radford et al., 2015, Arjovsky et al., 2017, Karras et al., 2017, Berthelot et al., 2017], image to image translation [Isola et al., 2017, Zhu et al., 2017] and simulated image refinement [Shrivastava et al., 2017]. Despite the many different GAN models reported in the literature, evaluation of GANs is still challenging. Some comprehensive reviews for GAN evaluation are available [Theis et al., 2015, Xu et al., 2018, Borji, 2018] but in summary the evaluation for GANs is divided into mainly two types, i.e. qualitative and quantitative metrics. The most representative qualitative metric is to use human annotation to determine the visual quality of the generated images. Quantitative metrics compare statistical properties between generated and real images. Both approaches have strengths and limitations. Qualitative metrics generally focus on how convincing the image is from a human perceptual perspective rather than detecting overfitting, mode dropping and mode collapsing problems [Metz et al., 2016]. Human annotation approaches are also time consuming because they require asking evaluators to generate behavioral responses on an image by image basis. Quantitative metrics in contrast, are less subjective but

[*] Equal contribution

the psycho-perceptual basis of image quality assessment is not well represented in the metrics hence the robustness of their performance is compromised. As a result, the field of research around evaluation methodologies for GANs is still a developing one and presents opportunities for new approaches. One such approach which we propose, is the introduction of brain signals in the context of a brain-computer interface.

A brain-computer interface (BCI) is a communication system in which an individual can send messages or commands to the external world without using the brain's normal output pathways of peripheral nerves and muscles [Wolpaw et al., 2002]. While there are several key BCI applications [Lees et al., 2018, Healy et al., 2017, Solon et al., 2017], there is a growing interest in using electroencephalography (EEG) signals in a BCI to help in searching sets of images. This is based on estimating image content by examining participants' neural signals in response to image presentation. The concept of rapid serial visual presentation (RSVP) can be introduced using a familiar example, that of rapidly riffling through the pages of a book in order to locate a needed image [Spence and Witkowski, 2013]. In RSVP, a rapid succession of target and standard (non-target) images are presented to a participant on a display at a rate of 4 Hz - 10 Hz. The location of target images within the high-speed presentation is not known in advance by users and hence requires them to actively look out for targets i.e. to attend to target images. This paradigm where users are instructed to attend to target images amongst a larger proportion of standard images is known as an oddball paradigm and is commonly used to elicit Event-related Potentials (ERPs) such as the P3, a positive voltage deflection that typically occurs between 300 ms - 600 ms after the appearance of a rare visual target within a sequence of frequent irrelevant stimuli [Polich, 2007]. Since participants do not know when target images will appear in the presentation sequence, their occurrence causes an attentional-orientation response that is characterized by the presence of a P3 (or P300) ERP. An example of a RSVP paradigm protocol is shown in Fig. 1.

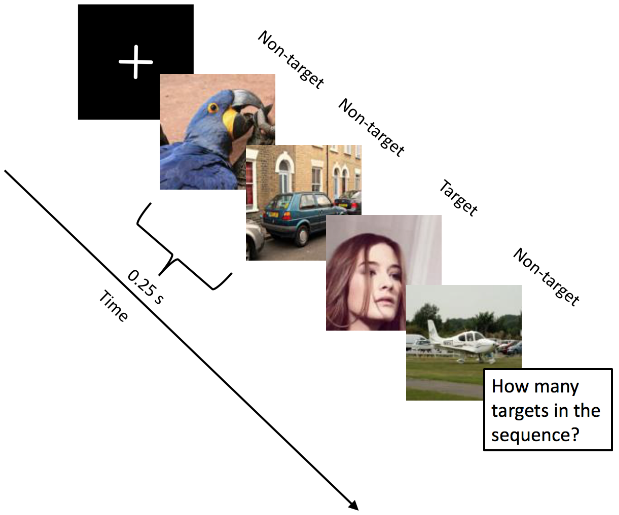


Figure 1: An RSVP image sequence showing juxtaposition of target and non-target images along with a response request.

Although some work in the literature has mentioned that quantitative metrics are correlated with human judgment [Salimans et al., 2016, Heusel et al., 2017], there is no specifically designed experiment explicitly described in the literature which compares quantitative metrics with those produced by human judgment. It should be noted that while the use of human judgment through annotation to evaluate GANs in terms of visual quality is very effective, current approaches are very time consuming and impractical in real-world application. Given this, our work has two primary contributions to make:

- The design and evaluation of a specific experiment to compare human assessments with leading quantitative metrics of GAN performance in terms of image quality.
- The demonstration of a fast and efficient BCI paradigm in which neural signals provide a superior measurement for evaluation of GANs.

