Cognitive Computation (2020) 12:13–24
https://doi.org/10.1007/s12559-019-09670-y

Use of Neural Signals to Evaluate the Quality of Generative Adversarial Network Performance in Facial Image Generation

Zhengwei Wang[1] · **Graham Healy**[1] · **Alan F. Smeaton**[1] · **Tomás E. Ward**[1]

Received: 20 February 2019 / Accepted: 3 July 2019 / Published online: 8 August 2019

Abstract

There is a growing interest in using generative adversarial networks (GANs) to produce image content that is indistinguishable from real images as judged by a typical person. A number of GAN variants for this purpose have been proposed; however, evaluating GAN performance is inherently difficult because current methods for measuring the quality of their output are not always consistent with what a human perceives. We propose a novel approach that combines a brain-computer interface (BCI) with GANs to generate a measure we call Neuroscore, which closely mirrors the behavioral ground truth measured from participants tasked with discerning real from synthetic images. This technique we call a neuro-AI interface, as it provides an interface between a human's neural systems and an AI process. In this paper, we first compare the three most widely used metrics in the literature for evaluating GANs in terms of visual quality and compare their outputs with human judgments. Secondly, 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. The correlation between our proposed Neuroscore and human perceptual judgments has Pearson correlation statistics: $r(48) = -0.767$, $p = 2.089\text{e}-10$. We also present the bootstrap result for the correlation i.e., $p \leq 0.0001$. Results show that our Neuroscore is more consistent with human judgment compared with the conventional metrics we evaluated. We conclude that neural signals have potential applications for high-quality, rapid evaluation of GANs in the context of visual image synthesis.

Keywords Generative adversarial networks · Rapid serial visual presentation · Human judgments · Brain-computer interface · Neuro-AI interface

Introduction

Artificial intelligence (AI) has significant impact on society yet research into the interaction between humans and AI deserves further exploration and has only recently become a research focus. Cognitive computation provides a way of using cognitively inspired techniques to solve a variety of real-world problems, and these become especially useful when the interface between an AI system and a human is via a brain-computer interface. Abbass [1] recently explored the last 50 years of the human-AI relationship with a focus on how the development of trust between the parties has been essential. He also covered the emergence of direct brain-computer interfaces based on electroencephalography (EEG).

As EEG can be the direct reflection of a human's mental processes, the use of EEG is widely studied and deployed in the cognitive computation literature, for example by [10, 24]. It has been demonstrated recently that EEG can be used effectively for reading emotion [24] and that a spiking neural network framework can be used to analyze a human's attention to a task by using EEG [10]. In this paper, we demonstrate a type of neuro-AI interface derived from

Zhengwei Wang and Graham Healy have equal contribution.

✉ Zhengwei Wang
zhengwei.wang22@mail.dcu.ie

Graham Healy
graham.healy@dcu.ie

Alan F. Smeaton
alan.smeaton@dcu.ie

Tomás E. Ward
tomas.ward@dcu.ie

[1] Insight Centre for Data Analytics, Dublin City University, Dublin 9, Ireland

cognitive computational perspective (as seen in Fig. 1), which uses neural signals, in this case EEG, to score the performance of generative adversarial networks (GANs). The relevance between our work and the existing literature such as [10, 24] is that a processing pipeline has been developed and demonstrated for transforming EEG signals into a value (score or accuracy), and this value matches well a human's cognitive response to a specific class of stimulus, in our case an artificially generated facial image. Moreover, our work contains experimental details and provides neuroscientific interpretation in the comparison of our EEG-based technique to existing approaches in the literature.

GANs [13] are attracting increasing interests across many different computer vision applications, for example the generation of plausible synthetic images [2, 5, 20, 32], image-to-image translation [19, 48], and simulated image refinement [34]. Despite the extensive work and the many different GAN models reported in the literature, evaluation of the performance of GANs is still challenging. Some comprehensive reviews for GAN evaluation are available including work in [7, 40, 46], and in summary, the evaluation for GANs is divided into two main types, *qualitative* and *quantitative*. 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 [30]. 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 the psychoperceptual basis of image quality assessment is not well represented in such metrics; hence, the robustness of their performance is compromised. As a result, the field of research around evaluation methodologies for GANs is still developing and presents opportunities for new approaches. One such approach which we propose is the introduction of a neuro-AI interface that uses brain signals for image evaluation in the context of a brain-computer interface.

A brain-computer interface (BCI) is a communication system in which an individual sends signals to the external world without using the brain's normal output pathways of peripheral nerves and muscles [45]. While there are several key BCI applications [15, 23, 35], there is a growing interest in using EEG signals in a BCI to help in searching through 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 [36]. In RSVP, a rapid succession of target and standard (non-target) images is presented to a participant via a display at a rate of 4 to 10 Hz. The location of target images within the high-speed presentation is not known in advance by participants and hence requires them to actively look out for targets, i.e., to attend to target images. This paradigm where participants 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 the P300 event-related potential (ERP), a positive voltage deflection that typically occurs between 300 and 600 ms after the appearance of a rare visual target within a sequence of frequent non-relevant stimuli [18, 31]. Since participants do not know when target images will appear in the presentation sequence, their

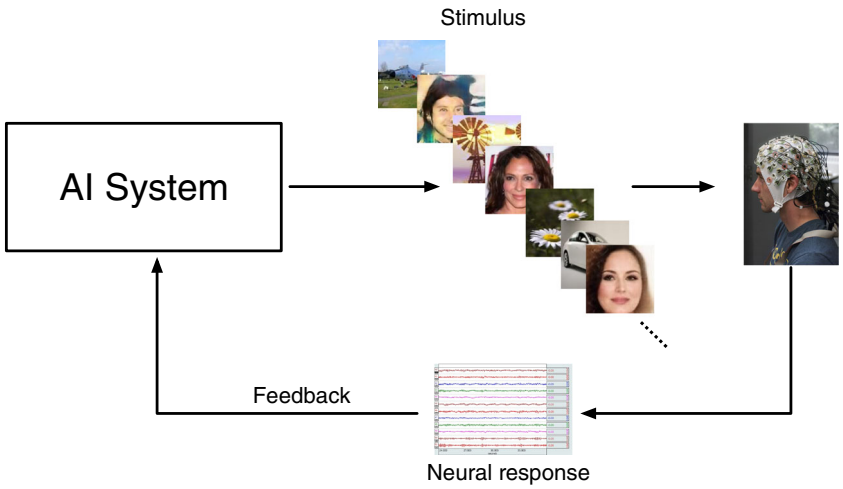

Fig. 1 Schematic of neuro-AI interface demonstrated in this study. A type of AI system (e.g., GANs used in this work) produces image stimulus to participants and the corresponding recorded neural response returns to scoring the performance of GANs

occurrence causes an attentional-orientation response that is characterized by the presence of a P300 (or P3) ERP. An example of a RSVP paradigm protocol is shown in Fig. 2 where the participant's task might be to count the number of images with faces or to recognize the face of a particular individual.

