UNIVERSITY OF CAPE TOWN

MASTERS THESIS

Generative Adversarial Networks for Fine Art Generation

Author:
Alan BERMAN

Supervisor:
A/Prof. Deshen MOODLEY

A thesis submitted in fulfilment of the requirements for the degree of Master of Science

in the

Department of Computer Science

Declaration

I, Alan Berman....., hereby declare that the work on which this dissertation/thesis is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university.

I empower the university to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signature: 

Date: 24/1/2020.....

Abstract

Generative Adversarial Networks (GANs), a generative modelling technique most commonly used for image generation, have recently been applied to the task of fine art generation. Wasserstein GANs and GANHack techniques have not been applied in GANs that generate fine art, despite their showing improved GAN results in other applications. This thesis investigates whether Wasserstein GANs and GANHack extensions to DCGANs can improve the quality of DCGAN-based fine art generation. There is also no accepted method of evaluating or comparing GANs for fine art generation. DCGAN's, Wasserstein GANs' and GANHack techniques' outputs on a modest computational budget were quantitatively and qualitatively compared to see which techniques showed improvement over DCGAN. A method for evaluating computer-generated fine art, HEART, is proposed to cover both the qualities of good human-created fine art and the shortcomings of computer-created fine art, and to include the cognitive and emotional impact as well as the visual appearance. Prominent GAN quantitative evaluation techniques were used to compare sample images these GANs produced on the MNIST, CIFAR-10 and Imagenet-1K image data sets. These results were compared with sample images these GANs produced on the above data sets, as well as on art data sets. A pilot study of HEART was performed with 20 users. Wasserstein GANs achieved higher visual quality outputs than the baseline DCGAN, as did the use of GANHacks, on all the fine art data sets and are thus recommended for use in future work on GAN-based fine art generation. The study also demonstrated that HEART can be used for the evaluation and comparison of art GANs, providing comprehensive, objective quality assessments which can be substantiated in terms of emotional and cognitive impact as well as visual appearance.

Dedication

To, in no particular order: Sonia, Eamonn, Jethro, Jess, Quintin, and Sam.

Acknowledgements

Firstly, I would like to thank the *Centre for Artificial Intelligence Research* wing of the *Council for Scientific and Industrial Research*, whose generous bursary enabled me to conduct this research.

I am extremely grateful to Piotr Black for his enthusiastic and friendly help with *PyTorch*, without which I would not have been able to perform many of the experiments in this work. I would also like to thank Jethro and Sonia for their unwavering support.

Finally, I would like to thank my supervisor Prof. Deshen Moodley for his invaluable guidance and patience.

Contents

List of Figures

Nomenclature

D Discriminator network of a GAN

G Generator network of a GAN

`LReLU` Leaky Rectified Linear Unit activation function

`ReLU` Rectified Linear Unit activation function

Base GAN DCGAN, WGAN, WGANGP (and conditional versions thereof), and the IWGAN

CAN Creative Adversarial Network

CC Computational Creativity

D Use of dropout in G GANHack

DCGAN Deep Convolutional GAN

F Flipped labels in D GANHack

FID Fréchet Inception Distance

FLD DCGAN with F, L, D GANHacks

FLND DCGAN with F, L GANHacks

FLND DCGAN with F, L, N GANHacks

FLND DCGAN with F, L, N, D GANHacks

GAN Generative Adversarial Network

GANHacks A set of GAN training tips

HEART Holistic Evaluation of Art

IS Inception Score

IWGAN Improved Wasserstein GAN

L Use of `LReLU` in G GANHack

MC Mode collapse

N Use of instance noise in G GANHack

VAE Variational Auto-Encoder

WGAN Wasserstein GAN

WGANGP Wasserstein GAN with gradient penalty

Chapter 1

Introduction

1.1 Generative Adversarial Networks

First proposed by Goodfellow *et al.*, GANs are a generative modelling technique that create data by means of a game between two players, a 'generator' G and a 'discriminator', D, as well as a data set [Goodfellow et al., 2014, Goodfellow, 2017]. In this game, G aims to generate data samples drawn from the same distribution as the data set, and D aims to perfectly discern fake data samples from real ones (samples from the data set) [Goodfellow et al., 2014, Goodfellow, 2017]. The game is commonly described using a counterfeiting analogy, where G is a counterfeiter and D is the police [Goodfellow et al., 2014, Goodfellow, 2017]. Throughout the game, the counterfeiter G tries to pass off batches of its created banknotes (generated samples) as true currency, and the police D examine batches of real currency and of counterfeit banknotes and label them accordingly. Over time, the counterfeiter G learns to make banknotes that are increasingly similar to true currency until the police D can no longer distinguish real currency from fake currency [Goodfellow et al., 2014, Goodfellow, 2017].

Using a simple prior distribution z, G generates a sample intended to come from p_{data}, the distribution of the training data [Goodfellow, 2017]. D, usually a binary classifier, receives two inputs sequentially: an input x, which is a sample from p_{data}, and then $G(z)$, a sample from G. For each sample received as input, D outputs the probability that the sample is real [Goodfellow, 2017]. D aims for $D(x)$ to approach 1 and $D(G(z))$ to approach 0, while G aims for $D(G(z))$ to approach 1. The game ends when the Nash equilibrium for the game is reached and $D(x) = \frac{1}{2}$ for all x, *i.e.* D cannot distinguish fake samples from real samples and G is able to generate data samples drawn from p_{data} [Goodfellow, 2017, Mirzaei et al., 2017]. Figure 1.1 shows a high-level overview of a GAN [Dickson, 2018].

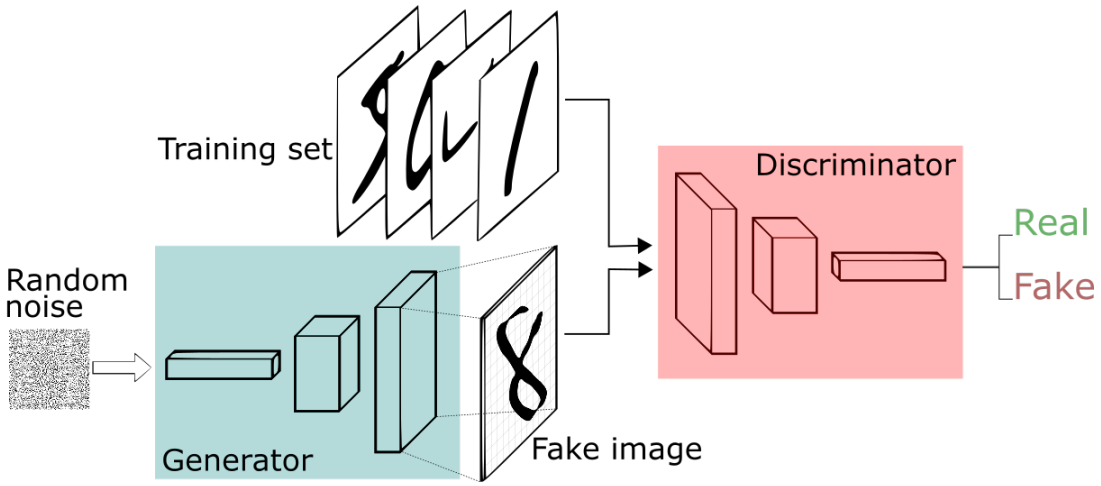


Figure 1.1: High-level overview of a GAN

GANs learn to generate samples from p_{data} through the following minimax value function, where V is the value function whose inputs are D, the discriminator, and G the generator [Goodfellow et al., 2014, Goodfellow, 2017]:

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}} \log D(x) + E_{z\sim p_z} \log(1 - D(G(z))) \tag{1.1}$$

The term 'value function', though typically associated with reinforcement learning, refers to the utility associated with a state under a particular policy [Pardo, 2007], but is used in GANs as the term for its underlying procedure for playing the game. In GANs, equation 2.1 is also interchangeably referred to as the GAN's 'objective function', or more simply, 'loss function' [Gormley, 2017]. The first term of equation 2.1 corresponds to D's ability to recognize samples from p_{data} as real samples, and the second term to D's ability to recognize samples from G as fake [Mirzaei et al., 2017].Goodfellow *et al.* show that the minimax game has a global optimum when $p_g = p_{data}$ through a proposition for the optimal D and a theorem for finding the global minimum for G [Mirzaei et al., 2017, Goodfellow et al., 2014].

The advent of Radford *et al.*'s Deep Convolutional GAN (DCGAN) [Radford et al., 2015] and improvements to GAN training via Wasserstein GANs [Arjovsky et al., 2017, Gulrajani et al., 2017] (hereafter collectively referred to as the base GANs), have allowed modern GANs to achieve state-of-the-art performance in image generation, using popular image data sets such as MNIST [LeCun et al., 1998], CIFAR-10 [Krizhevsky, 2009] and Imagenet-1k [Deng et al., 2009]. Other improvements to GAN training include the 'GANHacks', a set of GAN training tips [Chintala et al., 2016]. While GANs are most commonly applied to images, they have recently been applied to the task of fine art generation.

1.2 GANs for Fine Art Generation

The data sets typically used for training GANs contain images of singular structured objects, such as faces or animals. Though fine artworks may feature natural objects, they often feature more than one object per artwork. Moreover, many fine art styles are not photo-realistic, such as cubism, and thus the objects they feature may have shapes that are distorted or unconventional [Tan et al., 2017b]. For example, generating Chinese landscape art is noted as especially difficult, as

these paintings have many irregular shapes, an unclear foreground-background distinction, and the tendency to feature occlusions such as fog [Wang et al., 2017b]. Typical image data sets used to train GANs contain objects with high inter-class diversity, such as the digits 0 through 9 in the black-and-white MNIST handwritten digit data set. Though these data sets have images of different classes, intra-class diversity is low. For example, the images of the digit '3' are all the same colour, and have the same basic shape. In contrast, though they may depict the same object, different art styles will yield dramatically different images. Consequentially, fine art data sets such as Wikiart are different from those commonly used in training GANs, such as Imagenet-1k and CIFAR-10 [Deng et al., 2009, Krizhevsky, 2009].

The use of GANs for fine art generation is also interesting from a philosophical perspective. As with intelligence, creativity is considered a hallmark of the human condition [Boden, 2009]. Despite the general acceptance of artificial intelligence's attempt to model intelligence, there is a recognised unwillingness, or even outright rejection, of the idea that creativity can indeed be modeled [Boden, 2009]. Computer-generated creative works, such as the music-generating software *Emmy*, have often been dismissed entirely or viewed as mere computer output [Hofstadter, 2002, Boden, 2009]. The idea that a non-human entity, such as a GAN, could possibly be creative and create fine art is controversial.

1.3 Computational Creativity

Computational Creativity (CC) is a subset of Artificial Intelligence (AI) research that explores creativity specifically through the medium of computational systems [Colton and Wiggins, 2012, Cardoso et al., 2009]. Creativity is the process that creates *novel* and *useful* or *valuable* artefacts [Elgammal and Saleh, 2015]. Valuable artefacts include those created in the visual arts, such as fine art.

The three main types of visual art generation explored in CC research are evolutionary art, non-Photo-realistic Rendering (NPR), and automated painting [Colton, 2008]. Evolutionary art uses the mechanisms of evolutionary computation (EC) to generate visual art. (EC) adapts features of the processes of evolution that occur in nature to solve problems [Eiben et al.,]. In EC, individuals (candidate solutions) compete and mutate over generations [Eiben et al.,]. How well the individuals fare in their environments, their 'fitness', is calculated using a fitness function [Eiben et al.,]. Evolutionary art, which is usually human-guided, uses the aesthetic judgments of the user on the generated art to guide the evolution of the individuals [Colton, 2008]. This process continues until the user is satisfied with the generated art, which is often abstract art [Colton, 2008].

NPR seeks to create images that appear to be generated by humans, such as Impressionist versions of photographs and simulating artistic media such as charcoal, using the photo itself [Colton, 2008]. NPR is strongly focused on the output, and pays no heed to the "the many cognitive aspects of the artistic process, such as choosing subject matter and painting style" [Colton, 2008]. NPR systems, such as *Adobe Illustrator*, are frequently used by human visual artists [Colton, 2008]. Systems that use NPR systems and generate full artworks, considering high-level details such as subject matter are termed 'automated painters' [Colton, 2008]. GANs that create

fine art are thus automated painters.

1.4 Evaluating GANs

The evaluation of GANs is noted to be one of its biggest obstacles [Lucic et al., 2017]. While various quantitative methods have been widely used in the literature, there exists no single, standard metric. Moreover, computational budget limitations and the commonplace practice of using cherry-picked results render fair quantitative comparisons between GANs difficult. Qualitative evaluation of GANs is typically limited to judgments of the visual quality (VQ) of images produced by GANs. For GANs that generate art, qualitative evaluation has expanded upon VQ judgments to include judgment of the GANs' creativity and of the cognitive impact of the images they produce.

1.5 Research Questions

1. To what extent does a GAN's quantitative performance align with its qualitative performance on benchmark data sets?

2. Does the use of GANHacks [Chintala et al., 2016] improve qualitative, quantitative and/or runtime performance of DCGAN?

3. Does the use of Wasserstein GANs improve quantitative, qualitative and/or runtime performance of DCGAN?

4. Does improved qualitative performance of GANs on benchmark data sets translate to better qualitative performance on art data sets?

5. Can a proposed qualitative evaluation method successfully be used to evaluate the emotional impact, cognitive impact, visual quality and creativity of the creations of art GANs?

1.6 Tools and Approach

Wherever possible, existing implementations of the GANs reported in the literature are used in this research. Official implementations of quantitative metrics, such as the Fréchet Inception Distance, are also used where possible. The popular and prominent deep learning framework *PyTorch*, commonly used in GAN development, is used as the primary framework. The GANs are evaluated using benchmark data sets and prominent quantitative evaluation metrics, and are all trained on the same high-performance cloud computing environment with equal GPU resources.

1.7 Contributions

This work has the following contributions:

1. The proposal and small-scale testing of HEART (the Holistic Evaluation of Art), a qualitative evaluation method for art GAN that can be used to evaluate the emotional and cognitive impacts, as well as important characteristics of GAN-produced images.

2. The demonstration of Wasserstein GANs' higher VQ than DCGAN on Imagenet-1K and art data sets.

3. The demonstration that good qualitative performance on benchmark data sets does not necessarily translate to good qualitative performance on art data sets.

4. The demonstration that GANHacks do not improve quantitative performance of DCGAN, though some improve VQ.

5. The demonstration that typically, quantitative performance of GANs aligns well with qualitative performance on benchmark data.

This work provides quantitative results for GANs trained on a *limited* computational budget, which sheds light on the impacts of cherry-picked results and computational power on quantitative GAN evaluation.