2 Preliminaries

2.1 Generative adversarial networks

A generative adversarial network (GAN) has two components, the discriminator D and the generator G. Given a distribution $\mathbf{z} \sim p_{\mathbf{z}}$, G defines a probability distribution p_g as the distribution of the samples $G(\mathbf{z})$. The objective of GAN is to learn the generator's distribution p_g that approximates real data distribution p_r. Optimization of GAN is performed with respect to a joint loss for D and G

$$\min_G \max_D \mathbb{E}_{\mathbf{x}\sim p_r} log[D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}}[1 - D(G(\mathbf{z}))] \tag{1}$$

The evaluation for GANs can be considered to measure the dissimilarity between p_r and p_g. Unfortunately, the density of p_r is intractable. Thus, it is not possible to make a direct measurement between p_r and p_g. Another challenge for evaluation of a GAN is how to interpret that the evaluation metric for GANs indicates the visual quality.

2.2 Evaluation metrics

This paper uses three most widely used evaluation metrics for GANs in the literature for comparison.

2.2.1 The Inception Score

is the most widely used metric in the literature [Salimans et al., 2016]. It uses a pre-trained inception network [Szegedy et al., 2016] as the image classification model $\mathcal{M}$ to compute

$$IS = e^{\mathbb{E}_{\mathbf{x}}\sim p_g [KL(p_{\mathcal{M}(y|\mathbf{x})}||p_{\mathcal{M}(y)})]} \tag{2}$$

where $p_{\mathcal{M}}(y|\mathbf{x})$ is the label distribution of $\mathbf{x}$ that is predicted by the model $\mathcal{M}$ and $p_{\mathcal{M}}(y)$ is the marginal probability of $p_{\mathcal{M}}(y|\mathbf{x})$ over the probability p_g. A larger IS will have $p_{\mathcal{M}}(y|\mathbf{x})$ close to a point mass and $p_{\mathcal{M}}(y)$ close to uniform, which indicates that the Inception network is very confident that the image belongs to a particular ImageNet category and all categories are equally represented. This suggests that the generative model has both high quality and diversity. However, IS may fail in some cases [Barratt and Sharma, 2018]. 1/IS is used as a score in this work in order to be consistent with other two conventional scores.

2.2.2 Kernel Maximum Mean Discrepancy (MMD)

MMD [Gretton et al., 2007] is computed as

$$MMD^2(p_r, p_g) = \mathbb{E}_{\substack{\mathbf{x}_r, \mathbf{x}'_r, \sim p_r, \\ \mathbf{x}_g, \mathbf{x}'_g \sim p_g}} [k(\mathbf{x}_r, \mathbf{x}'_r) - 2k(\mathbf{x}_r, \mathbf{x}_g) + k(\mathbf{x}_g, \mathbf{x}'_g)] \tag{3}$$

It measures dissimilarity between p_r and p_g for some fixed kernel function k. Gaussian kernel, defined as $k(x, x') = e^{-\frac{|x-x'|^2}{2\sigma}}$ where σ is the bandwidth parameter, is often used for computation [Li et al., 2015]. A lower MMD indicates that p_g is closer to p_r, indicating a GAN shows has performance.

2.2.3 The Frechet Inception Distance (FID)

FID [Heusel et al., 2017] uses feature space extracted from a set of generated image samples by a specific layer of inception network. Regarding the feature space as multivariate Gaussian, the mean and covariance are estimated for both the generated data and real data. FID is computed as

$$FID(p_r, p_g) = ||\mu_r - \mu_g||_2^2 + Tr(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{\frac{1}{2}}) \tag{4}$$

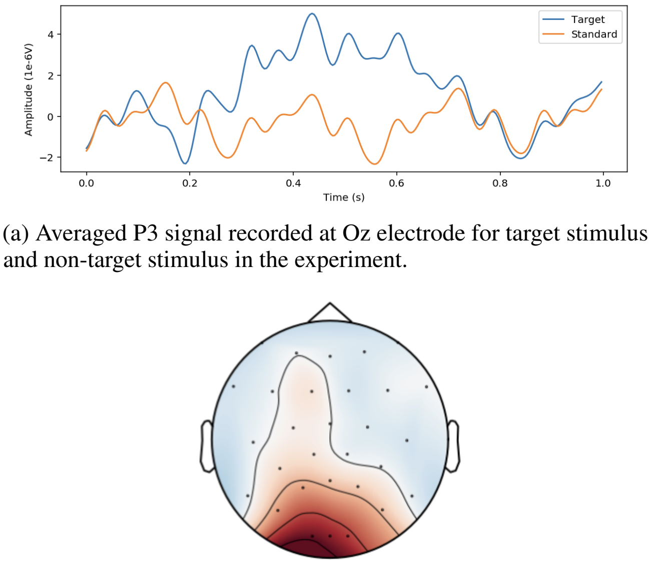
Similar to MMD, smaller FID indicates better GAN performance.