The P300 ERP can suffer from a low signal-to-noise ratio (SNR), and its appearance spans multiple electrodes on the scalp, which make the precise measurement of P300 activity in the raw, unprocessed EEG epoch difficult. Our previous work [43, 44] has shown that the P300 can be spatially filtered to improve SNR and reduce dimensionality. The work here will demonstrate a pipeline that uses a linear discriminant analysis (LDA) beamformer to reconstruct the P300 component for each type of GAN.

Although some work in the GAN evaluation literature has mentioned that quantitative metrics are correlated with human judgment [17, 33], there is no specifically designed work reported in the literature which compares quantitative metrics with those produced by human judgment. It should be noted that the use of human judgment through annotation to evaluate GANs in terms of visual quality is very effective. However, such approaches are very time-consuming and impractical in terms of scale, in real-world applications. Given the advantages of conventional human annotation approaches, we explore the area of BCI as we know that neural signals can reflect human perception. In this work, we propose a type of neuro-AI interface for evaluating GAN outputs and we deploy an oddball task for eliciting P300 components via a RSVP protocol, where human subjects are rapidly evaluating images produced by GANs. An evaluation metric called Neuroscore is proposed and the calculation of Neuroscore is demonstrated. Results show this neuro-AI interface is more efficient compared with conventional human annotation approaches and Neuroscore is highly correlated with behavioral human judgment. Given this, our work has two primary contributions:

- The design and evaluation of an experiment to compare human assessments with the leading quantitative metrics for GAN performance measurement in terms of image quality
- The demonstration of a fast and efficient neuro-AI interface in which neural signals provide a superior metric for the evaluation of GANs

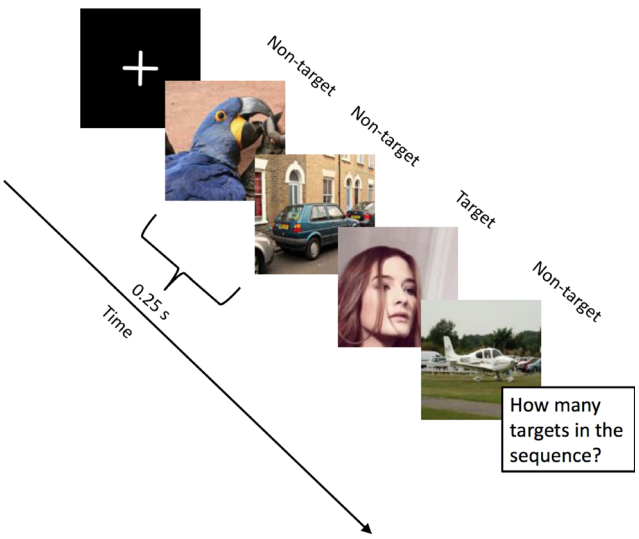

Fig. 2 A RSVP image sequence showing juxtaposition of target and non-target images along with a response request

Preliminaries

Generative Adversarial Networks

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

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

The evaluation of GANs can be considered an effort to measure the dissimilarity between p_r and p_g. Unfortunately, the accurate estimation of p_r is intractable, and thus, it is not possible to make a good estimation of the correspondence between p_r and p_g. Another challenge for the evaluation of a GAN is how to interpret that the evaluation metric indicates visual quality. Notwithstanding such challenges, metrics are available and we examine three well-known metrics as background and for comparative purposes.

GAN Evaluation Metrics

This paper uses three of the most widely used evaluation metrics for GANs in the literature for comparison, and we now examine these in turn.

Inception Score

The inception score is the most widely used GAN performance metric in the literature [33]. It uses a pre-trained inception network [39] as the image classification model $\mathcal{M}$ to compute

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

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 inception score 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. A larger inception score suggests that the generative model has both high quality and diversity. However, inception score may fail in some cases [4]. 1/IS (1/inception score) is used as the comparison score in the work in this paper, for consistency with the other two scores examined.

Kernel Maximum Mean Discrepancy

Maximum mean discrepancy (MMD) [14] is computed as

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

It measures the dissimilarity between p_r and p_g for some fixed kernel function k. A Gaussian kernel, defined as $k(\mathbf{x}, \mathbf{x}^\top) = e^{-\frac{||\mathbf{x}-\mathbf{x}^\top||^2}{2\sigma}}$ where $\mathbf{x}$ are input samples and σ is the bandwidth parameter, is often used for this purpose [25]. A lower MMD indicates that p_g is closer to p_r, indicating a GAN has better performance.

The Frechet Inception Distance

Frechet inception distance (FID) [17] uses a feature space extracted from a set of generated image samples by a specific layer of the 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

$$\mathrm{FID}(p_r, p_g) = ||\boldsymbol{\mu}_r - \boldsymbol{\mu}_g||_2^2 + \mathrm{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{\frac{1}{2}}) \tag{4}$$

Similar to MMD, a smaller FID indicates better GAN performance.

Comparing Metrics

In the case of the inception score, this is calculated through the inception model [39]. It has been shown previously that inception score is very sensitive to the model parameters [4]. Even scores produced by the same model trained using different libraries (e.g., Tensorflow, Keras, and PyTorch) differ a lot from each other. Inception score also requires a large sample size for the accurate estimation of $p_{\mathcal{M}}(\mathrm{y})$. FID and MMD both measure the similarity between training images and generated images based on the feature space [46], since the pixel representations of images do not naturally support the computation of meaningful Euclidean distances [12]. The main concern about the FID and MMD methods is whether the distributional characteristics of the feature space exactly reflect the distribution for the images [12].

In general, these conventional metrics are easily affected by small artifacts in either pixel space or feature space. For instance, some sharp artifacts in BEGAN may cause large difference between real and generated images regarding the distribution. However, such sharp artifacts would not affect image content and quality as human perception is more robust to conventional metrics regarding these issues.