1.8 Structure of Dissertation

Chapter 2 gives a thorough overview of GANs. Notable network architectures, value functions, optimisation and evaluation methods for GANs, both quantitative and qualitative, are described in detail. GANs for computational creativity, GANs that produce creative artefacts, are then discussed. The fine art GANs are explored, with a specific focus on the state-of-the-art in fine art GANs, the Creative Adversarial Network and the ArtGAN. Finally, the evaluation of visual art is discussed. Chapter 3 details the methodology of this work, including the GANs explored, and introduces HEART, a proposed qualitative evaluation method for art, including that of art GANs. In chapter 4, the experimental design and implementation are given, followed by a description of the data sets used, the quantitative experiments conducted and the HEART pilot study. The results for these experiments are given in chapter 5. Chapter 6 discusses the results, and the final chapter gives the conclusions and suggestions for future work.

Chapter 2

Literature Review

In this chapter, current GANs are explored in detail. The architecture, training, evaluation and challenges facing GANs are described, followed by a brief overview of the GANs considered to be the state-of-the-art. GANs used in creative tasks, including GANs for fine art generation, or art GANs, are discussed. Finally, the evaluation of visual art is discussed.

2.1 GANs

A GAN consists of two neural networks, a generator and a discriminator, which compete in a game. Each player (network) plays the game according to the given GAN's 'value function' (also termed a 'loss function'). The generator aims to minimise the value function and the discriminator aims to maximise it. The value function calculates the difference between the probability distribution of fake samples created by the generator and the probability distribution of real samples. If a generator's samples are very similar to real samples, and the discriminator identifies few fake samples, the difference between these probability distributions is minimised. Conversely, if the discriminator reliably classifies fake samples as not being real, the difference between the probability distributions is maximised. To better train GANs and produce better samples, experiments with various aspects of GANs have been investigated. These include different activation functions. These functions determine the output of a neuron of a neural network, given a particular input. The output of the discriminator given an input image is also determined by an activation function. Outputs for the same input will differ across activation functions, and the use of one activation function in the layers of a GAN's network over another may help or hinder its performance. Other aspects investigated include additional loss functions, which supplement the GAN value function. These loss functions are typically used in an effort to improve the quality of the fake samples made by the generator. Optimisation, whether through training tricks or choice of neural network optimiser, has also been explored in GANs to improve their performance or combat training instability.

All percentages quoted here are taken from the *GAN Zoo*, a reputable repository of over 350 GAN papers [Hindupur, 2017].

2.1.1 Structure

The structure of a GAN is given by the form of its generator (G) and discriminator (D) networks, and the activation functions and normalisation used therein. G, the generator, acts as a function that manipulates the values of a noise vector input to

create an image output. This noise vector has the same dimensions as the output image. The goal of the generator is to create output images that resemble those of the real image distribution, and in so doing, fool D, the discriminator. The discriminator acts as a function that receives image inputs, both real images from the data set, the probability distribution of which is termed p_{data}, and fake ones created by the generator, the probability of which is termed p_g, and outputs a 'real' or 'fake' label for each. The forms of G and D are typically deep neural networks [Goodfellow, 2017]. The original GAN uses multi-layer perceptrons (MLPs) as the structure for its G and D [Goodfellow et al., 2014]. Though MLPs are still used in modern GANs, the majority of modern GANs use deep, convolutional neural networks (CNNs) for their G and D [Hindupur, 2017, Lucic et al., 2017].

Additionally, the use of multiple discriminators and generators, as well as decomposing the generative process into multiple steps, have been noted to result in better training [Hou et al., 2017, Neyshabur et al., 2017, Zhao et al., 2018]. Figure 2.1 shows a GAN [Shaikh, 2017].

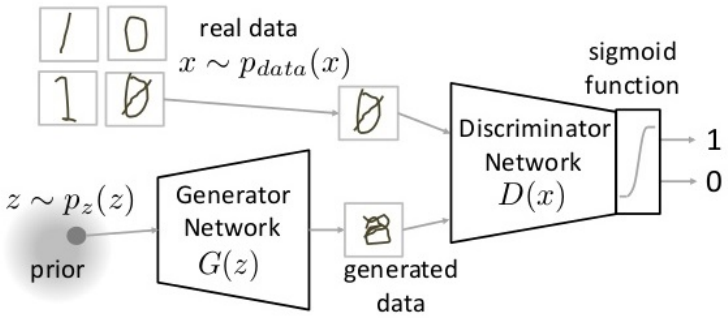


Figure 2.1: A Generative Adversarial Network

DCGAN

The benchmark GAN structure, which is a deep CNN, is undoubtedly Radford *et al.*'s Deep Convolutional GAN (DCGAN) [Goodfellow, 2017, Li et al., 2017a, Radford et al., 2015, Hindupur, 2017]. 29% of GANs use DCGAN at least partly as the structure of their GANs [Hindupur, 2017, Lucic et al., 2017]. DCGAN's fully-convolutional nature, absence of pooling layers and its choice of activation functions all allowed for improved performance over other deep, convolutional GANs of the time [Radford et al., 2015, Goodfellow, 2017]. Figure 2.2 shows the generator of DCGAN [Radford et al., 2015].

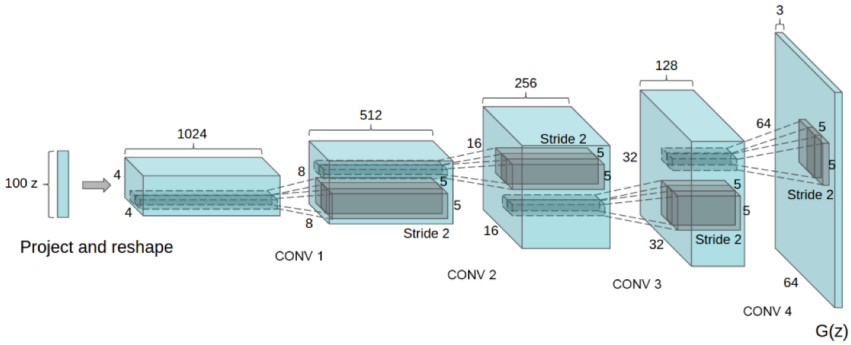


Figure 2.2: The generator of Radford *et al.*'s DCGAN

DCGAN is not only a common structure of modern GANs, but also frequently used as a baseline GAN against which new GANs are compared [Hindupur, 2017].

Variational Auto-Encoder

Another significant development in GAN structure is the merging of GANs with variational auto-encoder (VAEs), which has enjoyed popularity since the development of the VAE/GANs in 2015 [Larsen et al., 2015]. 30% of GANs feature VAEs as their G or D networks [Hindupur, 2017, Lucic et al., 2017, Berthelot et al., 2017]. VAEs, an alternative generative modelling technique to GANs, consist of two networks: an encoder network and a decoder network [Tripathy et al., 2017]. The encoder network generates (encodes) a hidden representation of the input, which is then processed (decoded) into the final output [Tripathy et al., 2017]. VAEs aim to recreate their inputs, and in doing so learn salient features of the data [Achlioptas et al., 2017, Choi et al., 2017]. The use of an auto-encoder based network, such as the popular U-net, also preserves spatial dependencies through its skip-connections [Fabbri et al., 2018]. The merging of GANs with VAEs aims to combine the strengths of both approaches: stable training and high sample quality [Tran et al., 2018, Ulyanov et al., 2017]. VAEs are noted to enjoy stable training free from the vanishing gradient and mode collapse problems often found in GANs [Mariani et al., 2018, Brock et al., 2016]. Vanishing gradient occurs when the error (output) of a neural network disappears as it propagates back through the network [Hochreiter, 1998]. Mode collapse refers to the failure case of GANs where instead of capturing the full richness of p_{data}, G focuses on a small subset of it. VAEs are also capable of inference, in addition to generation, due to their learning "a bidirectional mapping between a complex data distribution and a much simpler prior distribution" [Ulyanov et al., 2017]. In the field of image generation, GANs are noted to generate higher-quality samples than VAEs, whose samples are often blurry [Mariani et al., 2018, Ulyanov et al., 2017, Shang et al., 2017, Creswell et al., 2017].

2.1.2 Activation Functions

GANs typically use different activation functions in their G and D networks. A popular activation function configuration is the configuration used by DCGAN. DCGAN uses the rectified linear unit (`ReLU`) activation function in all layers of G, with the exception of its output layer, which uses `Tanh` [Radford et al., 2015]. Leaky ReLU (`LReLU`) is used as the activation function in all layers of D. `ReLU` and `LReLU` are the dominant activation functions, most often used in generators and discriminators respectively, with `ReLU` in 34% of GANs' generators and `LReLU` in 29% of all

GANs' discriminators. However, `ReLU` has been used in 10% of GANs' discriminators and `LReLU` in 14% of GANs' generators. The remaining popular activation functions, typically used in the output layers, are the `Tanh` activation function and the `sigmoid` activation function, found in 20% of GANs' generators and 19% of GANs' discriminators respectively.

2.1.3 Normalisation

Normalisation is widely used in GANs. 38% of GANs' generators and 28% of GANs' discriminators use normalisation [Hindupur, 2017, Goodfellow, 2017]. For both generators and discriminators, batch normalisation is by far the most common normalisation scheme, with 87% of generators and 82% of discriminators that are normalized using batch normalisation [Hindupur, 2017]. The remaining notable normalisation schemes are instance normalisation and layer normalisation.

Batch normalisation seeks to optimise the model by replacing the "complicated interaction between all of the weights of all of the layers" [Goodfellow, 2017] used to calculate the mean and variance of features, with *single* mean and variance parameters [Goodfellow, 2017]. Batch normalisation was seen by the authors of DCGAN as one of the key reasons for the model's success, used in both G and D, and is regarded as essential to the model [Radford et al., 2015, Salimans et al., 2016, Qi, 2017]. Batch normalisation is noted to conflict with GANs that employ a gradient penalty in their objective function, and is thus not used in these GANs [Bellemare et al., 2017, Arjovsky et al., 2017]. The use of batch normalisation in D has been found to lead to trivial or poorer solutions, and has been omitted from some GANs as a result [Wang and Gupta, 2016, Hjelm et al., 2017]. In some GANs, batch normalisation has been found to not be needed, or to decrease the quality of samples [Xian et al., 2017, Mao et al., 2016]. Interestingly, batch normalisation is not used among state-of-the-art GANs for image super-resolution [Neyshabur et al., 2017, Wang et al., 2018].

Instance normalisation, batch normalisation where the batch size is one, is noted to be an effective normalisation technique for image generation, improving upon batch normalisation in some cases [Zhu et al., 2017a, Wu et al., 2017a, Fabbri et al., 2018].

2.1.4 Value Functions

There have been numerous modifications to the standard, adversarial, GAN value function [Lucic et al., 2017, Goodfellow, 2017, Hindupur, 2017]. The primary goal of these modifications is to improve the noted training instability of GANs [Berthelot et al., 2017, Jaiswal et al., 2018, Barsoum et al., 2017]. However, many of these modifications also result in better-quality samples [Arjovsky et al., 2017, Gulrajani et al., 2017, Mao et al., 2016]. While these modified GANs' value functions do not prescribe a specific structure, they are commonly built on top of a DCGAN structure [Arjovsky et al., 2017, Lucic et al., 2017, Mao et al., 2016]. Notable GAN value function modifications include the use of an energy function and a least squares loss [Zhao et al., 2016, Mao et al., 2016]. However, the most important GAN value function is undoubtedly the Wasserstein GAN's (WGAN) value function [Hindupur, 2017, Lucic et al., 2017, Bellemare et al., 2017].

Wasserstein GANs

Wasserstein GANs are based on the Earth-Mover (EM) or Wasserstein-l distance, which is "continuous everywhere and differentiable almost everywhere" [Arjovsky et al., 2017]. The EM distance is viewed as superior to the Jensen-Shannon (JS) divergence used in the original GAN, as the former can converge in situations where the latter cannot, and the former is differentiable in more regions of the data manifold than the latter [Arjovsky et al., 2017, Yi et al., 2017]. The JS divergence is the distance between two probability distributions [Lin, 1991]. Instead of minimizing the JS divergence, the EM distance between p_{data}, the probability distribution of real samples, and p_g, the probability distribution of samples created by the generator, is minimised [Arjovsky et al., 2017, Gulrajani et al., 2017, Lucic et al., 2017]. Equation 2.1 shows the Wasserstein GAN value function [Gulrajani et al., 2017]:

$$\min_G \max_{D \epsilon \mathcal{D}} E_{x \sim P_r}\big[D(x)\big] - E_{\tilde{x} \sim P_g}\big[D(\tilde{x})\big] \tag{2.1}$$

Because the base EM distance is "highly intractable" [Arjovsky et al., 2017], WGAN, the original Wasserstein GAN, uses the Kantorovich-Rubinstein duality to calculate an approximation of the EM distance [Arjovsky et al., 2017, Lucic et al., 2017, Xian et al., 2017]. The Kantorovich-Rubinstein duality uses the space of K-Lipschitz functions to calculate the EM distance [Arjovsky et al., 2017], and thus D must be restricted in order to "lie within the space of 1-Lipschitz functions" [Gulrajani et al., 2017]. The EM distance is defined as the following:

$$W(P_r, P_g) = \inf_{\gamma \epsilon \Pi(P_r, P_g)} E_{(x,y) \sim \gamma}\big[\|x - y\|\big] \tag{2.2}$$

where "$\Pi(P_r, P_g)$ represents the set of all joint distributions $\gamma(x, y)$ whose marginals are respectively P_r and P_g" [Arjovsky et al., 2017]. $\gamma(x, y)$ indicates "how much 'mass' must be transported from x to y in order to transform the distributions from P_r to P_g" [Arjovsky et al., 2017]. The EM distance is then this transport plan's 'cost' [Arjovsky et al., 2017]. WGAN enforces the Lipschitz constraint by clipping the weights of D to lie within a small box (usually $[-0.01, 0.01]$) after each gradient update [Arjovsky et al., 2017]. This procedure is done after each gradient update as the EM distance changes after each update of G [Sun et al., 2017]. D is also updated n times for each G update [Yi et al., 2017, Lin, 2017]. The authors of WGAN argue that due to the EM distance's continuous nature and its high differentiability, D can and must be "train[ed] to optimality" [Arjovsky et al., 2017]. Unlike an optimal GAN's D trained using the JS divergence, which would quickly be able to perfectly discern fake from real samples and thus not be able to provide any gradient information to G, an optimal WGAN D does not saturate and "gives remarkably clean gradients everywhere" [Arjovsky et al., 2017].

The authors of WGAN state that "weight clipping is... a terrible way to enforce a Lipschitz constraint" [Arjovsky et al., 2017], as too large a box leads to long training times, while too small a box may lead to the vanishing gradient problem [Arjovsky et al., 2017]. Indeed, optimisation of WGAN has been noted to be extremely sensitive to choice of the clipping value [Gulrajani et al., 2017]. Though the authors of WGAN use weight clipping due to its 'satisfactory' performance and its easy implementation, the problem of vanishing gradients has nonetheless been observed in

WGAN [Arjovsky et al., 2017, Xian et al., 2017]. Moreover, using weight clipping was noted to have longer training times than DCGAN [Miyato et al., 2018].