2.3 The Event-related Potential and P3 (or P300) component

In neuroscience, Event-related Potentials (ERPs) can be detected by capturing very small voltage changes generated in the brain on the scalp in response to specific events or stimuli [Blackwood and Muir, 1990]. ERPs show EEG changes that are time locked to sensory, motor or cognitive events, and provide a safe and noninvasive approach to study psychophysiological correlates of mental processes [Sur and Sinha, 2009]. ERPs can be elicited by a wide variety of sensory, cognitive or motor events. The P3 is but only one ERP component that was discovered by Sutton [Sutton et al., 1965] and since then has been the major component of research in the field of ERPs. The P3 can be elicited when a participant is instructed to respond mentally or physically to a target stimulus and not respond otherwise in the experiment. In this way, it reflects a participant's attention, that is it can be modulated by the specific instruction given to a participant. Fig. 2 shows that the P3 response elicited by target stimulus it typically evident between 300 ms - 600 ms post stimulus presentation depending on the type of task.

We list the explanation of some terminology used in this work as below:

- *Trial:* Each individual presented image is called a trial.
- *Single trial P3 amplitude:* It is the P3 amplitude corresponding to each single image. The P3 amplitude is calculated by selecting the maximum value between 300 ms - 600 ms for each EEG epoch.
- *Averaged P3 amplitude:* It is the difference between the averaged P3 amplitude corresponding to target trials and the averaged P3 amplitude corresponding to standard trials (non-face).
- *Reconstructed single trial P3 amplitude:* It is the beamformed P3 amplitude corresponding to each single image. The beamformed P3 amplitude is calculated by selecting the maximum value between 300 ms - 600 ms for each beamformed EEG epoch.
- *Reconstructed averaged P3 amplitude:* It is the difference between the averaged LDA beamformed signal corresponding to target trials and averaged LDA beamformed signal corresponding to standard trials (non-face).



(a) Averaged P3 signal recorded at Oz electrode for target stimulus and non-target stimulus in the experiment.

(b) P3 topography for target stimulus in the experiment.

Figure 2: P3 response interpretation using time series and topography plot, *Participant 9*.

3 Methodology

3.1 Data acquisition and experiment

We use three GAN models to generate synthetic images: DCGAN[Radford et al., 2015], BEGAN[Berthelot et al., 2017] and Progressive Growing of GANs (PROGAN)[Karras et al., 2017] - see Fig. 3. Image streams in the experiment contain generated images of DCGAN, BEGAN, PROGAN, RFACE images and other category images. EEG was recorded for both of these two types of tasks along with timestamping information for image presentation and behavioural responses (via a photodiode and hardware trigger) to allow for precise epoching of the EEG signals for each trial [Wang et al., 2016]. EEG data was acquired using a 32 channel BrainVision actiCHamp at 1000 Hz sampling frequency, using electrode locations as defined by the 10-20 system. Pre-processing of some kind is generally a required step before any meaningful interpretation or use of any EEG data can be realized. Pre-processing typically involves re-referencing, filtering the signal (by applying a bandpass filter to remove environmental noise or to remove activity in non-relevant

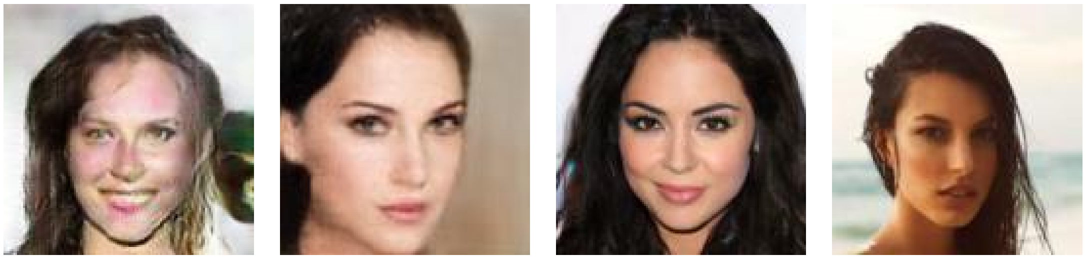


Figure 3: Face image examples used in the experiment. From left to right: DCGAN, BEGAN, PROGAN, RFACE.

frequencies), epoching (extracting a time epoch typically surrounding the stimulus onset) and trial/channel rejection (to remove those containing artifacts). In this work, a common average reference (CAR) was utilized and a bandpass filter (e.g. 0.5-20 Hz) was applied to the dataset. EEG data was then resampled at 250 Hz and the analysis of a behavior response is considered between 0 and 1 s after the presentation of a stimulus.

EEG data from up to 10 participants was used in this work. Data collection was carried out with approval from Dublin City University's Research Ethics Committee. Each participant completed BE and RSVP tasks. BE tasks contained three blocks, where each block contained 90 images (18 images for each face category thus 72 face images in total and 18 images for non-face images) thus there were 216 face images and 54 non-face images in BE task. Participants are presented with one image at a time and asked to press a button corresponding to a "Yes" if they perceive a real face (RFACE) or a "No" for anything they perceive as not being a real face. Next feedback is given on whether or not the presented image is indeed a real face so that the participants can learn from this to distinguish real faces from those which are artificially generated. The accuracy of the subject's responses are recorded and the performance is referred to subsequently as a "human judgment" metric.