The Event-Related Potential and P300 Component

In neuroscience, event-related potentials (ERPs) refer to low-amplitude voltage signals measured on the scalp which arise from current source dynamics in the brain whose changes reflect specific events or stimuli [6]. ERPs are characterized by EEG changes that are time-locked to sensory, motor, or cognitive events and provide a safe and non-invasive approach to study psychophysiological correlates of mental processes [37]. ERPs can be elicited by a wide variety of sensory, cognitive, or motor events. The P300 ERP component was discovered by Sutton et al. [38] and since then has been one of the most investigated ERP components. The P300 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. Figure 3 shows an averaged P300 response elicited by a target stimulus that is typically evident between 300 and 600 ms post presentation of a stimulus, depending on the type of task. A list of related physiologically relevant terminology and associated explanations used in this work is presented below:

- *Trial:* Each individual image presentation is called a trial.
- *Epoch:* An epoch is a specific time window which is extracted from the continuous EEG signal. Each epoch is time-locked with respect to an event (image stimulus presentation in our case).
- *Single-trial P300 amplitude:* This is the amplitude of the P300 component corresponding to each individual image. The P300 amplitude is calculated by selecting the maximum voltage value between 400 and 600 ms for each EEG epoch.
- *Averaged P300 amplitude:* This is the difference between the averaged target (for example a face) trial amplitudes and the averaged standard trial amplitudes (for example a non-face).

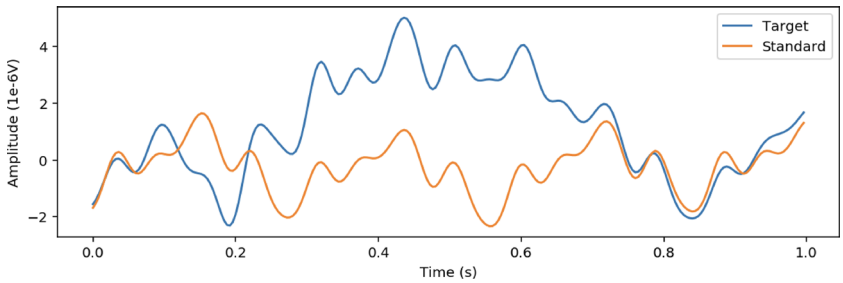

Fig. 3 Averaged ERP response for *participant 9* showing P300-related activity

– *Reconstructed single-trial P300 amplitude:* This is the P300 amplitude corresponding to each single target image. It is the LDA-beamformed single-trial P300 amplitude (the detail of the LDA beamformer method is introduced later in Section "P300 Reconstruction").
– *Reconstructed averaged P300 amplitude:* It is the difference between the averaged LDA-beamformed P300 corresponding to target trials and the averaged LDA-beamformed signal corresponding to standard trials (non-face).

Methodology

Data Acquisition and Experiment

We used three GAN models to generate synthetic images of faces: DCGAN [32], BEGAN [5], and progressive growing of GANs (PROGAN) [20] as shown in Fig. 4. Image streams in the experiment contain generated images from DCGAN, BEGAN, and PROGAN, as well as real face (RFACE) images and non-face category images. RFACE images were sampled from CelebA dataset [26]. Non-face category (standard) images were sampled from ImageNet dataset [9], similar to those used in other RSVP experiments such as [15, 16].

EEG data for 12 participants was gathered. Data collection was carried out with approval from Dublin City University Research Ethics Committee (REC/2018/115). Each participant completed two types of tasks which we call the behavioral experiment (BE) task and the rapid serial visual presentation (RSVP) task. The sequence of blocks presented in the experiment was BE → RSVP → BE → RSVP → BE.

The objective of the BE task was to record participants' responses to each type of image category while the RSVP task was to record EEG when participants were seeing the rapid presentation of images. The ultimate goal of this study was to compare whether the EEG responses in the RSVP task were consistent with participants' responses in the BE task.

The BE task consisted of three blocks, where each block contained 90 images (18 images for each face category resulting in 72 face images in total and 18 non-face images). Thus, there were 216 face images and 54 non-face images in the BE task in total. Participants were presented with one image at a time and asked to press a button corresponding to a "yes" if they perceived a real face (i.e., belonging to the RFACE set) or a "no" for anything they perceived as not being a real face (including fake face produced by GANs and non-face). Following each response, feedback was given on whether or not the presented image was indeed a real face to make participants pay more attention to the task. The accuracy (number of correct trials divided by number of presented images for that GAN type) of each participant's response was recorded, and their performance is referred to subsequently as a "human judgment" metric.

The RSVP task contained 26 blocks. Each RSVP block contained 240 images (6 images for each face category thus 24 face targets in total and 216 non-face images); thus, there were 6,240 images (624 face targets/5,616 non-face images) available for each participant. In the RSVP

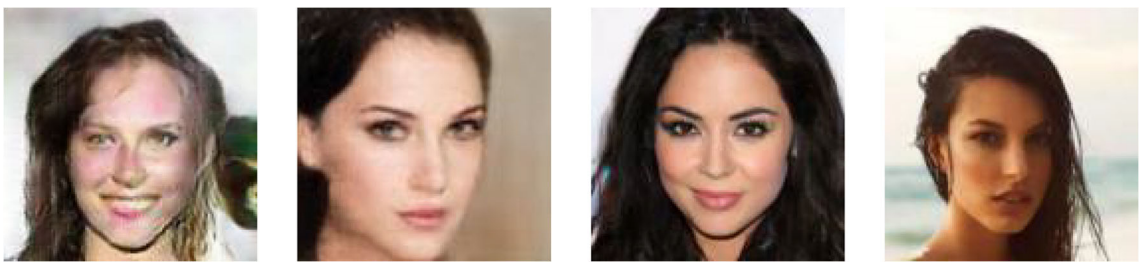

Fig. 4 Face image examples used in the experiment. From left to right: DCGAN, BEGAN, PROGAN, and real face (RFACE)

task, image streams were presented to participants at a 4-Hz presentation rate. Participants were asked to search for RFACE images in this task so as to elicit a P300. We compare the P300 amplitude in the RSVP task with the human judgment measure in the BE task to determine if they are consistent with each other.

EEG was recorded for both of the BE and RSVP tasks along with timestamping information for image presentation and behavioral responses (via a photodiode and hardware trigger) to allow for precise epoching of the EEG signals for each trial [42]. EEG data was acquired using a 32-channel BrainVision actiCHamp at 1000-Hz sampling frequency, using electrode locations as defined by the International 10-20 system.