A successor to WGAN, the Improved Wasserstein GAN (IWGAN), opts for a different approach to enforcing the Lipschitz constraint on D [Gulrajani et al., 2017, Xian et al., 2017, Lucic et al., 2017, Mroueh and Sercu, 2017]. Rather than clamping the weights of D, the norm of the gradient of D is penalized with respect to its input through a 'gradient penalty' [Gulrajani et al., 2017]. Though weight clipping is data-independent and a gradient penalty is data-dependent, the use of a gradient penalty gives IWGAN faster training and better sample quality than WGAN [Mroueh and Sercu, 2017, Gulrajani et al., 2017]. The authors of IWGAN note that deep WGAN commonly do not converge, even if batch normalisation is used [Gulrajani et al., 2017]. Though IWGAN too trains slower than DCGAN, it is able to use momentum-based optimisation methods such as ADAM for D [Gulrajani et al., 2017], unlike WGAN [Gulrajani et al., 2017]. Only IWGAN has a different structure from DCGAN [Gulrajani et al., 2017]. Using a ResNet architecture, IWGAN achieves benchmark quantitative evaluation performance.

Both WGAN and IWGAN are noted to produce images of high quality [Arjovsky et al., 2017, Gulrajani et al., 2017]. This has been demonstrated under benchmark data sets such as CIFAR-10, particularly so IWGAN, whose high quality samples allow it to achieve benchmark quantitative evaluation performance. A WGAN with gradient penalty and a DCGAN structure, hereafter referred to as a WGANGP, has also been found to produce higher-quality samples than a normal DCGAN on benchmark data sets.

Conditional GANs

The most common modification to the original GAN value function is to make the GANs conditional, with 35% of all GANs being conditional [Hindupur, 2017]. Conditional GANs (CGAN), first developed by Mirza *et al.*, place condition(s) on the generative process to steer it in specific direction(s) [Mirza and Osindero, 2014]. A common condition is a class label c as an additional input to G. For many tasks, such as image-to-image translation or artificial face aging, placing conditions on the generative process is essential [Zhu et al., 2017a, Antipov et al., 2017]. A translated image should retain features of the original image; thus the changed image should be conditioned on the original. Interestingly, it has been found that CGANs, such as class-conditional GANs, produce "systematically better quality samples" [Grinblat et al., 2017] than their unconditioned counterparts [Goodfellow, 2017]. This may be another reason for the high popularity of CGANs.

2.1.5 Additional Loss Functions

Along with an adversarial loss (objective) function, whether it be the original GAN value function or a modified one such as Wasserstein GANs', 50% of GANs use (at least one) additional loss function(s) [Hindupur, 2017]. Of this 50%, 66% use weighting parameters to emphasize or relax focus on each objective [Hindupur, 2017]. The majority of these additional loss functions may be classified as content, perceptual or reconstruction losses. These additional loss functions are used to improve the perceptual quality of G's outputs, and are frequently used in CGAN, particularly in GANs which use face images and in image-to-image translation GANs [Hindupur,

2017].

Common additional loss functions used in image-based convolutional networks (and GANs) are pixel-wise loss functions [Ma et al., 2017, Kupyn et al., 2017, Wang et al., 2017a]. These loss functions compare images pixel-by-pixel, using Euclidean distance measures between images [Wu et al., 2017a]. The most commonly used pixel-wise loss function is the L2 loss function, but L1 loss and mean square error (MSE) are also prominently used [Wu et al., 2017a]. Though the use of such loss functions may quicken training time, their tendency to produce unwanted artifacts is extremely well-documented [Wu et al., 2017b, Kupyn et al., 2017, Wang et al., 2017a, Zhu et al., 2017a, Olut et al., 2018]. The most-often mentioned problem of such losses is their tendency to produce blurry images [Wu et al., 2017a, Zhu et al., 2017a, Wang et al., 2017a, Wu et al., 2017b, Yin et al., 2017]. L2 loss is noted to suffer from this problem more than its L1 counterpart, which is the chief reason for L1 loss being used in twice as many GANs as L2 loss [Zhu et al., 2017a, Yin et al., 2017, Hindupur, 2017].

Pixel-wise loss functions have been noted to fail to capture high-level perceptual qualities, such as texture [Wu et al., 2017a, Ren et al., 2017, Wang et al., 2017a]. For such reasons, some view their use as inadequate for image comparison [Ma et al., 2017]. The Perceptual loss, proposed by Johnson *et al.*, aims to function as a loss function that *can* capture such high-level perceptual qualities, and is widely used [Johnson et al., 2016b, Di et al., 2017, Wang et al., 2017a, Ma et al., 2017]. Rather than operating on images pixel-by-pixel, the Perceptual loss calculates the L2 loss between the feature maps of a generated image and a target image, to "penalize the discrepancy between extracted high-level features" [Ren et al., 2017, Wang et al., 2017a, Kupyn et al., 2017]. These feature maps are extracted from a well-trained convolutional neural network [Johnson et al., 2016b, Di et al., 2017, Ma et al., 2017]. The Perceptual loss has been noted to generate high-quality images but be time- and memory-intensive [Kupyn et al., 2017, Wu et al., 2017a].

2.1.6 Optimisation

The most common optimisation method used in GANs is ADAM, a "first-order gradient-based optimiser of stochastic objective functions" [Kingma and Ba, 2014, Hindupur, 2017, Lucic et al., 2017]. The ADAM method is used in 50% of all GANs implementations [Hindupur, 2017, Lucic et al., 2017]. Under ADAM, GANs converge faster than traditional stochastic gradient descent, and it has been found to be useful in large-scale data sets [Zhang et al., 2017, Nowozin et al., 2016]. The second-most popular optimiser, and often used in contrast to ADAM, is RMSprop [Hindupur, 2017]. While RMSprop is only used in 7% of GANs, it is argued to have benefits over ADAM [Hindupur, 2017]. RMSprop is argued to be more stable than ADAM, and to not suffer from instability in the face of "highly non-stationary" problems [Mao et al., 2016, Yi et al., 2017, Neyshabur et al., 2017, Arjovsky et al., 2017]. RMSprop also allows for larger step-sizes than ADAM [Neyshabur et al., 2017]. Interestingly, the most prominent GANs survey found no obvious superior optimisation method between the two [Lucic et al., 2017]. However, the survey did note that under default parameters, ADAM performed better [Lucic et al., 2017].

Though not an optimisation method, the 'GANHacks', a prominent collection of GAN training tricks, were developed to combat the instability of GANs [Chintala et al., 2016]. Each GANHack typically benefits either the generator or discriminator [Chintala et al., 2016]. These hacks can be applied to any GAN, such as a DCGAN, and include changing activation functions and adding noise to inputs. Though the effects on the qualitative and quantitative performance following the addition of a GANHack, or multiple GANHacks, to a GAN have not yet been rigourously investigated, they are noted to make training of GANs more stable [Chintala et al., 2016].

2.1.7 Evaluation

The approaches to evaluation of GANs are either quantitative or qualitative. Figure 2.3 shows the most prevalent evaluation techniques or metrics.

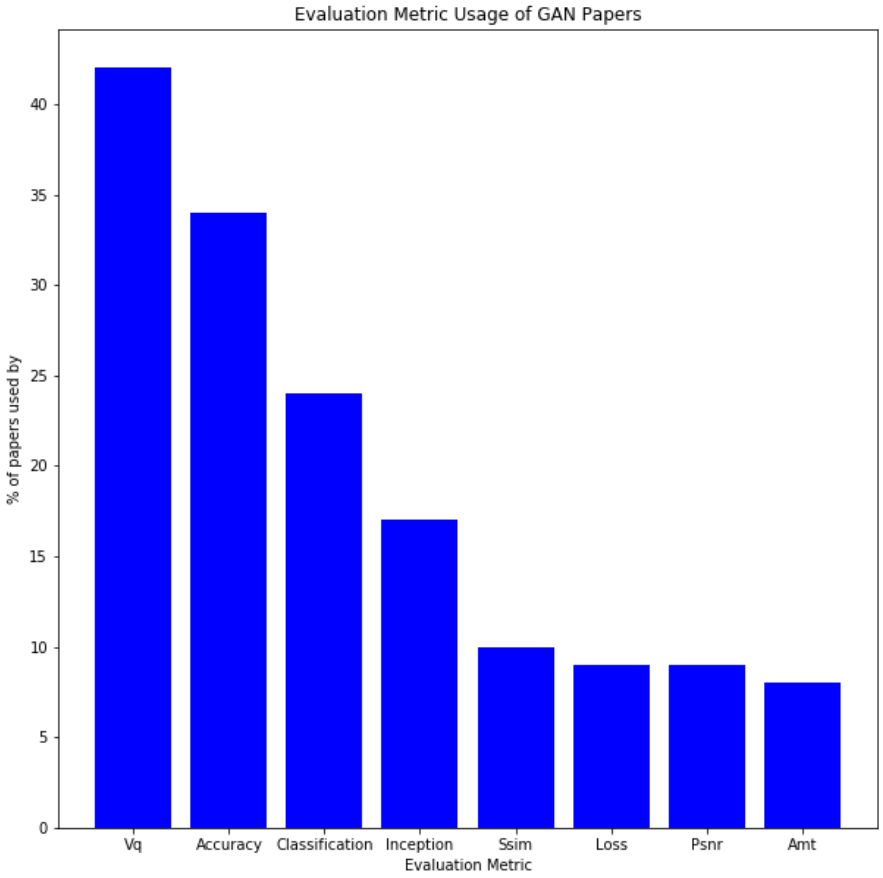


Figure 2.3: GAN Evaluation Use of Metrics

Quantitative Evaluation Techniques

While there are over 24 quantitative evaluation techniques used in GANs, there are a few quantitative evaluation techniques that enjoy widespread use [Borji, 2019, Hindupur, 2017]. These metrics, in descending order of popularity, are image recog-

nition, classification accuracy, the Inception Score, the Structural Similarity Metric (SSIM) and its multi-scale equivalent, MS-SSIM, and Peak Signal-to-Noise Ratio (PSNR) [Hindupur, 2017].

Image recognition includes measures such as facial recognition accuracy [Hindupur, 2017]. Classification accuracy refers to the use of a trained GAN's discriminator as an unsupervised feature extractor on top of which a classifier is built and trained [Radford et al., 2015]. For example, in Radford *et al.*'s DCGAN paper, the discriminator is used in this way to build a classifier for unseen images. This technique, which came to prominence through DCGAN, has been noted to be an indirect evaluation metric which is highly sensitive to the choice of classifier [Im et al., 2016]. SSIM and MS-SSIM, which evaluate the similarity of two images, assign higher scores to more similar images and lower scores to more distinct images [Mariani et al., 2018, Rosca et al., 2017]. However, their correlation with perceptual quality has been debated, and they are noted to not evaluate similarity between produced images and those in the training set [Russo et al., 2017, Dolhansky and Canton-Ferrer, 2017, Karras et al., 2017]. PSNR, commonly used for image reconstruction and image super-resolution, calculates the MSE between two images at the logarithmic decibel scale [Xie et al., 2018]. It has been noted to not correlate strongly with perceptual quality [van Amersfoort et al., 2017]. This, as with L1 loss, L2 loss and MSE, is due to its being a pixel-wise metric that cannot capture high-level perceptual qualities [Ledig et al., 2016, Wu et al., 2017a, Ren et al., 2017, Wang et al., 2017a].

Apart from accuracy calculations, the most widely-used quantitative evaluation metric is the Inception score (IS) [Hindupur, 2017, Spampinato et al., 2018]. The IS is based on a pre-trained 'Inception' model, a convolutional neural network trained on the Imagenet data set, and aims to evaluate the diversity of a GAN and its ability to generate meaningful objects [Juefei-Xu et al., 2017, Lucic et al., 2017, Salimans et al., 2016]. The Inception Score is defined as the following [Barratt and Sharma, 2018]:

$$IS(G) = exp\big(E_{x\sim p_g} D_{KL}(p(y|\mathbf{x})\|p(y))\big) \tag{2.3}$$

To obtain the IS, the pre-trained Inception model is applied to a set of images generated by a GAN to obtain a conditional label distribution $p(y|\mathbf{x})$, where y is the class label for a given input image x. A GAN that creates meaningful images will have a low entropy $p(y|\mathbf{x})$. In a GAN that creates diverse images from varied classes, the marginal $\int p(y|x = G(z))dz$ will have high entropy [Salimans et al., 2016]. The IS uses the Kullback-Leibler (KL) divergence of the conditional and the marginal probability distributions to combine these quality and diversity measures into a single relative entropy score that reflects both measures [Salimans et al., 2016].

Though the IS is widely used, including by state-of-the-art GANs, it nevertheless is *not* a perfect metric [Hindupur, 2017, Lucic et al., 2017]. A number of failure cases have been observed, where a high IS can be obtained for a poorly-performing GAN [Donahue et al., 2018]. A well-documented failing of the IS is its inability to detect mode collapse in GANs [Odena et al., 2016, Rosca et al., 2017, Johnson and Zhang, 2018, Che et al., 2016, Lucic et al., 2017]. A GAN that produces only one image per class will achieve a high IS score, as will one that has memorized an example per Imagenet class [Donahue et al., 2018, Odena et al., 2016, Che et al.,

2016, Lucic et al., 2017].The IS is stated by its inventors to correlate strongly with human perception, but this has been debated [Salimans et al., 2016, Olut et al., 2018, Dolhansky and Canton-Ferrer, 2017]. The IS was proposed for use with the CIFAR-10 data set, and is commonly used with this data set and the Imagenet data set. Its use for evaluation using other data sets has been noted to be misleading and is not advised [Barratt and Sharma, 2018]. The IS has also been noted to be sensitive to weights and batch sizes, and that small-size classes of data can result in misleading scores [Barratt and Sharma, 2018]. The IS's reduction of high quality samples to ones that are diverse and meaningful has also been questioned [Zhou et al., 2017].

The IS measures quality and diversity of generated images, but does not compare the generated images with actual images from the domain so as to assess their similarity. Heusel *et al.*'s Fréchet Inception Distance (FID), a proposed improvement to the IS, instead measures the distance between the embedding of p_{data} and p_g [Wu et al., 2017c, Lucic et al., 2017, Heusel et al., 2017]. The activations of the coding layer of the pre-trained 'Inception' net for the generated images is taken as a Gaussian distribution, which is then compared with that of actual images from the domain. The mean and covariance of these two distributions are then used to calculate the FID. A lower FID indicates greater similarity between these two distributions and thus better generated images. The Fréchet Inception Distance is defined as:

$$d^2((\mathbf{m}, \mathbf{C}), (\mathbf{m}_w, \mathbf{C}_w)) = \|\mathbf{m} - \mathbf{m}_w\|_2^2 + \mathrm{Tr}(\mathbf{C} + \mathbf{C}_w - 2(\mathbf{C}\mathbf{C}_w)^{1/2}) \tag{2.4}$$

where d^2 is the Fréchet distance between $(\mathbf{m}, \mathbf{C})$, the mean of the Gaussian of p_g, and $(\mathbf{m}_w, \mathbf{C}_w)$, the mean of the Gaussian of p_{data}.