RSVP task contained 28 blocks where last two blocks repeated one random selected previous block, one is exactly the same as the selected block while another randomly shuffled presented image sequence. Each RSVP block contained 240 images (6 image for each face category thus 24 face targets in total/216 non-face images), thus there were 6240 images (624 face targets/5616 non-face images) available for each participant. In the RSVP task, image streams are presented to participants at a 4 Hz presentation rate. Participants are asked to search for real face (RFACE) images in this task, and this will elicit a P3. We hypothesize that a larger P3 amplitude should be elicited when more realistic images are presented. Hence, P3 amplitude is a proxy measure for the visual quality of the images generated by the GANs. We also compare the P3 amplitude in the RSVP task to the human judgment measure in the BE task to determine if they are consistent with each other.

3.2 P3 reconstruction

EEG in our study is recorded by using a number of spatially distributed electrodes across the scalp (32 channels of EEG in this study). The P3 component is typically predominant on the posterior electrodes on the head, which also means the P3 is detected in multiple channels simutaniously. We uses the LDA Beamformer [Treder et al., 2016] to reconstruct the P3 in this work for the following reasons. First, it is difficult to compare the P3 in number of channels and the location of P3 varies with different participants. Second, P3 suffers from strong background brain activity so it has a very low signal-to-noise ratio (SNR) [Luck, 2014]. The LDA Beamformer method allows us to reconstruct the P3 from multi-dimensional recorded EEG signal i.e. transform 32 channels of EEG to one channel time series data so it is easy for us to make within-subject comparisons with the additional benefit of improving the SNR for the reconstructed P3 as well. Given an EEG epoch $\mathbf{X} \in \mathbb{R}^{C \times T}$ (C is the number of channels and T is time series points in that EEG epoch). The optimization problem for the LDA Beamformer is to find a projection vector (we call it a spatial filter in the area of EEG) $\mathbf{w}$ that solves the optimization as follows

$$\min_{\mathbf{w}} \mathbf{w}^{'}\mathbf{\Sigma}\mathbf{w} \ \ \text{s.t. } \mathbf{w}^{'}\mathbf{p} = 1 \tag{5}$$

where $\mathbf{\Sigma} \in \mathbb{R}^{C \times C}$ is the EEG epoch covariance matrix and $\mathbf{p} \in \mathbb{R}^{C \times 1}$ contains each EEG epoch channel value at specific time index. By solving equation 5, the optimal projection vector $\mathbf{w}$ can be calculated as

$$\mathbf{w} = \mathbf{\Sigma}^{-1}\mathbf{p}(\mathbf{p}^{'}\mathbf{\Sigma}^{-1}\mathbf{p})^{-1} \tag{6}$$

After getting the optimal $\mathbf{w}$, a high dimensional EEG epoch then can be projected to the one dimensional subspace (reconstructed signal) as

$$\mathbf{X}_{sub} = \mathbf{w}^{'}\mathbf{X} \tag{7}$$

The LDA Beamformer method can be applied to different time regions to reconstruct different individualized spatial profiles for ERP components present in that time frame. In this study, we used it to reconstruct the P3 component that appears in 400 ms - 600 ms time region, hence we apply LDA Beamformer between 400 ms - 600 ms in order to reconstruct P3.

3.3 Neural score

The *reconstructed averaged P3 amplitude* is going to be used for evaluating GANs. Because P3 latency is varying with seeing each image [Luck, 2014], we pick the maximum value between 400 ms and 600 ms for as a *reconstructed single trial P3 amplitude* and then average them across the trials to get the *reconstructed averaged P3 amplitude*. It should be noted that **neural score benefits from high SNR** comparing to the traditional single trial P3 for following reasons: 1) The LDA beamformer has been applied to raw EEG epoch in order to maximize the SNR; 2) Neural score is calculated by averaging trials which is able to mitigate the background EEG noise. Hence, neural score is a stable measurement in this work. It should be noted that higher neural score indicates better GAN performance which is reversed to the traditional score used in this work.

4 Results

4.1 Behavior task performance

We include 10 participants in the BE tasks and record the accuracy of their judgments for each face category. In Table 4.1 it can be seen that human achieve the lowest accuracy 0.7 for PROGAN and highest accuracy 0.994 for DCGAN i.e. participants rank PROGAN, BEGAN and DCGAN from high performance to low performance respectively. It is interesting that human judgment accuracy for RFACE is 0.686 and it is very low. This may be caused by participants being convinced by GAN generated images and subsequently feel less confident on the RFACE images, which indicates that GANs are able to convince participants in this case.