To enhance the low signal-to-noise ratio of the acquired EEG, pre-processing is required. 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 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 (i.e., 0.5–20 Hz) was applied prior to epoching. EEG data was then downsampled to 250 Hz. Only behavioral responses occurring between 0 and 1 second after the presentation of a stimulus were used. Trial rejection was carried out to remove those trials containing noise such as eye-related artifacts (via a peak-to-peak amplitude threshold across all electrodes). Details of the retained trials for each participant are shown in Table 1. A LDA beamformer [41] was applied to the retained EEG epochs for each participant to enhance the signal-to-noise ratio (SNR). Details of the application of the LDA beamformer method is described in Section "P300 Reconstruction."

Table 1 Number of trials for each stimulus type remaining after artifact rejection

ID	DCGAN	BEGAN	PROGAN	RFACE	Standard
1	116	108	107	113	4220
2	100	106	110	98	3215
3	156	153	154	154	5553
4	144	153	143	144	5168
5	110	101	92	80	4150
6	135	131	122	106	4521
7	138	139	143	141	4955
8	151	151	150	151	5290
9	146	149	140	149	4832
10	104	87	93	82	3286
11	149	138	144	142	5270
12	97	92	99	101	3859

P300 Reconstruction

EEG in our study was recorded using a number of spatially distributed electrodes across the scalp (32 channels of EEG in this study). The P300 is typically predominant on the posterior electrodes of the scalp, which also means the P300 is detected in multiple channels simultaneously. We use the LDA beamformer [41] to reconstruct the P300 in this work for the following reasons. Firstly, it is difficult to compare P300 between participants across a number of channels as the location of the P300 varies across participants. Secondly, the P300 suffers from interference from strong background brain activity so it has a very low SNR [27]. The LDA beamformer method allows us to reconstruct the P300 from a multi-dimensional set of EEG signals, i.e., transform 32 channels of EEG to a one-channel time series facilitating within-subject comparisons (with the additional benefit of improving the SNR for the reconstructed P300 as well). Given an EEG epoch $\mathbf{X}_i \in \mathbb{R}^{C\times T}$ (C is the number of channels and T is the time series points in that EEG epoch), let $\mathbf{p}_1 \in \mathbb{R}^{C\times 1}$ and $\mathbf{p}_2 \in \mathbb{R}^{C\times 1}$ be the spatial patterns of a particular component in two different experimental conditions, e.g., face stimuli versus non-face stimuli in this paradigm. We denote the difference pattern as $\mathbf{p} := \mathbf{p}_1 - \mathbf{p}_2$ and the covariance matrix as $\Sigma \in \mathbb{R}^{C\times C}$ [41]. The optimization problem for the LDA beamformer is to find a projection vector (we call it a spatial filter in the area of EEG/BCI) $\mathbf{w} \in \mathbb{R}^{C\times 1}$ that satisfies

$$\min_{\mathbf{w}} \mathbf{w}^\top \Sigma \mathbf{w} \;\; s.t. \mathbf{w}^\top \mathbf{p} = 1 \tag{5}$$

The optimal projection vector $\mathbf{w}$ (in Eq. 5) can be calculated as

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

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

$$\mathbf{S}_i = \mathbf{w}^\top \mathbf{X}_i \tag{7}$$

where $\mathbf{S}_i \in \mathbb{R}^{1\times T}$ is a one-trial reconstructed source signal. 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 [44]. In this study, we apply the LDA beamformer between 400 and 600 ms in order to best extract the P300.

Neuroscore

The *reconstructed averaged P300 amplitude* is used as the basis for our novel metric for evaluating GAN outputs. To address latency of the P300 which varies across participants, this work [44] has successfully demonstrated the use of LDA beamformer to search for the optimal P300 time index

in a RSVP experiment. We select the maximum value in the 200-ms time window which is centered at the optimal time index to represent the *reconstructed single-trial P300 amplitude* and then average these across the trials to get the *reconstructed averaged P300 amplitude*. This *reconstructed averaged P300 amplitude* is the Neuroscore. The process of calculating Neuroscore can be seen in the algorithmic block below.

It should be noted that Neuroscore benefits from a high SNR compared with the traditional single-trial P300 for the following reasons:

1. The LDA beamformer has been applied to raw EEG epoch data in order to maximize the SNR.
2. Neuroscore is calculated by averaging trials which is able to mitigate the background EEG noise.

Hence, our proposed Neuroscore is a relatively robust metric as defined for this work. It should be noted that higher Neuroscore values indicate better GAN performance which is inverse to the traditional scores used in this work.

Algorithm 1 Steps for calculating Neuroscore.

Input:

- $\mathbf{X} \in \mathbb{R}^{N\times C\times T}$ is the EEG corresponding to target stimulus, N is the number of target trials, C is number of channels, T is number of time points.
- $\mathbf{K} \in \mathbb{R}^{M\times C\times T}$ is the EEG corresponding to standard stimulus, M is number of standard trials, C is number of channels, T is number of time points.

Output: Neuroscore

$\Sigma = \frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i\mathbf{X}_i{}^{\top} + \frac{1}{M}\sum_{i=1}^{M}\mathbf{K}_i\mathbf{K}_i{}^{\top}$
for t_i in [400 ms, 600 ms] **do**
  $\mathbf{p} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_{i,t_i} - \frac{1}{M}\sum_{i=1}^{M}\mathbf{K}_{i,t_i}$
  $\mathbf{w} = \Sigma^{-1}\mathbf{p}(\mathbf{p}^{\top}\Sigma^{-1}\mathbf{p})^{-1}$
  $\mathrm{J}_{t_i} \leftarrow \mathbf{w}^{\top}\Sigma\mathbf{w}$
  $\mathrm{W}_{t_i} \leftarrow \mathbf{w}$
end for
$t_{optimal}$=argmin$_{t_i}$ J
$\mathbf{w}_{optimal}$=W$_{t_{optimal}}$
t$_{P300}$=[t$_{optimal}$ - 100 ms, t$_{optimal}$ + 100 ms] ▷ *This is time window being detected for P300.*
for $i = 1 : N$ **do**
  $\mathbf{s} = \mathbf{w}^{\top}\mathbf{X}_i$
  $a = \max(\mathbf{s}_{t_{P300}})$
  $\mathrm{A}_i \leftarrow$ a
end for
Neuroscore = $\frac{1}{N}\sum_{i=1}^{N}\mathrm{A}_i$

Experimental Results

Behavior Task Performance

We included 12 participants in the BE tasks and recorded the accuracy (calculated as the number of correctly labeled images divided by the total number of images) of their judgments for each face category. In Table 2 , it can be seen that participants achieve the lowest accuracy (0.705) for PROGAN and the highest accuracy (0.994) for DCGAN, i.e., participants rank PROGAN, BEGAN, and DCGAN from high performance to low performance respectively. While learning effects may be present, our result is robust regardless of learning effects as we examined using different groups of RSVP blocks combined with different parts of the BE task, and the results remained consistent. It is interesting that human judgment accuracy for RFACE is 0.686 which is comparatively low. This may be caused by participants being convinced by GAN-generated images and subsequently feeling less confident on RFACE images, which indicates that GANs are able to convince participants in this case.