FID is considered an improvement over the IS as it is less sensitive to noise, can detect mode collapse and is suitable for use on all data sets. A GAN that only produces one sample per class would achieve a perfect IS, but a poor FID, because the difference between its output images and real images from the domain will be high [Lucic et al., 2017]. Moreover, FID captures *both* precision and recall, while the IS captures only precision [Lucic et al., 2017].While FID is considered superior to the IS, it is interestingly only used in 3% of GANs [Hindupur, 2017]. However, the FID is not without drawbacks. It, much like the IS, fails to detect over-fitting [Lucic et al., 2017]. Moreover, a so-called 'memory GAN', which merely remembers and then reproduces all the training data, would achieve a perfect FID score and a perfect IS [Lucic et al., 2017].

Qualitative Evaluation Techniques

Perhaps due to the lack of a *single* quantitative metric, some GANs forego quantitative evaluation entirely [Lucic et al., 2017, Metz et al., 2016]. Instead, these GANs commonly use human judgment of sample quality, which we will term *visual quality* (VQ). VQ is the most popular GAN evaluation technique, used in 42% of GANs [Hindupur, 2017]. Some GANs also use Mechanical Turk qualitative surveys as evaluation [Hindupur, 2017]. Though such metrics are popular, they cannot be used as the (sole) evaluation technique for comparing GANs, due to their subjectivity [Lucic

et al., 2017, Im et al., 2016]. Other flaws of these qualitative evaluation measures are their cumbersome and expensive nature, and that due to their high variance, large sample sizes are required [Im et al., 2016]. Consequently, while some GANs use VQ alone, 96% of those that do use VQ use additional evaluation techniques [Hindupur, 2017]. The popular qualitative evaluation techniques used in GANs are Turing Test-based assessments and Likert scale-based surveys.

Challenges to Evaluation

Though GANs are an active area of research, the evaluation of GANs remains an important and unsolved problem [Lucic et al., 2017, Borji, 2019]. Evaluation, whether quantitative or qualitative, has been described as one of GANs' greatest challenges [Rosca et al., 2017]. The *de facto* quantitative evaluation metric for generative modelling techniques, log-likelihood measurements, cannot be used to evaluate GANs [Borji, 2019, Eghbal-zadeh and Widmer, 2017, Lucic et al., 2017]. There exists no universally-agreed-upon quantitative evaluation metric for GANs [Lucic et al., 2017]. The lack of a definitive, quantitative evaluation metric is significant, as it challenges fair and objective comparisons of different GANs, and makes it more difficult to identify poor models during training [Eghbal-zadeh and Widmer, 2017].

In addition to the lack of a standard quantitative evaluation metric, it is commonplace for GAN papers to report only the best scores achieved by the respective model, rather than averages [Lucic et al., 2017]. This can be misleading, as GANs have been found to be highly sensitive to the following:

- Data set
- Learning rate(s) of D and G
- Optimiser of D and G and its parameters
- Number of filters in D and G
- Random seed for network weights initialisation
- Network architecture

A given GAN will achieve dramatically different scores under a fixed evaluation metric when the above features are changed, even if in a seemingly minute manner [Lucic et al., 2017]. For example, DCGAN has been found to produce superior quality samples when the beta values of the ADAM optimiser for its networks are 0.5 and 0.999, rather than the default 0.9 and 0.999 [Radford et al., 2015, Chintala et al., 2016]. Though misleading, it is understandable why many GAN papers report only the best scores achieved. Many features of GANs have ranges of acceptable and plausible values, such as the learning rates, and to exhaustively explore the "combinatorial explosion in the number of choices [of GAN training] and their ordering" [Lucic et al., 2017] is impossible [Lucic et al., 2017]. This, coupled with the noted intensive GPU-processing requirements of GAN training, means GAN practitioners and researchers may only be able to explore a few model configurations, for both time and budgetary reasons [Lucic et al., 2017]. In order to compare GANs fairly, two guidelines have been proposed: a fixed computational budget, and the same network architecture across different models (where applicable) [Lucic et al., 2017]. An

increased computational budget has been observed to afford greater improvements to GAN performance than model or algorithmic changes [Lucic et al., 2017]. In fact, given enough computational resources, the majority of the state-of-the-art GANs exhibit equal performance under a given quantitative metric [Lucic et al., 2017]. Thus, by using a standardised budget, the effects on performance of the unique features of different GANs can be fairly observed [Lucic et al., 2017]. Similarly, as deeper and more complex architectures have been found to produce superior quality samples, using the same architecture for different GANs ensures that improvements to performance arise from algorithmic modifications and not architectural ones [Lucic et al., 2017].

Given the high variance in GANs performance depending on model configuration, GAN researchers have been encouraged to report *average* results [Lucic et al., 2017]. A cherry-picked result reported for a given GAN represents *only* one obtained under a particular seed and with a given computational budget, and *not* for that GAN *in general*. These limitations and caveats must be emphasized; to imply otherwise is misleading and also makes replication of results contingent on having not only an identical model, but also the same computational budget. Many GAN researchers have noted that their models are constrained by their limited budget, as wall-clock hours of a sufficiently-powerful GPU are very expensive [Jones, 2017, Lucic et al., 2017].

2.1.8 Applications and Data sets

While they have enjoyed use in audio and video tasks, GANs have been most widely used for image-related application areas, which may be grouped as follows:

- Categorical image generation
- Facial image generation
- Image super-resolution
- Image restoration
- Style transfer
- Image translation

Categorical image generation may be defined as the generation of images across a set of categories or classes, such as the digits 0 to 9. This, coupled with facial image generation tasks, are arguably the two most researched application areas of GANs. Table 2.1 shows the data sets most commonly used in these application areas.

Data Set	**Images Generated**
MNIST	Categorical
CIFAR-10	Categorical
Imagenet	Categorical
CelebA	Facial

Table 2.1: Benchmark Data Sets for Categorical and Facial Image Generation

As GANs are used for a wide variety of tasks, from underwater image restoration to artificial facial aging, over 90 data sets have been used in GAN implementations [Fabbri et al., 2018, Antipov et al., 2017, Hindupur, 2017]. Figure 2.4 shows GAN data set usage.

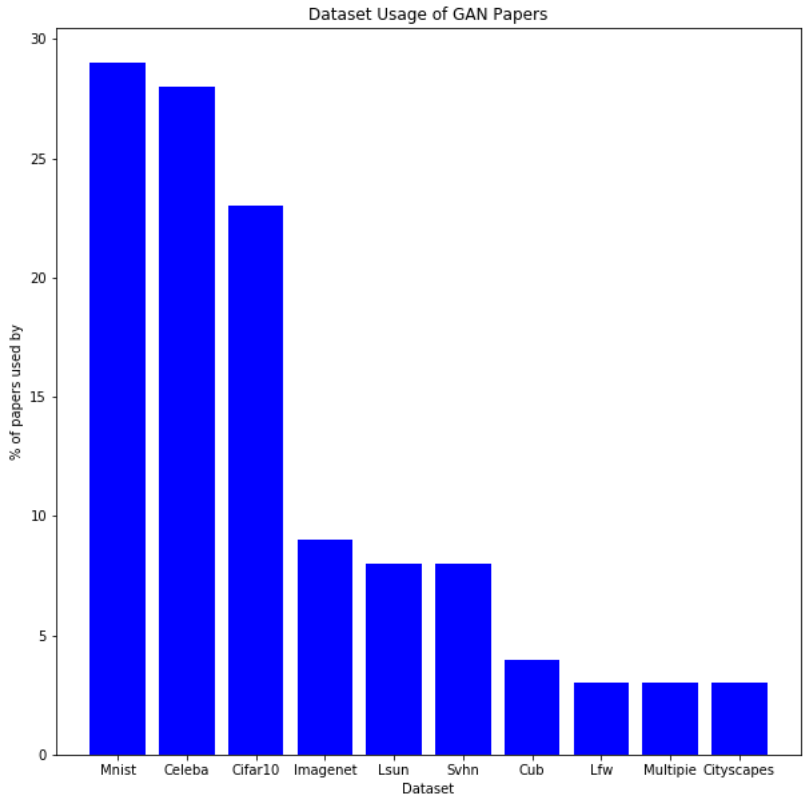


Figure 2.4: GAN Data set Usage

However, the benchmark data sets are arguably MNIST, CelebA and CIFAR-10, used in a combined 80% of GANs' implementations [Hindupur, 2017]. MNIST, a subset of the NIST database of handwritten digits, consists of 60000 training images and 10000 test images of size-normalized, centered 28x28 black-and-white handwritten digits [LeCun et al., 1998]. The MNIST dataset is frequently used by GANs as a baseline or sanity check dataset, as it contains small, simple images and can also be used for CGAN as it contains 10 classes (the digits 0-9). The CelebFaces Attributes Dataset (CelebA) is a collection of more than 200000 celebrity face colour images with 40 attribute annotations [Liu et al., 2015]. Though frequently used as the baseline dataset in GANs concerned with facial images, CelebA is also frequently used in GANs that are not facial image-specific, such as DCGAN [Hindupur, 2017, Radford et al., 2015]. CIFAR-10 is a labeled, 10-category subset of the 80 million tiny images dataset, and contains 50000 training images and 10000 test images, all 32x32 and equally split among the 10 classes [Krizhevsky, 2009].

2.1.9 Challenges

The main challenges of GANs are mode collapse (MC), training difficulties such as vanishing gradients and non-convergence, and the difficulty of generating high-resolution images. MC and training difficulties are the chief reasons for the modifications to the GAN objective function, as well as research into training strategy for GANs [Jaiswal et al., 2018]. MC, a well-studied problem of GANs, refers to the failure case of GANs where p_g is *far* less diverse than p_{data} [Rosca et al., 2017, Che et al., 2016]. Instead of representing the full diversity of p_{data} (*i.e.* all of its modes or classes), G concentrates on only a few modes of p_{data} and ignores the rest [Mariani et al., 2018, Zhu et al., 2017b, Gu et al., 2017]. Importantly, in the MC scenario, G does fool D, as its samples are realistic, but it creates few similar samples, which may not adequately cover p_{data} and thus is regarded as a failure case of GANs [Mariani et al., 2018, Chavdarova and Fleuret, 2017]. The cause of MC has been debated. Some attribute it to the nature of the JS divergence used in training GANs, even in the presence of 'perfect' training [Le et al., 2017]. Others believe the sometimes cyclical nature of the alternating gradient descent used in GANs, and the tendency of GANs to get stuck in local minima, to be the cause of MC [Kodali et al., 2017]. Though the exact cause of MC has not been definitively found, it remains a consistent problem of GANs [Gu et al., 2017, Zhu et al., 2017b].

A notable technique to combat MC is minibatch discrimination (MBD). MBD involves D examining many samples in combination, rather than one at a time [Salimans et al., 2016]. Though D retains its original task of single-sample discrimination, the use of multiple examples (the rest of the samples in the minibatch) gives D "side information" [Salimans et al., 2016]. Moreover, it allows D to detect samples in the minibatch that closely resemble others generated in the minibatch [Hoang et al., 2017]. MBD is viewed as strictly more powerful than using D for single images, and results in high-quality samples, but is computationally-expensive and requires extra parameters [Salimans et al., 2018, Hoang et al., 2017, Lin, 2017]. The ability of MBD to sufficiently punish a collapsed G has also been questioned, and it has not been used in any state-of-the-art GANs [Huang et al., 2016, Lucic et al., 2017, Hindupur, 2017].

GANs are well-documented to be difficult to train [Berthelot et al., 2017, Jaiswal et al., 2018, Shang et al., 2017, Barsoum et al., 2017, Che et al., 2016, Cao et al., 2017]. GAN training is noted to be exceedingly sensitive to hyper-parameter as well as network choice, with very few hyper-parameter choices leading to successful training [Che et al., 2016, Cao et al., 2017, Metz et al., 2016, Lin, 2017]. A common training problem is an imbalance between G and D [Durugkar et al., 2016]. D is typically far more powerful than G, especially at the beginning of training [Berthelot et al., 2017, Tran et al., 2018, Eghbal-zadeh and Widmer, 2017, Metz et al., 2016]. This has been referred to as the 'perfect discriminator problem' [Eghbal-zadeh and Widmer, 2017]. This imbalance causes vanishing gradients, a noted failure case of GANs [Jaiswal et al., 2018, Yi et al., 2017, Tran et al., 2018, Mao et al., 2016, Che et al., 2016, Eghbal-zadeh and Widmer, 2017]. Vanishing gradients occur when D perfectly, and too quickly, classifies G's samples as fake and D's loss becomes 0 [Tran et al., 2018, Che et al., 2016, Eghbal-zadeh and Widmer, 2017]. In this scenario, D becomes a perfect discriminator, saturating locally before G has the

chance to approximate p_{data} [Tran et al., 2018, Lin, 2017, Gulrajani et al., 2017, Yi et al., 2017]. Consequently, D cannot provide any useful feedback (a gradient) to G, and G stops learning completely [Lin, 2017, Eghbal-zadeh and Widmer, 2017, Tran et al., 2018, Miyato et al., 2018]. The vanishing gradient problem means G cannot possibly improve [Eghbal-zadeh and Widmer, 2017]. State-of-the-art GANs, such as WGAN and IWGAN, are thought to improve MC and training stability due to their constraining D's gradients [Kodali et al., 2017]. Generator-discriminator imbalance also leads to non-convergence, viewed by the creator of GANs as the most important GAN problem [Goodfellow, 2017]. Alternating gradient descent, used by GANs in its training of G and D, is noted as a cause of imbalance, as "an update made by one player can repeatedly undo the progress made by the other one" [Grnarova et al., 2017], and can cause GANs to get stuck in local Nash equilibria [Grnarova et al., 2017, Oliehoek et al., 2017]. Another flaw of GAN training, which may lead to non-convergence, is its lack of memory; D is "prone to forgetting past samples that the generator synthesizes" [Shrivastava et al., 2016, Kim et al., 2018].

Though far less-studied, the difficulty of using GANs for high-resolution images has been noted [Wang et al., 2017d, Bergmann et al., 2017, Wang et al., 2017c, Denton et al., 2016]. When generating high-resolution images, GANs are noted to introduce unwanted artifacts and fail to provide sufficient detail and texture [Wang et al., 2017d, Wang et al., 2017c]. This difficulty has been attributed more to GPU memory constraints than GAN architecture, though GAN architecture typically results in a fixed output size [Bergmann et al., 2017, Curto et al., 2017, Li et al., 2017b].

2.1.10 Benchmark GANs

Given GAN usage for different tasks across varied domains, the lack of a standard quantitative evaluation metric and the high computational requirements of GANs, it is difficult to provide a single benchmark GAN [Lucic et al., 2017]. The definitive study of GANs [Lucic et al., 2017] lists the benchmark GANs as: Wasserstein GANs (WGAN), improved WGAN with gradient penalty (IWGAN), Least Squares GANs (LSGAN), DRAGAN and Boundary Equilibrium GANs (BEGAN) [Lucic et al., 2017].