ID	DCGAN	BEGAN	PROGAN	RFACE
1	1.0	0.759	0.704	0.759
2	0.981	0.741	0.537	0.537
3	1.0	0.796	0.778	0.537
4	0.981	0.889	0.704	0.667
5	1.0	0.667	0.648	0.759
6	1.0	0.926	0.704	0.759
7	1.0	0.815	0.611	0.759
8	0.981	0.815	0.870	0.759
9	1.0	0.796	0.685	0.704
10	1.0	0.815	0.759	0.722
Mean	0.994	0.802	**0.7**	0.686

Table 1: Accuracy for each GAN in the BE task. Ranked by PROGAN, BEGAN, DCGAN. Lower accuracy indicates better image quality for GANs i.e. it convinced the participant.

4.2 Traditional evaluation metrics

Three traditional methods are employed to evaluate GANs used in this study and their results are shown in Table 2. It can be seen that all three methods are consistent with each other and they rank the GANs in order of PROGAN, DCGAN and BEGAN from high performance to low performance (and they all give the highest performance evaluation for RFACE). By comparing the three traditional evaluation metrics to the BE task performance, it is not consistent with the human judgment of GANs performance. It should be noted that IS is able to measure the quality for the generated images [Salimans et al., 2016] while the other two methods can not. However, IS here still rates that DCGAN outperforms BEGAN. For the RFACE score given by MMD and FID, they are both types of distance measurements of two probability distribution. So they will all return 0 if you feed the same inputs to them. Hence, they are unable to compare the image quality between images generated by GANs and real face images and they are only able to compare GANs performance.

Methods	DCGAN	BEGAN	PROGAN	RFACE
1/IS	0.44	0.57	0.42	0.30
MMD	0.22	0.29	0.12	0
FID	63.29	83.38	34.10	0

Table 2: Traditional score of each participant for each category. Lower score indicates better performance of GAN.

4.3 Rapid Serial Visual Presentation task performance

In order to employ neural signals to evaluate the performance of GANs, we use the RSVP paradigm to elicit the P3 ERP. Fig. 4 shows the *reconstructed averaged P3 signal* of one selected participant using LDA Beamformer in the RSVP experiment. It should be noted here that the *reconstructed averaged P3 signal* is calculated as the difference between target trial averaging and standard trial averaging after applying LDA Beamformer i.e. $\frac{1}{n}\sum_{i=1}^{n}\mathbf{w}'\mathbf{X}_i^{target} - \frac{1}{m}\sum_{i=1}^{m}\mathbf{w}'\mathbf{X}_i^{standard}$, where $\mathbf{w}$ is the spatial filter calculated by LDA Beamformer, $\mathbf{X}$ is the EEG epoch, n and m are the numbers of targets and standards respectively. It can be seen that the reconstructed P3 has different amplitudes corresponding to different stimuli.

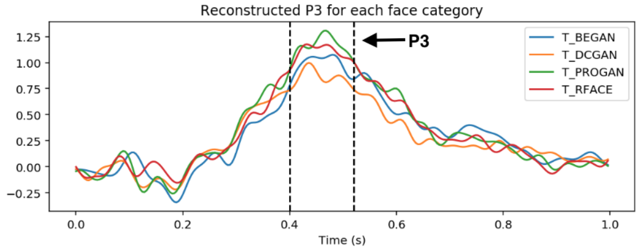


Figure 4: Reconstructed averaged P3 signal for each category of *participant 9* in this study.

We also show the neural score for each participant (for each GAN) in Table 3 for this study and the *reconstructed single trial P3 amplitude* distribution corresponding to each individual image of each category for each participant in Fig. 5. From Fig. 5, it can be seen that different categories have different *reconstructed single trial P3 amplitude* distributions and in particular can have different relative numbers of small reconstructed single trial P3 amplitudes. As the neural score is calculated by *averaging reconstructed single trial P3 amplitude* across all trials for a GAN type for a participant, those presented images that are not able to elicit *a single trial P3* or which only elicit a weak *single trial P3* will produce a smaller neural score.

ID	DCGAN	BEGAN	PROGAN	RFACE
1	0.576	0.645	0.736	0.686
2	0.598	0.773	0.882	0.885
3	0.434	0.623	0.686	0.584
4	0.522	0.549	1.017	0.998
5	0.492	0.683	0.699	0.714
6	0.656	0.808	0.905	0.796
7	0.460	0.630	0.941	0.866
8	0.666	0.684	0.696	0.678
9	0.752	0.823	1.057	1.007
10	0.710	0.765	1.097	0.964
Mean	0.587	0.698	**0.872**	0.818

Table 3: Computed neural score of each participant for each category. Higher score indicates better performance of GAN.