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 P300 ERP. Figure 5 shows the *reconstructed averaged*

Table 2 Accuracy for face images generated from three GANs and real face images in the BE task

ID	DCGAN	BEGAN	PROGAN	RFACE
1	1.000	0.759	0.704	0.759
2	0.981	0.741	0.537	0.537
3	1.000	0.796	0.778	0.537
4	0.981	0.889	0.704	0.667
5	1.000	0.667	0.648	0.759
6	1.000	0.926	0.704	0.759
7	1.000	0.815	0.611	0.759
8	0.981	0.815	0.870	0.759
9	1.000	0.796	0.685	0.704
10	1.000	0.815	0.759	0.722
11	1.000	0.907	0.759	0.685
12	1.000	0.963	0.704	0.796
Mean	*0.995*	0.824	*0.705*	0.695

Lower accuracy for GAN-generated images indicates better image quality, i.e., participants were often convinced that synthesized faces were in fact real

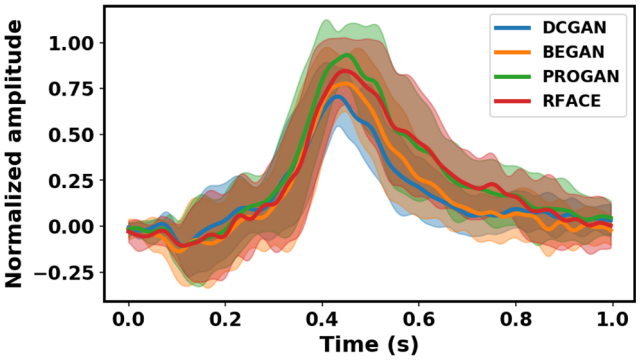

Fig. 5 Reconstructed averaged (via LDA beamformer) P300 signal across 12 participants in this study

P300 across all participants (using LDA beamformer) in the RSVP experiment. It should be noted here that the *reconstructed averaged P300 signal* is calculated as the difference between averaged target trials and averaged standard trials after applying the LDA beamformer method, i.e., $\frac{1}{n}\sum_{i=1}^{n}\mathbf{w}^{\top}\mathbf{X}_i^{target} - \frac{1}{m}\sum_{i=1}^{m}\mathbf{w}^{\top}\mathbf{X}_i^{standard}$, where $\mathbf{w}$ is the spatial filter calculated by the LDA beamformer, $\mathbf{X}$ are the EEG epochs, and n and m are the numbers of targets and standards respectively. The solid lines in Fig. 5 are the means of the *reconstructed averaged P300* signals for each image category (across 12 participants) while the shaded areas represent the standard deviations (across participants). It can be seen that the *reconstructed averaged P300* (across participants) clearly distinguishes between different image categories.

Figure 6 shows topographical plots (of averaged ERP activity) for the different image categories for each participant and for an average across participants. This demonstrates that the spatial topography of P300-related activity varies across participants. It is for this reason that we use the LDA beamformer approach to reconstruct the source P300 for each participant in this study (so as to eliminate erroneous measurement of the P300 by using a specific common channel). We also show a topographic representation of F-values from an ANOVA test that assesses statistical differences between the means of the four categories (one ANOVA for each channel). Larger F-values indicate a larger statistical effect when examining reconstructed P300 values across the four categories for a participant. It can be seen that spatial locations with high F-values are closely aligned to the P300's spatial topography.

We also show the Neuroscore for each participant in the study (for each GAN) in Table 3. A higher Neuroscore indicates better performance of a GAN. Ranking the

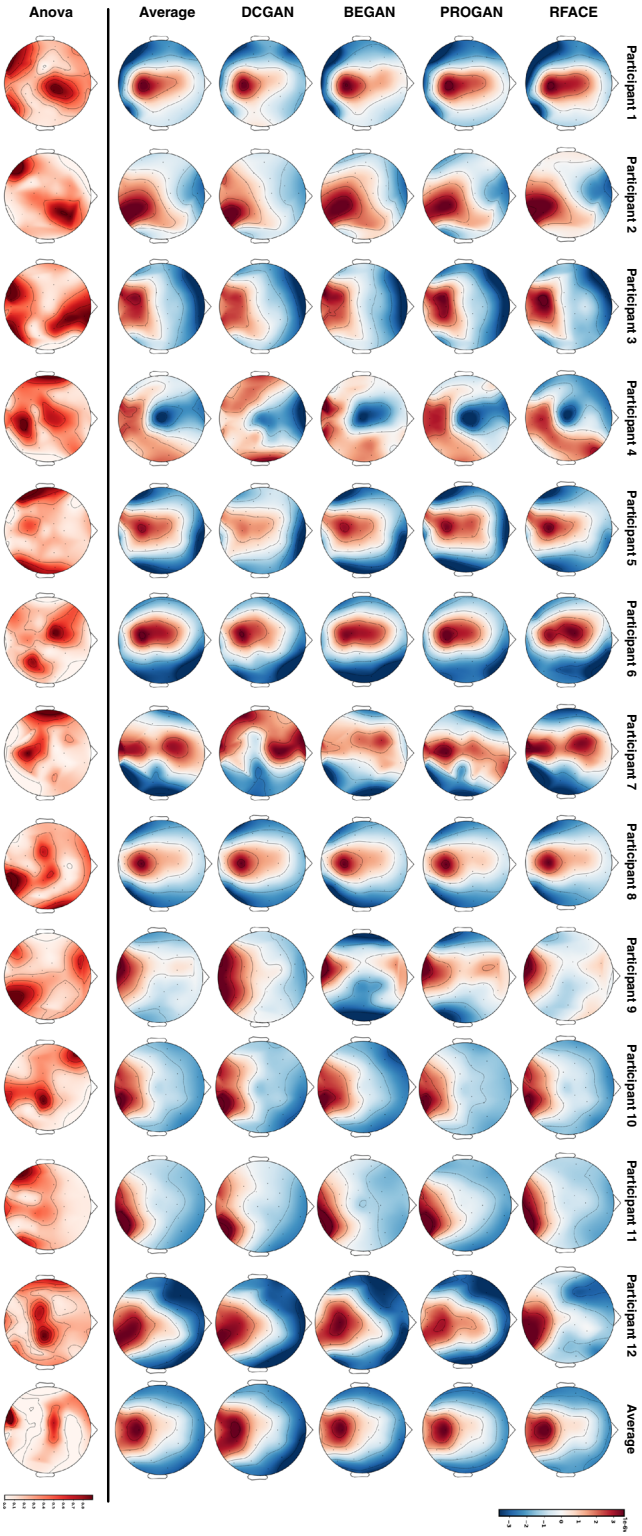

Fig. 6 Averaged P300 topography of each participant for each category. F-values from an ANOVA test were computed for each channel across four categories. Topography is created at the optimal P300 time index for each participant which is demonstrated in [44]