DCGAN remains the form of choice for benchmark GANs, including DRAGAN and WGAN [Lucic et al., 2017]. Though IWGAN is typically used with a more complex CIFAR ResNet instead, it achieves state-of-the-art performance when using a DCGAN structure, *i.e.* as a WGANGP. Indeed, CNN are the form for all state-of-the-art GANs with the exception of BEGAN, which uses VAE [Lucic et al., 2017]. The activation functions used in the state-of-the-art are `ReLU` and `LReLU`, used in all but BEGAN. IWGAN also makes use of the `softplus` and `Tanh` activation functions in its ResNet [Lucic et al., 2017]. Batch normalisation is used in all state-of-the-art GANs except for BEGAN, though layer normalisation is also explored in IWGAN [Lucic et al., 2017].

All state-of-the-art GANs use a value function modified from the original GAN [Lucic et al., 2017, Hindupur, 2017]. The most significant value function is the Wasserstein value function of WGAN. WGAN has been described as "the most seminal GAN-related work since the inception of the original GAN" [Juefei-Xu et al.,

2017]. This is not without cause; WGAN and IWGAN are used in 19% of all GANs [Gulrajani et al., 2017, Hindupur, 2017, Lucic et al., 2017]. Furthermore, IWGAN has been referred to as *the* state-of-the-art GAN "as it was shown to rival or outperform a number of previous methods" [Johnson and Zhang, 2018]. Interestingly, none of the state-of-the-art GANs use additional loss functions [Lucic et al., 2017]. With the exception of WGAN, which uses RMSProp as its optimisation method, all state-of-the-art GANs use ADAM [Lucic et al., 2017].

CelebA, CIFAR-10 and LSUN have each been the data set in at least one of the state-of-the-art GANs, as has MNIST and Imagenet [Lucic et al., 2017]. Interesting, there is little overlap in data sets used, with only CelebA and LSUN being used in evaluating more than one state-of-the-art GAN. The most common evaluation method used is visual quality, used in BEGAN, LSGAN and IWGAN, though the Inception Score and D loss have each been used in more than one state-of-the-art GAN [Lucic et al., 2017].

2.2 GANs for Computational Creativity

It could be argued that all GANs are creative, as they all produce samples that are novel. However, some GANs are specifically focused on tasks related to the arts. These are image translation, music generation and fine art generation.

2.2.1 Image Translation

Image translation involves the translation of input images into target output images while preserving the concept or content of the input images [Zhu et al., 2017a]. The range of desired transformations is wide. The `pix2pix` software package, which uses a conditional GAN, has been used successfully in the following image translation tasks: colourising black-and-white images, extracting the outline or edges of an image and transforming daytime scenes into night [Zhu et al., 2017a]. Image blending, the synthesizing of two input images into one, has also been investigated using GANs [Wu et al., 2017b]. Style transfer, which uses the stylistic features such as colour palette and textures of one image (often a painting) to transform another, is another common image translation task that has been approached using GANs [Johnson et al., 2016a, Zhu et al., 2017a]. Image super-resolution, the process of converting one or more low-resolution images into an equivalent high-resolution image, has also been performed using GANs [Johnson et al., 2016a, Ledig et al., 2016]. The GANs used in these image translation tasks are deep convolutional GANs, as convolutional neural networks are commonly used in image translation tasks [Johnson et al., 2016a].

2.2.2 Music Generation

Though GANs have predominantly been used for the generation of images, a number of GANs have also been used for music generation [Mogren, 2016, Chen et al., 2017, Dong et al., 2017]. These GANs have primarily used recurrent neural networks such as long short-term memory networks as their network structure, though deep convolutional neural networks have also been used [Mogren, 2016, Chen et al., 2017, Dong et al., 2017]. Human quality judgments, as well as various music-based characteristics such as pitch duration and tone span, have been used to evaluate these GANs [Mogren, 2016, Chen et al., 2017, Dong et al., 2017].

2.2.3 GANs for Fine Art Generation

The use of GANs for fine art generation is a new research area; only five fine art GANs have thus far been produced. These systems generate full art works and are thus automated painters. The fine art GANs are Elgammal *et al.*'s 'Creative Adversarial Network' (CAN) [Elgammal et al., 2017], Tan *et al.*'s 'ArtGAN' [Tan et al., 2017a, Tan et al., 2017b], Donahue and McAuley's 'semantically decomposed' GAN for art (SDAGAN) [Donahue and McAuley, 2017], Bonafilia and Jones' 'GANGogh' [Jones, 2017] and Wang *et al.*'s Chinese Painting GAN (CPGAN) [Wang et al., 2017b]. All use the WikiArt data set, a set of 81,449 style-labeled paintings from the 15th to the 20th century [Elgammal et al., 2017], with the exception of Wang *et al.*, who construct a data set by scraping images from Google and Baidu [Wang et al., 2017b]. The two state-of-the-art fine art GANs, by virtue of their citation count and influence on other fine art GANs, are CAN and ArtGAN. This work focuses on these two GANs. Table 2.2 shows the art GANs.

Art GAN	**Art Specificity**	**Evaluation**	**Data Set**
CAN	Ambiguous	Likert surveys	Wikiart
GANGogh	Genre	Classification accuracy VQ	Wikiart
ArtGAN	Style Genre Artist	Log-likelihood estimates VQ	Wikiart
CPGAN	Style Genre	D Loss VQ	Scraped (Google, Baidu)
SDAGAN	Ambiguous	Tool Evaluation	Subset of Wikiart

Table 2.2: The Art GANs

2.2.4 Approach

There are two contrasting approaches taken by CAN and the ArtGAN. ArtGAN's approach, the dominant approach among fine art GANs, is the creation of *specific* artworks through a conditional GAN. Most commonly, this approach seeks to create artworks of a specific art style or genre, such as Impressionist art or Chinese landscape paintings [Tan et al., 2017a, Wang et al., 2017b, Jones, 2017]. Generating artworks of a specific artist, such as Vincent van Gogh, has also been explored in these fine art GANs [Tan et al., 2017b, Tan et al., 2017a]. ArtGAN has the widest scope of these GANs, and aims to create genre-, artist- and style-based artwork [Tan et al., 2017a, Tan et al., 2017b].

However, in CAN, G is expressly tasked with *not* generating artworks of a specific style, genre or artist [Elgammal et al., 2017]. Rather, CAN seeks to create art of an *ambiguous* style [Elgammal et al., 2017]. CAN is motivated by the creative theories of arousal and the psychologist Colin Martindale's theory of new art creation [Elgammal et al., 2017], and is the only fine art GAN with an approach motivated by creative theory. The generator G of CAN's goal is to generate art that satisfies two properties simultaneously:

1. The generated art sufficiently looks like art.

2. The generated art has an ambiguous style.

The first property ensures that G does not create artwork that is too different from the artwork D knows about, the art in the WikiArt data set. This property is analogous to the classic objective of G; D must not be able to easily classify its creations as fake. This property aligns with the theory of arousal, which examines the level of excitement, the 'arousal potential', of a person in response to a stimulus. The properties of a stimulus that are the most important to aesthetic phenomena are the "collative properties": novelty, surprisingness, complexity, ambiguity, and puzzlingness [Elgammal et al., 2017]. It has been shown that stimuli with moderate arousal potential are preferred [Elgammal et al., 2017]. Stimuli with low arousal potential are viewed as boring, and people are averse to those with too high an arousal potential [Elgammal et al., 2017]. Thus, G must create art with not too high an arousal potential. This property is similar to the standard GAN value function; the discriminator D classifies G's outputs as either art or not art [Elgammal et al., 2017]. This classification, a signal to G, guides G towards creating images of art, by exploring "parts of the creative space that lay close to the distribution of art" [Elgammal et al., 2017].

The second property refers to D's ability to associate G's outputs with an art style. Prior to the adversarial process, D is trained on the distribution so that it learns to classify artworks by their style [Elgammal et al., 2017]. The second property aims to make D unable to identify the style of G's outputs. It is this last property that Elgammal *et al.* believe renders CAN creative [Elgammal et al., 2017]:

> We can think of a GANs that can be designed and trained to generate images of different art styles or different art genres by providing such labels with training. This might be able to generate art that looks like, for example, Renaissance, Impressionism, or Cubism. However that does not lead to anything creative either. No creative artist will create art today that tries to emulate the Baroque or Impressionist style, or any traditional style, unless doing so ironically.

Ambiguity of style is distinct from the inherent ambiguity Elgammal *et al.* note is often present in artificially-produced artworks, which "typically [do] not have clear figures or an interpretable subject matter" [Elgammal et al., 2017]. CAN employs an additional two losses to the standard GANs loss: a style classification loss and a style ambiguity loss. These losses work together to achieve the CAN objective.

The style classification loss is a simple classification loss that requires D to classify each sample into one of C classes (art styles) by minimizing the cross entropy between softmax posterior $D(c|x)$ and real art style labels [Elgammal et al., 2017, Donahue and McAuley, 2017]. Through this loss, D learns about art styles and how to distinguish them [Elgammal et al., 2017]. The style ambiguity loss is the key loss to CAN's approach. G is tasked with minimizing the cross entropy between a uniform distribution and $D(c|x)$ [Donahue and McAuley, 2017, Elgammal et al., 2017]. When this is achieved, D cannot reliably identify the art styles of samples, viewing each style as equally likely - and ambiguity is achieved [Elgammal et al., 2017].

2.2.5 Structure

The majority of fine art GANs use DCGAN as their structure. Both CAN and ArtGAN use a DCGAN as the base structure for their generators [Elgammal et al., 2017, Tan et al., 2017a]. CAN also uses a DCGAN for its discriminator. The structure of CAN is shown in Figure 2.5.

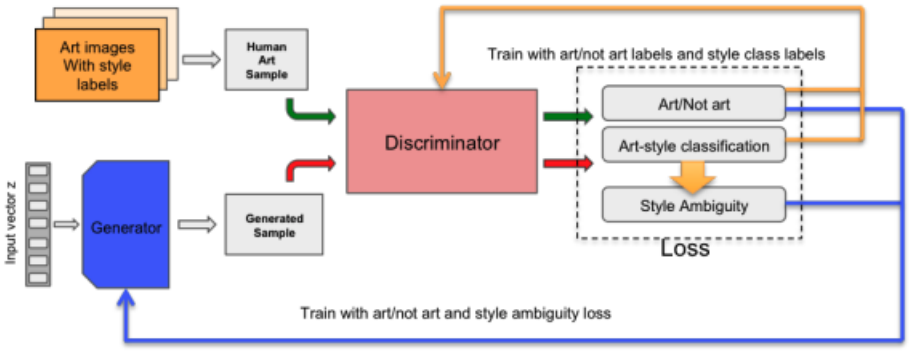


Figure 2.5: Structure of the CAN

In contrast, while earlier versions of ArtGAN used a DCGAN for the discriminator, the current ArtGAN uses a categorical auto-encoder as the basis for its discriminator [Tan et al., 2017a, Tan et al., 2017b]. This change was made due to the training improvements offered by AEs [Tan et al., 2017b]. ArtGAN uses the `LReLU` activation function and batch normalisation in both G and D. Tan *et al.* experiment with three variants of categorical AEs: an Energy-based GAN (EBGAN), a Denoising Feature Matching (DFM) GANs which feature a denoising autoencoder, and a standard AE GAN [Tan et al., 2017b]. The structure of ArtGAN can be found in the appendices. Like ArtGAN, CAN makes use of batch normalisation as well as `LReLU` in both G and D.

2.2.6 Training

Despite WGAN's and IWGAN's recognition as among the state-of-the-art in GANs, neither has been used in CAN nor in ArtGAN. Only the Chinese Painting GAN has investigated the use of Wasserstein-based GANs, and found such a GAN to produce higher-quality samples than DCGAN [Wang et al., 2017b]. However, the impact of the Wasserstein GANs on image sample quality has not been much explored.

ArtGAN uses a modified version of the conditional GAN's objective function, where D outputs a probability using softmax that a sample is one of K classes, plus an additional class that denotes a fake sample [Tan et al., 2017b]. Unlike CAN, which uses a style ambiguity loss and a style classification loss, ArtGAN does not feature additional loss functions. Both CAN and ArtGAN use the ADAM optimiser [Elgammal et al., 2017, Tan et al., 2017b].

2.2.7 Evaluation

There is not much overlap in the evaluation of art GANs. However, the majority of fine art GANs evaluations do use other GAN variants such as DCGAN for baseline comparisons [Elgammal et al., 2017, Tan et al., 2017b, Wang et al., 2017b].

Moreover, the majority of art GANs have VQ judgments as part of their evaluation [Jones, 2017, Tan et al., 2017b, Wang et al., 2017b]. Only ArtGAN's evaluation features a prominent GAN quantitative evaluation technique, the Inception Score [Lucic et al., 2017, Elgammal et al., 2017, Tan et al., 2017b].

Qualitative Evaluation

CAN is focused on the quality and characteristics of the generated images, as opposed to the robustness and stability of the model. CAN is evaluated through surveys of MTurk respondents and art history students, using five-point Likert scales to record their responses to the survey questions. These questions include the Turing test, an overall judgment of the images, and ones that ask the respondents to rate the following qualities of the images, seen below in Table 2.3.

Quality			
Inspiration	Novelty	Suprisingness	Ambiguity
Complexity	Intentionality	Visual structure	Communication

Table 2.3: Qualities examined in CAN's evaluation

CAN is the only fine art GAN which questions whether its creations *are indeed* art. Such questions are answered by Elgammal *et al.* through a Turing test question in their surveys. Tan *et al.*'s qualitative evaluation of their ArtGAN consists of examining its generated images for noise and artifacts, and ArtGAN's ability to mimic given artists, genres and styles. The author of GANGogh, as well as those of the CPGAN, also make visual quality judgments as part of their evaluations.

There is little overlap in quantitative evaluation of fine art GANs, and no overlap between CAN and ArtGAN. CAN is not evaluated by conventional GAN quantitative evaluation measures, such as log-likelihood estimation. CAN's evaluation includes t-tests to estimate the likelihood that the answers from the Likert scales used in survey questions come from the same distribution [Elgammal et al., 2017].The answers to these questions are used to evaluate whether or not CAN's images can be considered art, and their standing with respect to current and previous fine art. Despite the widely-held view that log-likelihood estimations is inappropriate for use with GANs and has the tendency to mislead, ArtGAN is first evaluated using this method [Tan et al., 2017b, Radford et al., 2015, Goodfellow, 2017]. ArtGAN is later evaluated using the Inception Score (IS) [Tan et al., 2017b]. Tan *et al.* note shortcomings of the IS, but use it due to the absence of another quantitative evaluative measurement for generative models [Tan et al., 2017b]. They concentrate on the 'objectness' part of the IS, the part that evaluates the ability of the generative model to generate meaningful objects [Tan et al., 2017b]. It is important to note that these IS calculations are *not* used to evaluate the (art) images of the ArtGAN, as is the norm with the IS. Rather, the GAN model, trained on a non-art data set, that achieves the highest 'objectness' score is *then* selected for training on Wikiart [Tan et al., 2017b]. The IS is thus used by Tan *et al.* only as a model selection criterion.