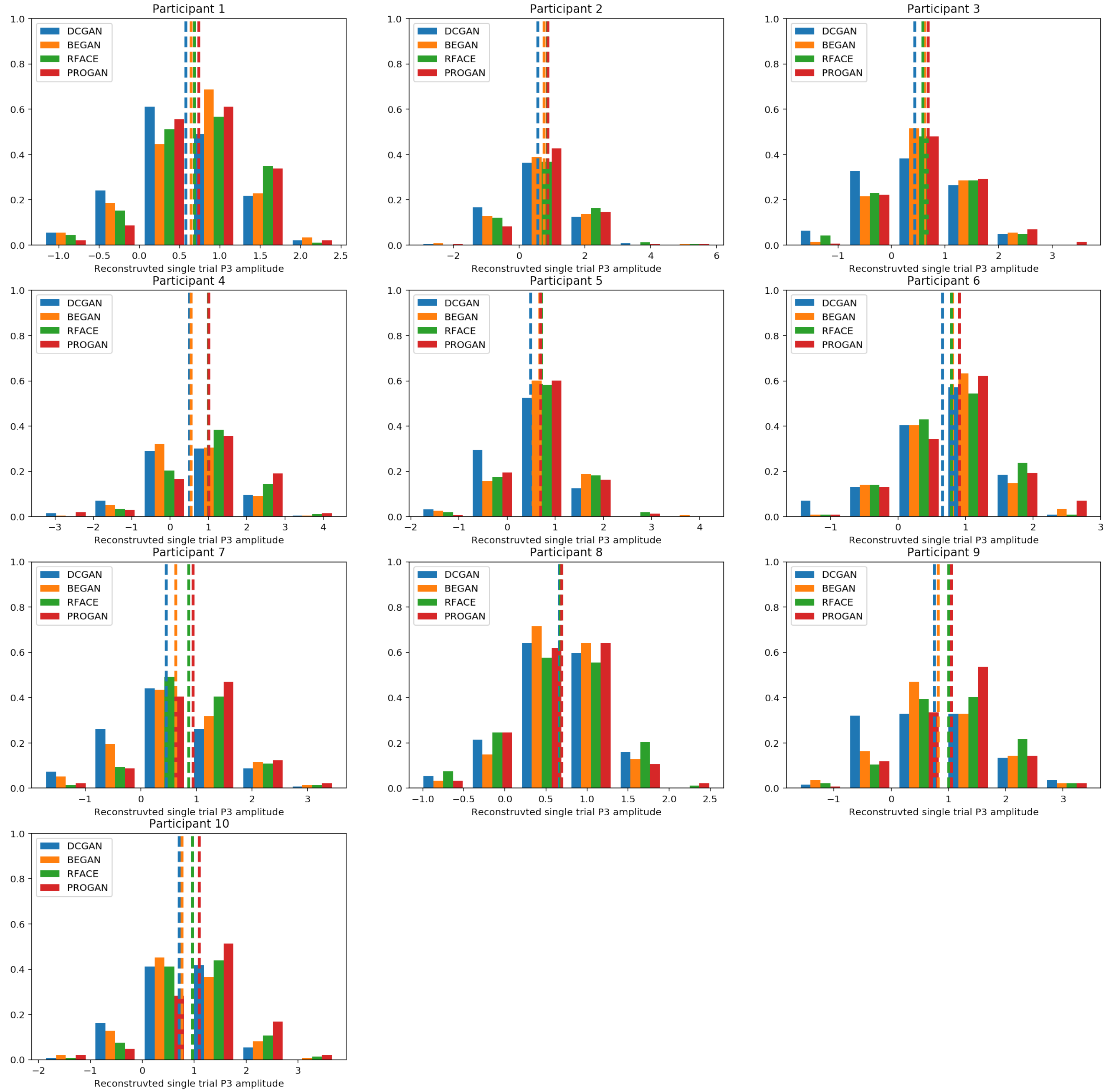


Figure 5: Single trial P3 amplitude of each category (DCGAN, BEGAN, RFACE, PROGAN) for each participant. Vertical dashed lines show the average neural score for the respective face image category.

From the mean value of the neural score and BE accuracy, it can be seen that the neural score is consistent with the BE accuracy (human judgment) i.e. PROGAN > BEGAN > DCGAN. We explore deeper this correlation between neural score and human judgment (see Fig. 6) as follows. First, we have calculated a Pearson correlation between neural score and BE accuracy. Second, we illustrate Spearman correlation between neural score and BE accuracy. Third, given that our sample size in this paper is limited, we use bootstrap [Efron and Tibshirani, 1994] to validate the correlation of both Pearson and Spearman. This bootstraping process was used as in both Pearson and Spearman correlation calculations we are pooling repeated measurements for participants (i.e. mixing DCGAN, BEGAN and PROGAN measurements), and hence are violating an assumption of both statistical tests. Bootstrapping both measures like this allows us to sidestep this violation of assumptions of both techniques and still obtain a reliable statistic. We do this by repeatedly randomly shuffling the BE accuracy values and neural scores, and then use Pearson linear correlation. After following this process 1000 times, we count how many calculated p values (i) are smaller than the original p value. $\frac{i}{1000}$ now becomes the bootstrapped Pearson/Spearman p value and it demonstrates that the probability to get a smaller p value with a randomized dataset. For the Pearson correlation, Pearson correlation coefficient and p value are -0.6216 and 0.0005 respectively while the Spearman correlation coefficient and p value are -0.5905 and 0.0012 respectively. This strongly supports the interpretation that our neural score is predictive of human judgment. This is consistent with our hypothesis that higher neural scores indicates better GAN models which is also indicated by lower BE accuracy. The bootstrapped p values for Pearson and Spearman are both 0.001, which means that it is unlikely to have obtained these correlation results by chance. Bootstrapped results for both Pearson and Spearman are shown in Fig. 7 and Fig. 8.