Table 3 Computed Neuroscore for each participant for each category

ID	DCGAN	BEGAN	PROGAN	RFACE
1	0.577	0.668	0.685	0.641
2	0.613	0.769	0.939	0.820
3	0.446	0.630	0.689	0.591
4	0.432	0.576	0.974	0.930
5	0.658	0.907	0.938	0.722
6	0.603	0.774	0.964	0.811
7	0.462	0.584	0.856	0.812
8	0.824	0.838	0.882	0.789
9	0.683	0.722	0.911	0.908
10	0.637	0.643	0.962	0.825
11	0.419	0.350	0.425	0.447
12	0.646	0.654	0.819	0.784
Mean	*0.583*	0.676	*0.837*	0.757

Higher score indicates better performance of GAN

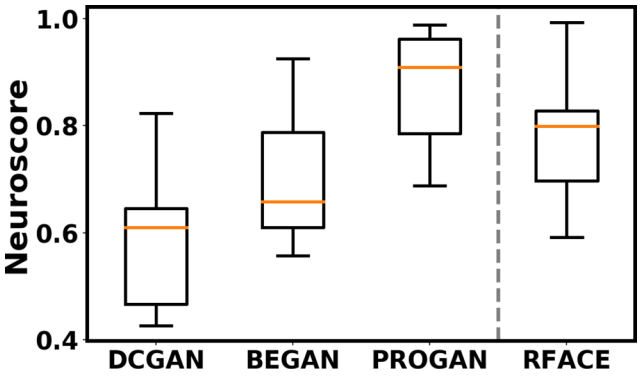

Fig. 7 Box plot of Neuroscore for each image category across 12 participants

performance of GANs by Neuroscore we see PROGAN > BEGAN > DCGAN, which is consistent with human judgment in the BE task.

Figure 7 summarizes the details from Table 3. The median values of the Neuroscore for each category across participants give the same rank as the mean value in Table 3.

From the averaged subtracted values (on a per-participant basis) of the Neuroscore and BE accuracies, it can be seen that the Neuroscore is correlated with the BE accuracy (human judgment), i.e., PROGAN > BEGAN > DCGAN (see Fig. 8).

In order to statistically measure this correlative relationship, we calculated the Pearson correlation coefficient and p value (two-tailed) between Neuroscore and BE accuracy and found ($r(48) = -0.767$, $p = 2.089\text{e}-10$).[1]

We used a bootstrap procedure [3, 11] to validate our Pearson correlation coefficient test since aggregating repeated measurements for participants (i.e., treating DCGAN, BEGAN, PROGAN, and RFACE measurements as being independent) like this results in a violation of assumptions for our statistical test (violation of independence). Using a bootstrap procedure with our correlation measure allows us to sidestep this violation of assumptions and still obtain a reliable statistic. We do this by repeatedly randomly shuffling the BE accuracy values and Neuroscore (within each participant) and then applying a Pearson correlation coefficient test. After following this process 10,000 times, we count how many p values calculated on randomly shuffled values (using within-participant shuffling) (i) are smaller than the original p value (where within-participant shuffling is not applied). $\frac{i}{10000}$ now becomes the bootstrapped Pearson p value, i.e., it estimates the probability of getting the calculated p value by chance. For the Pearson correlation coefficient test, this strongly supports the interpretation that our Neuroscore is predictive of human judgment. Due to time-based constraints in running the bootstrap procedure, we stopped at 10,000 iterations. This is consistent with our hypothesis that higher Neuroscore indicates better GAN models which is also indicated by lower BE accuracy. The bootstrapped p value for the Pearson correlation coefficient test is significant ($p \leq 0.0001$), which means that it is unlikely we have obtained these correlation results by chance.[2]

It is notable that PROGAN achieved a higher Neuroscore than RFACE. There are differences between the RFACE and GAN-generated images that are likely impacting the P300 amplitudes for the RFACE images. In the RFACE images, there are a wide range of background textures (e.g., sky, sea, and indoor environments) that are not present in the GAN-generated images. The GAN-generated images tend to have homogeneous backgrounds, where in most cases they are almost monochromatic and/or out of focus. Furthermore, the RFACE images contain a greater variety of other artifacts (e.g., jewellery) that tend not to be discernibly reproduced by the GANs. The lower Neuroscore for RFACE (i.e., RFACE < PROGAN) images is likely a result of these non-task-related visual components in the RFACE images increasing the discrimination difficulty. It is known that increasing task difficulty results in a diminished P300 amplitude [21]. For instance, increasing the amount of visual distractors in an image in a target

[1] We also did the Pearson statistical test and bootstrap on the correlation between Neuroscore and BE accuracy only for GANs, i.e., DCGAN, BEGAN, and PROGAN. Pearson statistic is ($r(36) = -0.827$, $p = 4.766\text{e}-10$) and the bootstrapped $p \leq 0.0001$.

[2] Without per-participant mean subtraction, the Pearson correlation statistic is ($r(48) = -0.556$, $p = 4.038\text{e}-05$) and the bootstrapped $p \leq 0.0001$.