Limitations

Among the art GANs, CAN's evaluation is arguably the most in-depth. It examines various qualities of the generated samples as well as the creativity of the CAN itself. However, it does not make reference to qualities noted by other art GANs, such as noise, nor to the hallucinatory nature of its creations, a quality noted to be common to artificially-generated artworks [Tan et al., 2017a, Tan et al., 2017b, Elgammal et al., 2017]. Moreover, CAN's evaluation does not probe the emotional impact of its creations, and only examines the surprise evoked by its creations. Conversely, while Tan *et al.* note qualities such as noise and colour in the ArtGAN's creations, they do not explore the cognitive impact nor the emotional impact of the ArtGAN's creations. The remaining art GANs also do not explore the cognitive impact nor the emotional impact of their GANs' creations. Thus, there is a lack of thorough emotional and cognitive impact analysis among the art GANs, and the qualities of GAN-produced images, such as noise and structure, have not been combined in analyses.

2.2.8 Findings

CAN's evaluation showed it not only satisfied the definition of creativity, but was also viewed on par, if not better, with contemporary, human-authored art. CAN's evaluation also found that people are more likely to view the generated images as being human-made if first asked about the qualities of the images [Elgammal et al., 2017]. This substantiates the theory that CC researchers should be forthcoming about the artificial origin of the outputs of computational creative systems such as GANs. While Tan *et al.* state their created images are of high quality, this claim is not substantiated [Tan et al., 2017b]. Though ArtGAN is found to be able to recognise semantic similarities in genre-specific art works (such as landscapes), it is noted to perform worse at this task than at artist-specific and style-specific art works. Interestingly, mode collapse is encountered in ArtGAN experiments only in the Oxford-102 flower data set, and not in experiments using the Wikiart data set [Tan et al., 2017b].

2.2.9 Data Sets

The benchmark art data set is Wikiart. Each image in the Wikiart data set is one of 27 art styles. Each image is also assigned a genre label, such as 'Portrait'. The Wikiart data set is used by all art GANs, with the exception of the CPGAN. The relationship between the VQ of samples produced by a GAN when trained on a benchmark data set and the VQ of samples produced by a GAN trained on Wikiart has not yet been explored.

2.2.10 Challenges

There exist challenges specific to fine art GANs such as CAN and ArtGAN. The main technical challenge that has been identified are high-performance hardware requirements, which affect both the created images and the GAN's structure. The significant GPU power required to generate images of a size larger than 64×64 pixels has constrained the generated samples of fine art GANs [Jones, 2017]. This does impact sample quality, and may also impact outside observers' opinions of samples' 'artness'.

Significant data-related challenges are posed by art itself. Fine artworks may fea-

ture natural objects, though they often feature more than one object per artwork. Moreover, many fine art styles are not photo-realistic, such as cubism, and thus the objects they feature may have shapes that are distorted or unconventional [Tan et al., 2017b]. Consequentially, prominent fine art data sets, such as the Wikiart data set, are different from those commonly used in training GANs, such as Imagenet. Other data-based challenges that have been identified are the lack of data sets for some styles of fine art, such as Chinese paintings [Wang et al., 2017b]. Small class membership for some art styles and genres have also been noted as a challenge to style- or genre-specific generation [Jones, 2017]. This is particularly problematic for high variance and complex genres [Jones, 2017], in which fine art GANs have exhibited poorer performance, though it is believed that a larger sample size for such genres will alleviate this problem [Jones, 2017].

2.2.11 Extensions

Though CAN and ArtGAN have been found to be able to generate high-quality images, there nevertheless exist gaps between the state-of-the-art of GANs, and these state-of-the-art fine art GANs. These gaps include technical gaps and evaluative gaps. It is reasonable to assume that the state-of-the-art GAN techniques would improve the quality of samples produced by fine art GANs.

Technical Gaps

Despite the popularity of the Wasserstein distance as the divergence measure in GANs, it has not enjoyed a similar popularity in fine art GANs. Though Wasserstein distance via gradient penalty is the dominant objective function amongst the state-of-the-art GANs it has not been explored in fine art GANs. Given IWGAN's dominance in state-of-the-art GANs, and its notable benefits over the standard, binary cross entropy value function, the investigation of IWGAN in CAN and ArtGAN is much needed. It would also be interesting to compare the results of a WGAN-based fine art GAN to a IWGAN-based fine art GAN, to observe whether IWGAN's dominance over WGAN holds for art data sets such as Wikiart. Initial results suggest this would be true, as CPGAN found modified and non-modified WGAN to produce more colourful and sharper images than the standard DCGAN [Wang et al., 2017b]. Furthermore, while the DCGAN and WGAN exhibited mode collapse, the modified WGAN exhibited none whatsoever, and both Wasserstein GANs were also more stable during training than DCGAN [Wang et al., 2017b]. Thus the use of Wasserstein value functions in CAN and ArtGAN may lead to improved image quality and help avoid mode collapse in the latter. Though CAN *does* use additional loss functions, the use of popular supplementary GAN losses such as the Perceptual loss and L2 pixel-wise loss has not been explored in any fine art GAN. This too could lead to improvements in sample quality in both the ArtGAN and CAN.

Evaluative Gaps

Though ArtGAN uses the IS, it has not been evaluated using the IS improvement, the Fréchet Inception Distance. The FID is arguably better suited for use in ArtGAN, as it is appropriate for use in data sets other than Imagenet and CIFAR-10, unlike the IS. However, these methods require that a GAN is able to create *varied* and *specific* images, or be able to classify images into different categories. Thus it is not appropriate to evaluate CAN with these quantitative evaluation method, as

CAN *explicitly avoids* creating varied, specific and classifiable images.

2.3 Evaluation of Visual Art

Evaluation of visual arts, especially of modern visual art, is noted as *not* merely being a evaluation of the aesthetics of an artwork [Colton, 2008]. Moreover, even if this were the case, there exists "no collective notion of beauty within art intelligencia" [Colton, 2008]. It is argued that with the advent of photography, fine art ceased to be the medium of choice for representation, such as the use of portrait photography's replacing portrait paintings to represent a person [Colton, 2008]. This led to artists' changing from being "craftsmen to intellectuals who use artistic techniques as their medium of expression" [Colton, 2008]. This shift is argued to have resulted in the creativity of the artist's replacing the aesthetics of her work as the primary criterion of evaluation [Colton, 2008]. Though this evolution is most clearly exhibited in conceptual art, it is held that for all art styles, "artists are expected to create both at the conceptual and the craft level, and art-lovers are expected to appreciate both" [Colton, 2008]. In the evaluation of *artificial* visual art, we are interested in the answers to the following two questions:

1. Is the art *good*?
2. Is the art *creative* (*i.e.* is it art)?

The following approaches to evaluation of visual art are examined: those of the art GANs, those of computational creativity, and a formal, structural approach motivated by art scholars.

2.3.1 Computational Creativity Theory

Evaluative theory is an extremely active research area in CC [Gervás, 2009]. A high-level validation of artefacts of a computational creative system (CCS), such as selling an artificial painting, is argued to be insufficient; "day-to-day" evaluation methods are needed [Colton and Wiggins, 2012]. Computational creativity theory is primarily interested in the answer to the question "Is the art *creative*?".

The common, and perhaps intuitive, approach to evaluation of a CCS is to use a Turing test (TT) [Colton et al., , Gervás, 2009]. The TT tests whether a human outside observer can reliably mistake a computational system for a human. In the context of CC research, the TT has been frequently applied as a method for evaluating computational systems of visual art [Elgammal et al., 2017, Gervás, 2009, Colton et al.,]. The premise behind using the TT is that a CCS is successful if it produces artworks that are consistently thought of as being human-authored. In CC research, the TT is most commonly applied in the traditional sense; the outside observer is *not* told beforehand that the artwork is artificially-generated, and is then asked to classify the artwork as being produced by a human or by a computer [Colton et al., 2009]. However, some CC researchers, such as Simon Colton, believe that the makers of CCS should explicitly state to the observer that the outputs *are indeed* artificial, before asking whether she would have thought of the artwork as a human-produced artefact [Colton et al., 2009]. This, it is argued, will decrease the negative bias towards artificially-generated creative artefacts over time [Colton et al., 2009].

Though the TT has been extensively used as a method of evaluation of creative

systems, and continues to be used today [Elgammal et al., 2017], many believe the use of the TT is "setting computers up for a fall" [Gervás, 2009]. There exists a strong negative bias towards artificially-generated creative artefacts, such as artificial artworks, and thus these artefacts will be judged less kindly than their human-produced counterparts [Gervás, 2009, Colton, 2008]. This is arguably why the majority of TTs used in CC research are indeed traditional, or 'blind'; to inform the observer beforehand that the artworks are artificial is likely to prime the observer into evaluating them harshly.

Colton argues that the strong negative bias toward artificially-generated art, and belief that computers cannot ever be creative leads to a 'vicious cycle' [Colton, 2008]. The negative bias, that CCS are not creative, leads to a poor evaluation of artificially-generated artworks. According to CC theorists, that which produces poor works *cannot* be viewed as creative [Colton, 2008]. It is also argued that the use of the TT for evaluation of CCS is inappropriate as it implies the goal of CC research is to attain "human-level creativity" [Colton and Wiggins, 2012], as opposed to exploring creativity in non-human ways.

2.3.2 Art Theory

A prominent art theory-motivated approach to art evaluation is that of Hagtvedt *et al.*, who seek to build a structural equation model for the evaluation of art [Hagtvedt et al., 2008]. According to Hagtvedt *et al.*:

> [A]rt... may be identified as works perceived as embodying human expression, where a perceived main feature of the work is the manner of its creation and/or execution rather than just a concept, idea, or message underlying it or conveyed by it, and where this manner is not primarily driven by any other contrived function or utility [Hagtvedt et al., 2008].

Central to Hagtvedt *et al.*'s model are the ideas of cognition and affect, and the interplay between the two [Hagtvedt et al., 2008]. Cognition refers to the perceived attributes of the art work that form the foundation of its appeal to the viewer [Hagtvedt et al., 2008]. This appeal may be intellectual or aesthetic, and includes aspects of the work such as its novelty, surprise and the arousal it sparks in the viewer [Hagtvedt et al., 2008]. Affect refers to the emotional response of the viewer to the art work in question, a quality that is well-established to occur in visual art [Hagtvedt et al., 2008]. These emotional responses are characterised by their arousal level and valence [Hagtvedt et al., 2008]. Valence describes whether the response is negative or positive, and typically correlates with the evaluation of the art [Hagtvedt et al., 2008]. Similarly, the weight of the evaluation is tied to the level of arousal induced by the art [Hagtvedt et al., 2008]. Hagtvedt *et al.* support the view that the well-documented interplay takes the form of cognitive appraisals being dependent on emotions [Hagtvedt et al., 2008].

Hagtvedt *et al.* conduct an empirical investigation into the factors involved in art evaluation and their relations to one another and the summary judgment of the artwork [Hagtvedt et al., 2008]. First, they identify the primary emotional responses and perceived attributes associated with art. This list is then refined in collaboration with art scholars and existing literature [Hagtvedt et al., 2008]. Following this, a small group of respondents are shown a sample of five artworks and asked

to classify the artworks as evoking either positive or negative emotional responses and whether the artworks elicit high or low arousal, both on a nine-point semantic differential (Likert) scale. This is followed by a larger-scale survey by Hagtvedt using 150 undergraduate students. Each respondent is shown one of the five artworks and indicates the extent to which each of the emotional responses and attributes collected by Hagtvedt *et al.* is evoked by that artwork using a nine-point Likert scale [Hagtvedt et al., 2008]. Additionally, the survey respondents indicate their overall evaluation of the artwork using another nine-point Likert scale, which ranges from strongly negative to strongly positive [Hagtvedt et al., 2008]. Exploratory factor analysis is then performed to observe the link between emotions elicited and respondents' impressions and judgments of the artworks, with the factor loadings of the of 15 emotions and 15 perceived attributes being calculated [Hagtvedt et al., 2008].

The 15 emotions are split into four emotion factors: NH, NL, PH, and PL, where N is negative emotion, P is positive emotion, H is high arousal and L is low arousal [Hagtvedt et al., 2008]. Similarly, the 15 perceived attributes are each assigned to one of the following four cognitive factors: curiosity appeal, aesthetic appeal, creativity and skill [Hagtvedt et al., 2008]. The interplay between the four emotion factors, the four cognitive factors, and the final judgment of the artwork form the structure of Hagtvedt *et al.*'s equation model, seen below in Figure 2.6:

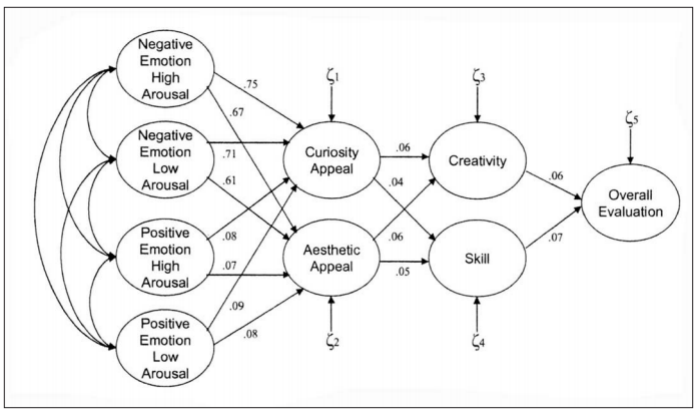


Figure 2.6: Hagtvedt *et al.*'s Equation Model

The cognitive factors identified by Hagtvedt *et al.* echo prior research, though are noted to exclude factors often associated with visual arts, such as complexity [Hagtvedt et al., 2008]. While this may seem a deficit, they note the current research focus on *universal* aspects of art, and that some art styles or forms may not benefit from higher complexity [Hagtvedt et al., 2008]. Similarly, Hagtvedt *et al.*'s model does not mention typicality, as some art styles may have little variation. Hagtvedt *et al.* stress that more research on emotion factors is necessary, as there is not yet consensus on whether overall affect of art, or separate emotional states evoked by art, should be prioritised [Hagtvedt et al., 2008].

Though Hagtvedt *et al.* provide an interesting model that includes both emotion and cognition, the model is not without shortcomings. Most importantly, Hagtvedt *et al.* note that a universal model of art evaluation may be impossible, due to the inherent variation and complexity of art [Hagtvedt et al., 2008]. They also note that

those surveyed in their work are not trained in art appreciation or evaluation, and that their results may differ if a trained group were used [Hagtvedt et al., 2008]. This strengthens the notion that a universal model is unattainable. Their model focuses exclusively on visual art, and does not examine the impacts and weightings of emotions and attributes across different visual art styles [Hagtvedt et al., 2008].