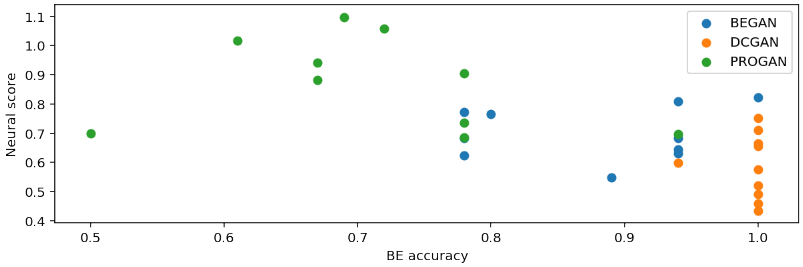


Figure 6: Scatter plot between neural score and BE accuracy.

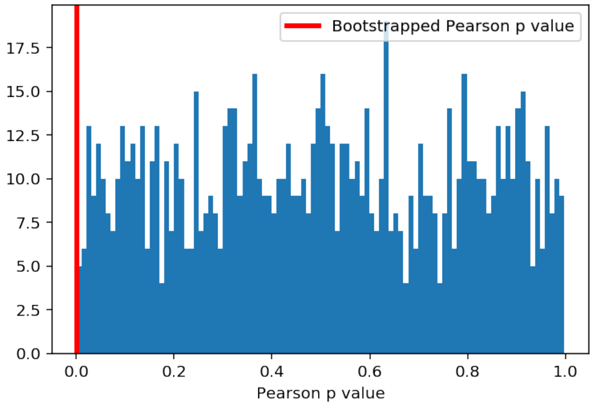


Figure 7: Histogram of bootstrapped Pearson p values.

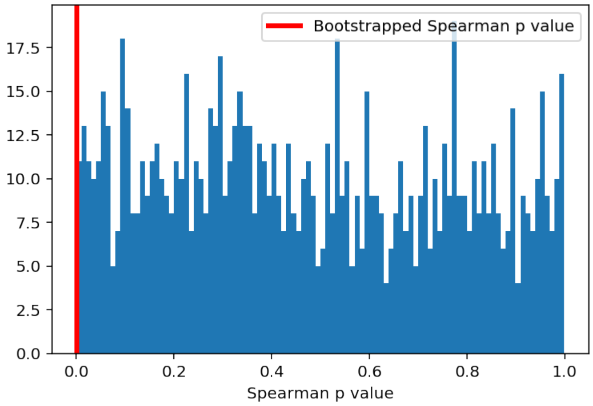


Figure 8: Histogram of bootstrapped Spearman p values.

It is notable that PROGAN achieved a higher neural score than RFACE. However, from the human assessment results in the previous section, it can be noticed that participants find the PROGAN output quite convincing rating faces produced by the GAN similar to real facial image (RFACE). Indeed the neural score here tells us that the PROGAN output used in the RSVP task accrues more attention from participants than even RFACE which indicates that PROGAN generates convincing high quality synthetic face images.

5 Discussion

We have compared human assessment with three representative quantitative metrics and used these for comparison with our proposed neural scoring approach. In short, our neural score conveys a measure of the visual quality of the generated images based on our hypothesis that a generated image which looks more like a real face image will elicit larger P3 in our RSVP task. Although the other three evaluation methods can provide insight into several aspects of GAN performance, we study their effectiveness from an image visual quality perspective only as this is the focus of this work. The results are compelling in their demonstration that the proposed neural score is better correlated with human judgment than the three quantitative metrics. This is important as an evaluation of the visual quality of a generated image is useful in understanding performance characteristics of specific GAN designs and training data sets. The method proposed can meet this need and is independent of any data modelling assumptions. In contrast, conventional quantitative metrics may fail in this regard. For example, IS is a model based evaluation method and the model is very sensitive to adversarial samples [Kurakin et al., 2016]. IS will produce a very high score if the generated images are produced using adversarial training [Barratt and Sharma, 2018]. Our neural score approach would not be compromised with such images in comparison. It is worth noting that compared with MMD and FID, both IS and our neural score provide a potentially good means of comparing the visual quality between generated images and real images i.e. IS and neural score may give higher score for the generated image that has better visual quality than the real image. IS however, unlike the neural scoring approach is not able to improve on the ranking of the three GANs compared to MMD and FID. It is also worth commenting that while GANs for generating facial images are explored in this study, our approach could potentially be used for any type of generated images because the appearance of the P3 feature does not depend on which type of images are presented to participants e.g. neural score may be applicable in the evaluation of GANs in bedroom image generation.