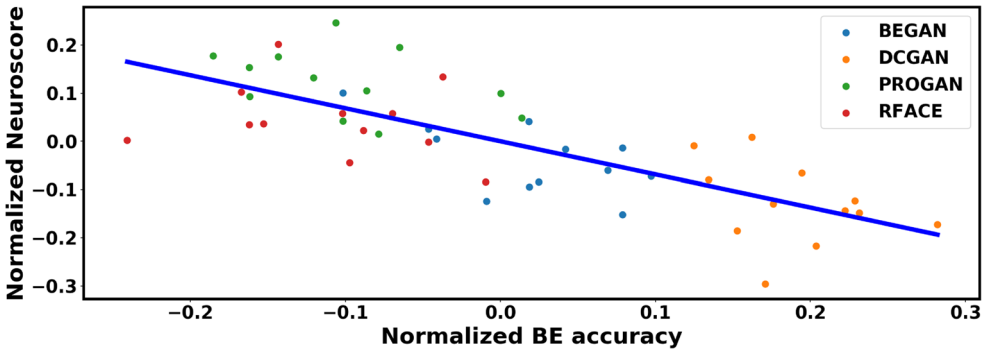

Fig. 8 Correlation between Neuroscore and BE accuracy. Neuroscore and BE are both mean centered within each participant

detection task reduces the P300 amplitude [28]. A further contributing factor may be the stereotyped visual structure of the GAN images (i.e., a face with a bland background), which facilitates the GAN images to be detected more easily in the fast RSVP paradigm used. From the human assessment results in the previous section, it can be seen that participants find the PROGAN output quite convincing, rating faces produced by the GAN similarly in accuracy as the RFACE images.

Comparison to Other Evaluation Metrics

Three traditional methods are also employed to evaluate the GANs used in this study. Table 4 shows the scores from the three traditional metrics, Neuroscore, and human judgment for three GANs. To be consistent with other metrics (smaller score indicates better GAN performance), we use 1/Neuroscore for comparison. It can be seen that all three methods are consistent with each other and they rank the GANs in the same order of PROGAN, DCGAN, and BEGAN from high to low performance. By comparing the three traditional evaluation metrics with the human, it can be seen that they are not consistent with human judgment of GAN performance. It should be remembered that inception score is able to measure the quality of the generated images [33] while the other two methods cannot do so. However, inception score still rates DCGAN as outperforming BEGAN. Our proposed Neuroscore is consistent with human judgment.

Table 4 Score comparison for each GAN category

Methods	DCGAN	BEGAN	PROGAN
1/IS	*0.44*	*0.57*	0.42
MMD	*0.22*	*0.29*	0.12
FID	*63.29*	*83.38*	34.10
1/Neuroscore	1.715	1.479	1.195
Human	0.995	0.824	0.705

Lower score indicates better performance of GAN

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 Neuroscore conveys a measure of the visual quality of facial images generated from GANs. This is based on our hypothesis that a generated image which looks more like a real face image will elicit a larger *reconstructed averaged P300 amplitude* in a RSVP task. Although the other three traditional evaluation methods do provide insight into several aspects of GAN performance, we study their effectiveness from a visual image quality perspective only as this is the focus of our work. The results are compelling in their demonstration that the proposed Neuroscore is better correlated with human judgment than any of 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 datasets. The method proposed can meet this need and is independent of any data modeling assumptions. In contrast, conventional quantitative metrics may fail in this regard.

For example, inception score is a model-based evaluation method, and the model is very sensitive to adversarial samples as shown in [22]. Inception score will also produce a very high score if the generated images are produced using adversarial training [4]. Our Neuroscore approach would not be compromised with such images in comparison. It is worth noting that compared with MMD and FID, both inception score and our Neuroscore provide a potentially good way of comparing the visual quality between generated images and real images, i.e., inception score and Neuroscore may give higher scores for the generated image that has better visual quality than the real image. Inception score, however, unlike the neural scoring approach, is not able to improve on the ranking of the three GANs compared with MMD or FID.

As mentioned earlier, more realistic GANs will produce a higher Neuroscore. This is because Neuroscore is sensitive to different stimulus processing requirements for different

types of GANs, i.e., the larger averaged single-trial P300 amplitudes for GANs reflect properties related to different stimulus information processing requirements [37]. It is also worth commenting that while GANs for generating facial images are explored in this study, our approach could be used for other types of generated images because the P300 ERP can be elicited using a wide variety of significantly different visual stimuli, e.g., Neuroscore may be applicable in the evaluation of GANs in bedroom image generation [20, 29, 32, 47].

The work presented here focuses on evaluating image visual quality only. Consequently there are some limitations when using the Neuroscore 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 quantitative methods with our Neuroscore 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 addressed by the GAN. If the goal of the GAN application is the generation of 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 a better choice. Therefore, the inclusion of a neural scoring approach as we have demonstrated should be considered in the context of the application's requirements.

Neuroscore is produced from human EEG signals and directly reflects human perception and neural processes. Compared with human judgment on images generated from GANs, our paradigm has several advantages as follows. Firstly, it is much faster than human judgment as a rapid image stream is presented to participants as part of the RSVP protocol. Traditional human judgment approaches entail the evaluation of images one-by-one whereas our paradigm supports batch evaluation of images. Secondly, as the EEG recorded corresponds to individual images, the method allows the tracking of image quality at the level of the individual image rather than the aggregated quality of a group of images. Thirdly, Neuroscore produces a continuous value while human judgment is binary ("real" or "fake"). Finally, it is possible to use EEG signals such as P300 as supervised information for improving training of GANs in the future.

In this work, we focus on the evaluation of images generated from GANs. However, time series evaluation of GANs is even more challenging and even less discussed in the literature. We believe that our paradigm may extend to use the auditory BCI [8] for auditory evaluation for GANs in the future.

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 image synthesis. We proposed and assessed a neural interfacing approach in which a Neuroscore is introduced as an alternative evaluation of GANs in terms of image visual quality. We interpret our results to conclude that Neuroscore is more consistent with assessments made by humans when compared with the three established quantitative metrics, and we show that the correlation between our Neuroscore and human judgment is not produced by chance, i.e., $p \leq 0.0001$. We believe that our proposed neuro-AI interface based on a rapid serial visual presentation approach is more efficient and less prone to error compared with 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.

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

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Ethical Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Formal approval for this work was given from Dublin City University Research Ethics Committee (REC/2018/115).

References

1. Abbass HA. Social integration of artificial intelligence: functions, automation allocation logic and human-autonomy trust. Cogn Comput. 2019;11:159–71.
2. Arjovsky M, Chintala S, Bottou L. Wasserstein GAN. arXiv:170107875. 2017.
3. Bakdash JZ, Marusich LR. Repeated measures correlation. Front Psychol. 2017;8:456. https://doi.org/10.3389/fpsyg.2017.00456.
4. Barratt S, Sharma R. A note on the inception score. arXiv:180101973. 2018.
5. Berthelot D, Schumm T, Metz L. BEGAN: boundary equilibrium generative adversarial networks. arXiv:170310717. 2017.
6. Blackwood D, Muir W. Cognitive brain potentials and their application. Br J Psychiatry. 1990;157(S9):96–101.
7. Borji A. Pros and cons of GAN evaluation measures. arXiv:180203446. 2018.
8. Cai Z, Makino S, Rutkowski TM. Brain evoked potential latencies optimization for spatial auditory brain–computer interface. Cogn Comput. 2015;7(1):34–43.

9. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: a large-scale hierarchical image database. In: Proceedings of the conference on computer vision and pattern recognition. IEEE; 2009. p. 248–55.
10. Doborjeh ZG, Doborjeh MG, Kasabov N. Attentional bias pattern recognition in spiking neural networks from spatio-temporal EEG data. Cogn Comput. 2018;10(1):35–48.
11. Efron B, Tibshirani RJ. An introduction to the bootstrap. CRC Press. 1994.
12. Forsyth DA, Ponce J. Computer vision: a modern approach, 2nd Ed. Pearson Education. 2012.
13. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: Advances in neural information processing systems, p. 2672–80. 2014.
14. Gretton A, Borgwardt KM, Rasch M, Schölkopf B, Smola AJ. A kernel method for the two-sample-problem. In: Advances in neural information processing systems, p. 513–20. 2007.
15. Healy G, Wang Z, Gurrin C, Ward T, Smeaton AF. An EEG image-search dataset: a first-of-its-kind in IR/IIR. NAILS: neurally augmented image labelling strategies. 2017.
16. Healy G, Ward TE, Gurrin C, Smeaton AF. Overview of NTCIR-13 nails task. In: The 13th NTCIR 2016-2017 evaluation of information access technologies conference. Tokyo. 2017.
17. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: Advances in neural information processing systems, p. 6626–37. 2017.
18. Hu J, He K, Xiong J. Comparison of event-related potentials between conceptually similar chinese words, english words, and pictures. Cogn Comput. 2010;2(1):50–61.
19. Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, p. 1125–34. 2017.
20. Karras T, Aila T, Laine S, Lehtinen J. Progressive growing of GANs for improved quality, stability, and variation. arXiv:171010196. 2017.
21. Kim KH, Kim JH, Yoon J, Jung KY. Influence of task difficulty on the features of event-related potential during visual oddball task. Neurosci Lett. 2008;445(2):179–83.
22. Kurakin A, Goodfellow I, Bengio S. Adversarial examples in the physical world. arXiv:160702533. 2016.
23. Lees S, Dayan N, Cecotti H, McCullagh P, Maguire L, Lotte F, Coyle D. A review of rapid serial visual presentation-based brain-computer interfaces. J Neural Eng. 2018;15(2):021,001.
24. Li J, Zhang Z, He H. Hierarchical convolutional neural networks for EEG-based emotion recognition. Cogn Comput. 2018;10:1–3.
25. Li Y, Swersky K, Zemel R. Generative moment matching networks. In: International conference on machine learning, p. 1718–27. 2015.
26. Liu Z, Luo P, Wang X, Tang X. Deep learning face attributes in the wild. In: IEEE International conference on computer vision (ICCV). 2015.
27. Luck SJ. An introduction to the event-related potential technique. MIT Press. 2014.
28. Luck SJ, Hillyard SA. Electrophysiological evidence for parallel and serial processing during visual search. Percept Psychophys. 1990;48(6):603–17.
29. Mao X, Li Q, Xie H, Lau RY, Wang Z, Smolley SP. Least squares generative adversarial networks. In: IEEE International conference on computer vision, p. 2813–21. 2017.
30. Metz L, Poole B, Pfau D, Sohl-Dickstein J. Unrolled generative adversarial networks. arXiv:161102163. 2016.
31. Polich J. Updating P300: an integrative theory of P3a and P3b. Clin Neurophysiol. 2007;118(10):2128–2148.
32. Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv:151106434. 2015.
33. Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X. Improved techniques for training GANs. In: Advances in neural information processing systems, p. 2234–42. 2016.
34. Shrivastava A, Pfister T, Tuzel O, Susskind J, Wang W, Webb R. Learning from simulated and unsupervised images through adversarial training. In: Proceedings of the conference on computer vision and pattern recognition. IEEE; 2017. p. 5.
35. Solon AJ, Gordon SM, Lance B, Lawhern V. 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; 2017. p. 5–8.
36. Spence R, Witkowski M. Rapid serial visual presentation: design for cognition. Heidelberg: Springer; 2013.
37. Sur S, Sinha V. Event-related potential: an overview. Indus Psych J. 2009;18(1):70.
38. Sutton S, Braren M, Zubin J, John E. Evoked-potential correlates of stimulus uncertainty. Science. 1965;150(3700):1187–88.
39. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the conference on computer vision and pattern recognition. IEEE; 2016. p. 2818–26.
40. Theis L, Oord A, Bethge M. A note on the evaluation of generative models. arXiv:151101844. 2015.
41. Treder MS, Porbadnigk AK, Avarvand FS, Müller KR, Blankertz B. The LDA beamformer: optimal estimation of ERP source time series using linear discriminant analysis. Neuroimage. 2016;129:279–291.
42. Wang Z, Healy G, Smeaton AF, Ward TE. An investigation of triggering approaches for the rapid serial visual presentation paradigm in brain computer interfacing. In: 27th Irish signals and systems conference. IEEE; 2016. p. 1–6.
43. Wang Z, Healy G, Smeaton AF, Ward TE. A review of feature extraction and classification algorithms for image RSVP based BCI. Signal Processing and Machine Learning for Brain-machine Interfaces, 243–70. 2018.
44. Wang Z, Healy G, Smeaton AF, Ward TE. Spatial filtering pipeline evaluation of cortically coupled computer vision system for rapid serial visual presentation. Brain-Comput Interf. 2018;5:132–45.
45. Wolpaw JR, Birbaumer N, McFarland DJ, Pfurtscheller G, Vaughan TM. Brain-computer interfaces for communication and control. Clin Neurophysiol. 2002;113(6):767–91.
46. Xu Q, Huang G, Yuan Y, Guo C, Sun Y, Wu F, Weinberger K. An empirical study on evaluation metrics of generative adversarial networks. arXiv:180607755. 2018.
47. Yu F, Seff A, Zhang Y, Song S, Funkhouser T, Xiao J. LSUN: construction of a large-scale image dataset using deep learning with humans in the loop. arXiv:150603365. 2015.
48. Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, p. 2223–32. 2017.