2.3.3 Art GANs

Elgammal *et al.*'s qualitative evaluation consists of four experiments on a specialised group of art history undergraduate students [Elgammal et al., 2017]. The first experiment is a Turing test in which those surveyed are tasked with deciding whether image samples appear human-authored or computer-generated. The next two experiments use five-point Likert scales to assess percieved attributes and elicited responses of the art, mirroring the approach of Hagtvedt *et al.*. In the second experiment, Elgammal *et al.* ask those surveyed to assess the shown samples on a small set of perceived attributes: appeal, novelty, surprise, ambiguity and complexity [Elgammal et al., 2017]. Thus, Elgammal *et al.*'s investigation into the *affect* of the artwork is shallower than Hagtvedt *et al.* However, Elgammal *et al.*'s third experiment allows for a thorough *cognitive* assessment of the artwork in a manner different from Hagtvedt *et al.*. In this third experiment, those surveyed are asked to indicate the extent to which they: observe the intention of the artwork, observe (the) structure in the artwork, feel communicated with by the artwork, and are inspired or elevated by the artwork [Elgammal et al., 2017]. In their final experiment, Elgammal *et al.* ask those surveyed to choose the more novel and more aesthetically-appealing of two artworks, one human-authored and one generated by the CAN [Elgammal et al., 2017].

Tan *et al.* do not perform any explicit or formal experiments to evaluate the GAN's samples qualitatively. Rather, they informally note the following attributes of their samples: noise, realism, and the use and relevance of colour [Tan et al., 2017b]. Additionally, Tan *et al.* comment on whether the samples are compelling. Though Tan *et al.*'s qualitative evaluation is indubitably the least comprehensive of the approaches listed in this work, their approach does offer unique advantages. Their examination of noise is relevant to VQ of samples generated by GANs, which often exhibit noise, and theirs is the only approach that notes not only the presence of colour, but whether its use is appropriate for the type of artwork generated [Tan et al., 2017b]. For example, a portrait-style artwork is unlikely to prominently feature the colour blue, and thus samples generated in this vein by ArtGAN, are atypical of art, though more creative. This consideration is important, whether one is judging samples by a human standard or under the view that computer-based creativity is separate.

2.4 Can Computers Be Creative?

It is believed in CC literature that it is natural to equate the terms 'human creativity' and 'creativity'. This belief is supported by the view held by much of the general public that creativity is a uniquely human quality that cannot be possessed by a computational system [Boden, 2009, Cardoso et al., 2009]. Many believe creativity is possessed only by humans, and "project cold, heart-less, simplistic (often just random) processing onto [computational systems]" [Colton and Wiggins, 2012]. The common argument levied against computers' being creative is that they per-

form only what they are instructed to, and that which can only follow rules cannot be creative [Boden, 2009, Cardoso et al., 2009]. Thus, any artefact produced by a computational system cannot render the system creative, but only its programmer [Boden, 2009]. Some CC researchers, such as Graeme Ritchie, believe that declaring the outputs of a creative computational system is more difficult than declaring the system itself creative [Cardoso et al., 2009]. Interestingly, Ritchie argues that CC research should defer the former until "we are substantially more capable in general automated reasoning and knowledge representation" [Gervás, 2009]. The view that computational systems cannot be creative is steadfastly held, irrespective of the performance of the computational system [Boden, 2009]. This indicates that people are biased against computational systems [Cardoso et al., 2009].

Many definitions of CC state that outputs of a software system can be said to be creative if observers would deem these outputs to be creative, if the latter outputs were made by a human [Cardoso et al., 2009, Colton and Wiggins, 2012, Colton et al., 2009]. However, it has been demonstrated that upon learning an artefact was artificially generated, and not generated by a human, observers of that artefact dramatically change their opinion of the value of the artefact. This change is almost universally negative, and has even resulted in the computational system's discontinuation [Boden, 2009]. Unsurprisingly, the bias against artificial creative artefacts has been challenged by both CC researchers and AI researchers. Seymour Papert's 'superhuman human fallacy' warns against rejecting the value of AI merely because it has not (yet) equalled the heights of human intelligence [Boden, 2009]. This would be akin to rejecting any painting or musical composition produced by anyone other than the world's finest artists. Not only would this be absurd, but it has not been the case with the outputs of AI research as a whole, such as self-driving cars.

2.5 Summary

Radford *et al.*'s DCGAN has become the standard deep convolutional GAN, and is frequently used as a baseline GAN against which new GANs are compared. The modified value function of the Wasserstein GANs have allowed them to be among the state-of-the-art, and achieve impressive quantitative and qualitative results noted to be superior to DCGAN. The GANHacks, a set of GAN training tips, were developed to help combat the noted training instability of GANs, though have not been rigorously investigated.

Though quantitative evaluation methods such as the Inception Score have enjoyed widespread use, evaluation of GANs remains one of their biggest challenges. The most widely-used evaluation metric, GAN sample quality judgments, are subjective, and due to high computational resource requirements and the difficulty of testing all possible parameter configurations, it is common for singular, cherry-picked results to be reported. This renders fair comparison of GANs, whether quantitative or qualitative, difficult. To fairly compare GANs, using the same structure for all GAN models, and the same computational budget, is encouraged.

The use of GANs for fine art generation is a new research area. Features of state-of-the-art GANs, such as Wasserstein GAN value functions, have not been extensively explored, and art generation presents both technical and philosophical challenges.

While the authors of some art GANs have explored the cognitive impact of their GAN's samples as part of their evaluation, in addition to the samples' creativity, a method to evaluate necessary considerations of art (cognitive impact, emotional impact and creativity) as well as common GAN sample defects, has not yet been devised.

Chapter 3

Methodology

3.1 GAN Implementations

Where possible, this work endeavoured to use official implementations of GANs and evaluation metrics, rather than implementing these from scratch. Official implementations offered a number of advantages: legitimacy, existing and active fora on their hosted platforms (most commonly GitHub), ease-of-deployment and documentation. To accurately replicate the experiments in DCGAN, this work follows the architectural guidelines as stated in Radford *et al.*'s work for the implementation of the DCGAN [Radford et al., 2015]:

- Strided convolutions in D
- Fractional-strided convolutions in G
- Batch normalisation in both G and D
- `ReLU` as the activation function in all layers of G, with the exception of the output layer, in which `Tanh` is used
- `LReLU` as the activation function in all layers of D

3.1.1 GAN Sets

In this work, three sets of GANs are used for experiments: a base set and two sets based on the GANHacks.

Base GANs

The Base GAN set consists of:

- DCGAN, CDCGAN
- WGAN, CWGAN
- WGANGP, CWGANGP
- IWGAN

Table 3.1 shows the base GANs, as well as the evaluation methods and data sets used in their papers.

GAN	Evaluation	Data Sets
DCGAN	Classification through D VQ	CIFAR-10 Imagenet-1k LSUN-Bedroom Celeb-A SVHN
WGAN	G and D loss Wasserstein distance	LSUN-Bedroom
IWGAN	Run time IS VQ D loss	CIFAR-10 LSUN-Bedroom Imagenet BillionWord

Table 3.1: The Base GANs

In addition to the DCGAN, the other GANs used in this set are the WGAN as per Arjovsky *et al.* [Arjovsky et al., 2017], a WGAN modified to use a gradient penalty as per Gulrajani *et al.*'s Improved Wasserstein GAN (WGANGP) [Gulrajani et al., 2017], and the Improved Wasserstein GAN (IWGAN) itself [Gulrajani et al., 2017]. With the exception of IWGAN, 64×64 versions of the base GANs were developed, to match the official *PyTorch* DCGAN and WGAN architectures. Conditional versions of DCGAN, WGAN and WGANGP (CDCGAN, CWGAN and CWGANGP) were developed to investigate the quantitative and qualitative differences (if any) between conditional GANs and their non-conditional counterparts. The IWGAN uses a CIFAR-10 ResNet and accommodates 32×32 images [Gulrajani et al., 2017]. IWGAN's network architecture is more complex than the others', as it has extra pooling layers, and thus has dramatically longer training times. Due to the longer training times of IWGAN, a conditional version thereof was not explored in this work. The architecture of the DCGAN's generator and discriminator, as well as the architectures of all GANs used in this work, can be seen in the appendices.

3.1.2 GANHacks Study

Two GANHack sets were investigated, one in which individual modifications (or GAN 'hacks') were applied in turn to a 32×32 DCGAN, and one where multiple modifications were added to a 32×32 DCGAN. It was expected that as the GANHacks combat training instability in GANs, that these modified DCGANs would lead to better quantitative and qualitative performance. Due to budgetary constraints, all DCGANs modified to use GANHacks were trained to generate 32×32 images.

The following GANHacks were investigated:

- Dropout (0.5) in multiple layers of G
- `LReLU` as the activation function of layers of G
- Adding Gaussian noise to each layer of G, decayed at $(1 + epoch)^{0.55}$
- Flipping the labels of samples given to D with probability $P = 0.5$
- Smoothing the labels of real samples given to D (from 1 to 0.9)

- RMSProp as the optimiser of D

However, not all of these GANHacks are reported in this work; the addition of Gaussian noise to G and the use of RMSProp as the optimiser of D were omitted. It was found that adding Gaussian noise to G, even when decayed over the course of training, collapsed G within the first epoch of training, with samples generated by G never fooling D and D having perfect accuracy. A GANHack that benefited the generator, such as flipped labels in the discriminator, had to be used in conjunction to prevent the generator from collapsing. The results of training with RMSProp added are not reported in this work as the addition had no effect on DCGAN training. Table 3.2 shows each GANHack investigated and the network that benefits from them.

GANHack	**Network Benefited**
Gaussian Noise in G	Discriminator
Dropout (0.5) in G	Discriminator
`LReLU` in G	Generator
Flipped labels in D	Generator
Smoothed labels in D	Generator

Table 3.2: *GANHacks* Network Benefits

3.1.3 GANHack Combinations

The following combinations of GANHacks were tested:

- Flipped labels in D + `LReLU` in G + Gaussian noise in G + Dropout in G (FLND)
- Flipped labels in D + `LReLU` in G + Gaussian noise in G (FLN)
- Flipped labels in D + `LReLU` in G + Dropout in G (FLD)
- Flipped labels in D + `LReLU` in G (FL)

3.2 Evaluation Ethos

Due to a fixed, limited computational budget, we could not perform an exhaustive search over the possible hyper-parameters for each GAN. Similarly, due to budget constraints, the various GAN models were only trained once, and thus averages are *not* given. Rather, the suggested hyper-parameters for each of the base GANs as listed in their original papers and official implementations were used. An exhaustive search over the possible hyper-parameters for any GAN, trained on just one data set, is infeasible [Lucic et al., 2017]. Thus, in comparing GANs, one can at best compare the optimal, worst and average results over a given number of hyper-parameter settings, for one or more sets of training data [Lucic et al., 2017]. Since such comparisons are still imperfect due to the many hyper-parameter settings not investigated, this work uses only the suggested hyper-parameters for each of the base GANs, as listed in their original papers and official implementations. Though this is a limitation of this work, it is important to remember that budgetary constraints are a significant constraint on GAN experimentation, and reporting singular results

is not uncommon [Lucic et al., 2017]. The GANs are then compared based on training under different data sets, as outlined in the next chapter. Table 3.3 shows the hyper-parameters of the base GANs.

Hyper-parameter	**(C)DCGAN**	**(C)WGAN**	**(C)WGANGP**	**IWGAN**
Learning rate	2×10^{-4}	1×10^{-4}	2×10^{-4}	2×10^{-4}
Noise vector dimension	[100,1]	[100,1]	[100,1]	[128,1]
D clamp value	-	± 0.01	-	-
D to G training ratio	1:1	5:1	5:1	5:1
Number of filters in D and G	64	64	64	64
G and D Optimiser	ADAM	RMSProp	ADAM	ADAM
Training Epochs	25	25	25	25

Table 3.3: Base GANs' Hyper-parameters

3.3 HEART - The Holistic Evaluation of Art

In this section, a qualitative evaluation method, the Holistic Evaluation of ART (HEART) is proposed. HEART combines the rigour and theoretical grounding of computational creativity and art theory into a single, structured evaluation technique. HEART not only assesses the relevant and important qualities of art, but accommodates the special considerations of GANs.

Through HEART's synthesis of the approaches of computational creativity theory, art theory and the art GANs, it is possible to answer the following question: Is the artwork of GANs *good*?

3.3.1 Is the artwork good?

This question is a summarised and simplified version of two sections of questions, the first mostly inspired by the research of Hagtvedt *et al.* and of Elgammal *et al.*, and the second by the special considerations of GANs.

Art Theory Factors

It is hoped that the first section of HEART, synthesized from previous approaches to evaluation of visual art, captures that which is necessary and sufficient to evaluate individual visual art works both cognitively and emotionally.

The first part of this section gauges the emotional impact of the artwork using the 15 emotion factors curated by Hagtvedt *et al.* Table 3.4 shows these factors arranged by valence and strength [Hagtvedt et al., 2008]:

Negative Emotion High Arousal	**Negative Emotion Low Arousal**	**Positive Emotion High Arousal**	**Positive Emotion Low Arousal**
Unease	Sadness	Excitement	Happiness
Anxiety	Despair	Enthusiasm	Joy
Uncertainty	Gloom	Thrill	Gladness
Disquiet	Loneliness		Serenity

Table 3.4: Hagtvedt *et al.*'s 15 Emotion Factors

For each of these emotion factors, the respondent is asked the extent to which she feels they are evoked on the following five-point Likert scale: [Strongly disagree, disagree, neither agree nor disagree, agree, strongly agree]. Through this, HEART captures both the range and intensity of the emotions evoked by the artwork, and represents the evaluation of the emotional impact of the artwork. The second part of the section represents the cognitive appraisal of the artwork, predominantly through Hagtvedt *et al.*'s cognitive factors [Hagtvedt et al., 2008]. For each of these attributes, using the same five-point Likert scale used in the first part of the section, the respondent is asked the extent to which she thinks it features in the artwork. Table 3.5 shows the cognitive factors used in HEART.

Cognitive Factors			
Interesting	Arouses curiosity	Fascinating	Intellectually stimulating
Aesthetically appealing	Appealing to the senses	Original	Distinct
Creative	Inventive	Of excellent workmanship	Well-crafted
Skillfully-made	Compelling (added)	Shows intent (added)	Good (excluded)
Favourable (excluded)	Positive (excluded)	Beautiful (excluded)	Attractive (excluded)

Table 3.5: Cognitive Factors used in HEART

All of Hagtvedt *et al.*'s perceived attributes and descriptors, or cognitive factors, are included in HEART, with the exception of the following: beautiful, attractive, good, favourable, positive. The descriptors 'beautiful' and 'attractive' were excluded from HEART as it was felt that artworks that appear frightening, unnerving or gloomy, *i.e.* ones intended *not* to be beautiful, may be penalised. However, the descriptor 'aesthetically appealing' was kept as it was more inclusive. Similarly, the descriptors 'favourable' and 'positive' were excluded, as they could be prejudiced against artworks with controversial, upsetting or negative subject matter. The attribute 'good' was excluded, as it was felt to be too vague and its inclusion not beneficial. The descriptors and attributes 'compelling' and 'showing intention', though not taken from Hagtvedt *et al.* but rather Tan *et al.* [Tan et al., 2017b], were included to supplement the cognitive appraisal. Tan *et al.*'s question of whether the artwork is compelling is added, as 'compelling' is a quality that could be applied to any style of artwork, and is a quality not necessarily assured if the particular artwork is considered good. For example, a landscape may look realistic and thus

positively evaluated, but this does not mean it is also compelling or memorable. Finally, the respondents are asked the extent to which they feel the artwork shows intent. For example, a portrait indicates the intent of the artist to render the face of a particular individual. However it must be noted that this may not favour abstract art, as it is easier to view the intent of a portrait than an abstract piece.