The work presented here focuses on evaluating image visual quality only. Consequently there are some limitations when using the neural score to evaluate GANs in this way. Overfitting, mode dropping and mode collapsing are very important aspects of GAN performance and most quantitative methods are able to assess these in some way. However for these broader assessments, we can augment these quantitative methods with our neural score to gain a better assessment of overall GAN performance. In reality, choosing the appropriate evaluation metric for GANs depends on the application and which type of problem is being solved by the GAN. If the goal of the use of the GAN is to generate high visual quality images e.g. super resolution image reconstruction, a qualitative metric is preferred in that case. If the GAN is to be trained to capture the categories of large image datasets, a quantitative metric would be the better choice. Therefore the inclusion of a neural scoring approach as we have demonstrated should be considered in the context of the application requirement.

6 Conclusion

We have conducted a comprehensive comparison between human assessments and three quantitative metrics for the comparison of image quality in the specific GAN application of facial imagery synthesis. We proposed and assessed a neural interfacing approach in which a neural score is introduced as an alternative evaluation of GANs in terms of image visual quality. We interpret our results to conclude that our neural score is more consistent with assessments made by humans when compared to the three quantitative metrics and we show that the correlation between our neural score and human judgment is not produced by chance i.e. $p = 0.001$. We believe that our proposed BCI paradigm based on a rapid serial visual presentation approach is more efficient and less prone to error compared to conventional human annotation. Consequently we suggest that approaches using such neural signals may complement or for some specific applications, replace, conventional metrics for evaluation of GAN performance.

7 Acknowledgement

This work is funded as part of the Insight Centre for Data Analytics which is sup-ported by Science Foundation Ireland under Grant Number SFI/12/RC/2289.

References

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

David Berthelot, Thomas Schumm, and Luke Metz. Began: boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint*, 2017.

Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *CVPR*, volume 2, page 5, 2017.

Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844*, 2015.

Qiantong Xu, Gao Huang, Yang Yuan, Chuan Guo, Yu Sun, Felix Wu, and Kilian Weinberger. An empirical study on evaluation metrics of generative adversarial networks. *arXiv preprint arXiv:1806.07755*, 2018.

Ali Borji. Pros and cons of gan evaluation measures. *arXiv preprint arXiv:1802.03446*, 2018.

Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163*, 2016.

Jonathan R Wolpaw, Niels Birbaumer, Dennis J McFarland, Gert Pfurtscheller, and Theresa M Vaughan. Brain–computer interfaces for communication and control. *Clinical neurophysiology*, 113(6):767–791, 2002.

Stephanie Lees, Natalie Dayan, Hubert Cecotti, Paul Mccullagh, Liam Maguire, Fabien Lotte, and Damien Coyle. A review of rapid serial visual presentation-based brain–computer interfaces. *Journal of neural engineering*, 15(2): 021001, 2018.

Graham Healy, Zhengwei Wang, Cathal Gurrin, Tomas Ward, and Alan F Smeaton. An eeg image-search dataset: a first-of-its-kind in ir/iir. nails: neurally augmented image labelling strategies. 2017.

Amelia J Solon, Stephen M Gordon, BJ Lance, and VJ Lawhern. Deep learning approaches for p300 classification in image triage: Applications to the nails task. In *Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-13, Tokyo, Japan*, pages 5–8, 2017.

Robert Spence and Mark Witkowski. *Rapid serial visual presentation: design for cognition*. Springer, 2013.

John Polich. Updating p300: an integrative theory of p3a and p3b. *Clinical neurophysiology*, 118(10):2128–2148, 2007.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

Shane Barratt and Rishi Sharma. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018.

Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, pages 513–520, 2007.

Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. In *International Conference on Machine Learning*, pages 1718–1727, 2015.

DHR Blackwood and WJ Muir. Cognitive brain potentials and their application. *The British Journal of Psychiatry*, 157 (S9):96–101, 1990.

Shravani Sur and VK Sinha. Event-related potential: An overview. *Industrial psychiatry journal*, 18(1):70, 2009.

Samuel Sutton, Margery Braren, Joseph Zubin, and ER John. Evoked-potential correlates of stimulus uncertainty. *Science*, 150(3700):1187–1188, 1965.

Zhengwei Wang, Graham Healy, Alan F Smeaton, and Tomas E Ward. An investigation of triggering approaches for the rapid serial visual presentation paradigm in brain computer interfacing. In *Signals and Systems Conference (ISSC), 2016 27th Irish*, pages 1–6. IEEE, 2016.

Matthias S Treder, Anne K Porbadnigk, Forooz Shahbazi Avarvand, Klaus-Robert Müller, and Benjamin Blankertz. The lda beamformer: optimal estimation of erp source time series using linear discriminant analysis. *Neuroimage*, 129:279–291, 2016.

Steven J Luck. *An introduction to the event-related potential technique*. MIT press, 2014.

Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.