GAN Art Factors

The second section, derived partly from Tan *et al.*, has additional characteristics explored of the images applied to collages of GAN-produced artworks. This section evaluates not only the GANs as artists, but allows for qualitative comparisons *between* art GANs themselves.

Tan *et al.*'s qualitative evaluation of the ArtGAN's art specifically mentions the lack of noise in the GANs generated samples [Tan et al., 2017b, Tan et al., 2017a], such as blurring or speckled particles dotted throughout the sample. As such noise is *not* present in human-generated artworks, the presence and degree of noise in GAN-produced artworks is discouraged and negatively impacts the evaluation of a sample under HEART, especially if passing the Turing test (TT) is prioritised. This is the basis for the first characteristic explored in this section.

Tan *et al.* explicitly mention the prevalence of colour in the evaluation of their artworks. As GANs seek to emulate the data they are shown, we should expect their samples to mirror the human-generated artworks they are trained on. For example, a portrait-style sample would be expected to feature predominantly skin tone colours. Were such a sample to feature predominantly atypical colours such as green, purple or turquoise, the GAN would have arguably failed to have learnt how to render a portrait, even if the sample closely resembled a face. If the sample is abstract, or does not feature an object with a structure with a set colour scheme, then HEART examines the presence and variety of colour in the sample. Similar to arousal potential, too many colours, or too-intense colours, negatively impact the evaluation of the sample.

The samples of a GAN are meant to be representative of the training data. If trained on Wikiart, a diverse set of artworks of 27 art styles and multiple genres, it is reasonable to expect the trained GAN to be able to produce samples reminiscent of many styles. A GAN that produces homogenous art, such as samples that are exclusively portrait-like, has not only fallen victim to mode collapse, but is arguably a worse artist for it. An artist capable of producing artworks of various and distinct genres or styles is arguably more skilled and more creative than one only able to produce artworks of a single style or genre. Through this characteristic, HEART evaluates not only the samples of the GAN but the creativity and success of the GAN *itself*.

Artificially-generated art has been criticised for its tendency to appear hallucinatory [Elgammal et al., 2017]. This is viewed negatively, as the ambiguous and formless nature of these creations are viewed as having too high an arousal potential [El-

gammal et al., 2017]. This extreme arousal, coupled with hallucinatory qualities, is noted to be a clear mark that an artwork is *not* human-generated [Elgammal et al., 2017]. Thus, especially if a successful TT is prioritised, hallucinatory artworks will lead the corresponding GAN to be negatively evaluated.

It is also important to allow for direct comparisons of samples from different GANs, in which overall judgments are given. Such overall judgments would be informed by reference to the above characteristics.

From the above analysis, the following characteristics are explored in GAN images in HEART:

1. Blurriness and noise
2. Appropriate/varied colour
3. Diversity
4. Hallucinatory
5. Overall quality

3.3.2 Advantages

As a synthesis of existing approaches to visual art evaluation, HEART has a strong theoretical grounding and support. Both experts and non-experts (whether in art or in generative modelling) can use HEART to evaluate visual art, as it requires no domain knowledge. This is advantageous, as the interaction with visual art is considered a part of the human condition, and not merely those of an elite, learnt few [Hagtvedt et al., 2008]. Due to its depth, such as in the array of emotions included in the affect appraisal, it allows for those skilled in art or generative modelling to express and document their thoughts of an artwork in a detailed manner. Both high-level judgments and detailed feedback can be obtained from HEART. HEART also requires no computational resources nor experts to perform it, and is an inexpensive and quick method of visual art evaluation. Encompassing both the affect and cognitive impacts of an artwork, HEART provides an avenue for comprehensive criticism.

HEART can be used to evaluate GAN-generated visual art *without* requiring modifications, owing to its GAN-motivated extra considerations of hallucinatory properties, diversity and noise, and its investigation of mode collapse. HEART is thus a simple, comprehensive and versatile tool for visual art evaluation, both human- and computer-generated.

3.3.3 Limitations

The biggest limitation of HEART is due to the nature of visual art itself. Hagtvedt *et al.* note the impossibility of a single, perfect tool for visual art evaluation:

> Indeed, considering the complexity and variety of visual art, it may not even be feasible to capture the perception of visual art in its entirety with a single model. [Hagtvedt et al., 2008]

Thus, even the synthesized approach of HEART is not a universal, complete approach to visual art evaluation. HEART may not be appropriate for forms of art other than visual art, and research is needed on the differences, if any, in evaluation among different art forms [Hagtvedt et al., 2008]. Despite the acknowledged focus on untrained users, HEART may be insufficiently expressive for use with highly-trained users [Hagtvedt et al., 2008]. Hagtvedt *et al.* also note that any tool must necessarily compromise between comprehensiveness and parsimony, and that further research into emotions and their impact on cognition is needed [Hagtvedt et al., 2008].

Because it is designed with the problems of GANs (such as noise) and of computer-generated art (such as hallucinatory qualities) in mind, it may be that HEART is *too* specialised for use on human-generated art, and that a simplified version of HEART would be better suited for the task of human-generated art evaluation. Indeed, 'hallucinatory' is a quality noted to appear in computer-generated *visual* art, so HEART may be inappropriate for use with other forms of computer-generated art, such as music. Similarly, noise, while noted to appear in GAN-generated images, is not known to appear in other visual art, whether computer- or human-generated. Similarly, human artists are not actively or consciously considered less-creative or less-skilled for creating artworks of a single style or genre, unlike GANs. This additional requirement of art produced by GANs, that it be diverse, is thus not appropriate for use in human-generated art evaluation.

3.3.4 Is the artwork creative?

Through HEART, it is also possible to answer the question: is the GANs artwork creative? HEART includes the Turing test (TT), used by Elgammal *et al.*'s Creative Adversarial Network. The TT has already been used to evaluate GANs for fine art, so its adoption by HEART is neither radical nor untested. Though domain knowledge of art would no doubt bestow upon the viewer a more informed answer to the TT, the TT can be used by those not trained in art and does not require any specialised training. This is advantageous, as the art evaluation technique can be used by anyone. The TT is also not sensitive to the choice of art style, and is intuitive.

3.4 Summary

Three sets of GANs were used in experiments: the base GANs, DCGANs with individual GANHacks added, and DCGANs with multiple GANHacks added. Official implementations of GANs and quantitative evaluation metrics were used wherever possible. GANs used hyper-parameters as suggested in their papers. Due to budgetary constraints, experiments were only run once. The HEART tool for qualitative evaluation of art including GAN-generated art, was proposed and discussed.

Chapter 4

Experimental Design and Implementation

This chapter outlines how the GANs and GANHacks study were implemented and evaluated. The results are presented in the next chapter.

4.1 Frameworks and Libraries

As deep convolutional GANs are a deep learning technique, machine learning libraries that support deep learning were used in the development of experiments. A number of such libraries were considered, such as *PyTorch* [Paszke et al., 2019] and *TensorFlow* [Abadi et al., 2015].

PyTorch, the second-most popular deep learning framework for GAN development, was ultimately used as the primary machine learning library for the majority of experiments [Hindupur, 2017]. Not only is *PyTorch* user-friendly and high-level, but its extensive documentation and active community made it a more attractive option than the popular but low-level *TensorFlow*, the most popular deep learning framework for GAN development [Hindupur, 2017]. The most compelling factor in the decision to use *PyTorch* was that existing, official implementations of prominent GANs could be used for experiments and extended for further experiments. The *PyTorch* GitHub features code for a 64×64 DCGAN which was written in part by Soumith Chintala, one of the authors of the original DCGAN paper [Radford et al., 2015]. Additionally, the official code accompanying the WGAN paper is written in *PyTorch* and hosted on the author's GitHub page[Arjovsky et al., 2017].

Other machine learning libraries were used in this work. For experiments requiring Fréchet Inception Distance (FID) calculations, the official code accompanying the paper in which FID is introduced was used. This code is written in *TensorFlow*, and is the only instance of *TensorFlow* used in this work. This is in contrast to experiments requiring Inception Score calculations, which used a prominent *PyTorch* port of the original *TensorFlow* code. *Scikit-learn* was used as the machine learning library for the classification experiments, as *PyTorch* does not natively support the linear models needed for these experiments. *Hyperopt-sklearn* [Komer et al., 2019], an automatic hyperparameter optimisation tool for *scikit-learn*, was also used in these experiments.

4.2 Environment

As GANs are exceptionally resource-intensive and typically run on a GPU, a high-performance environment was needed. The smallest GAN model used required a minimum of 7GB of VRAM, and could thus not be run on most desktop PCs. All

GANs were instead trained on the cloud-based deep learning platform *FloydHub*[1], using a Nvidia Tesla V100-SXM2-16GB GPU. *FloydHub* was also used to host all data sets used for training GANs.

4.3 Data Sets

In this work, the GANs were trained on four data sets: CIFAR-10, Imagenet-1k, MNIST and Wikiart-based data sets. The SVHN data set [Netzer et al., 2011] was used for quantitative evaluation of GANs via unsupervised classification and was not used to train GANs. As per Elgammal *et al.*, the Wikiart data set was augmented to form a new data set, 'Augmented Wikiart', by performing five crops for each image: top left, top right, center, bottom left and bottom right [Elgammal et al., 2017]. For each crop, the crop size was 90% of the original image's size. This increased the size of the data set five-fold, to 407.295 images. Thereafter, for each of these images, a mirrored image was produced, doubling the size of the data set to a final 814.590 images. Following the cropping process, each image was resized to 64×64, to ensure all inputs to the network would be exactly the same size. Table 4.1 shows a breakdown of the Augmented Wikiart data set by genre.

Art Genre	**Number of Images**
Abstract Painting	60.156
Cityscape	55.752
Genre Painting	133.260
Illustration	22.872
Landscape	163.416
Nude Painting	23.976
Portrait	172.258
Religious Painting	79.440
Sketch and Study	53.328
Still Life	33.768

Table 4.1: Breakdown of Augmented Wikiart by genre

The other Wikiart-based data sets used in this work were the following: Portrait, Cityscape and Landscape. Portrait and Landscape were chosen as they are the most populous genres, and Cityscape was selected to observe conditional GAN performance on smaller art data sets. All experiments were run on a maximum image size of 64×64 due to intensive computational requirements. Table 4.2 shows the data sets used in this work.

[1] `https://www.floydhub.com`

Dataset	Total Images	Image Dimensions
Wikiart	81.449	64×64
Augmented Wikiart	814.449	64×64
SVHN	98.389	32×32
CIFAR-10	60.000	32×32
MNIST	50.000	32×32
Imagenet-1K	14.197.122	64×64
Faces (Portraits)	172.258	64×64
Cityscapes	55.752	64×64
Landscapes	163.416	64×64

Table 4.2: Data Sets

4.4 Quantitative Evaluation

As mentioned previously, evaluation of GANs, whether qualitative or quantitative, remains a difficult and unsolved problem [Borji, 2019]. Quantitative evaluation has been the more extensively explored arm of evaluation, with more than 24 techniques proposed [Borji, 2019, Lucic et al., 2017]. This is due to the inherent subjectivity of qualitative evaluation, and the desire for *objective* comparisons of GANs [Borji, 2019, Lucic et al., 2017].

As outlined in Chapter 3, there are many quantitative evaluation techniques used in GANs. In this work, classification of unseen data by a trained discriminator, the Inception Score (IS) and the Fréchet Inception Distance (FID) were selected to compare the GAN variants investigated. The accuracy metric is the primary method used by Radford *et al.* to evaluate their DCGAN, while the IS has since become the most accepted quantitative evaluation metric for GANs. The FID, which has not gained as much prominence as the IS, was also used as it is alleged to be a better metric than the IS, and to compare it with the IS [Lucic et al., 2017, Heusel et al., 2017].

4.4.1 Classifying Unseen Data Using GAN Discriminators as Feature Extractors

The objective of this evaluation method is to examine how well a trained GAN can classify images from another, *unseen* data set. The GANs' classification power are compared in this experiment. Classification accuracy in this experiment is defined as:

$$classification\ accuracy = \frac{correct\ image\ classifications}{all\ image\ classifications} \tag{4.1}$$

To best evaluate the GANs in this experiment, the exact approach of Radford *et al.* was followed [Radford et al., 2015]. The features from the discriminator of a trained GAN are used to build a classifier to classify images from unseen data sets [Radford et al., 2015]. The classification power of this classifier is assessed on two prominent image datasets, CIFAR-10 and SVHN. To extract the features learnt by the trained discriminator, each convolutional layer of the discriminator is extracted and flattened into a 1-dimensional vector using 4×4 max-pooling. These flattened vectors are then concatenated and used as the input to a regularized SVM classifier.

In their paper, Radford *et al.* only state that an SVM classifier is used; no other hyper-parameters or details of the linear model are given. Initially, *scikit-learn*'s `LinearSVC` classifier was explored for use as the linear model. However, this failed, due to the large memory requirements of this classifier and the inability to run on a GPU with *scikit-learn*. Instead, the `SGDClassifier` was used, so as to copy the approach of Radford *et al.* Other aspects of their experiments are not stated, namely:

- The size of the convolutional layers in the trained GAN
- The proportion of Imagenet-1K trained upon
- Learning rates
- The proportions of CIFAR-10 and SVHN used to fit and test the SVM classifier
- Hyper-parameter optimisation of the SVM classifier (if any)

In an attempt to match or surpass Radford *et al.*'s results, each model was trained both without any hyper-parameter optimisation and with optimisation via *hyperopt-sklearn*. The optimised runs consisted of 25 trials, each with a timeout of 1200 seconds, and also used the `SGDClassifier`. These restrictions ensured the optimised model would be a faithful and fair replication of Radford *et al.*'s experiment.

Radford *et al.* first train their DCGAN on Imagenet-1K, a prominent image dataset with 14 million images across 1000 categories [Deng et al., 2009]. They use the classifier to classify images from the prominent CIFAR-10 and SVHN image datasets [Radford et al., 2015]. CIFAR-10 consists of 60.000 32×32 RGB images across 10 categories, with 50.000 images in the training set and the remaining 10.000 images in the test set [Krizhevsky, 2009]. Similar to the prominent handwritten digit dataset MNIST, the Street View House Numbers (SVHN) dataset consists of 32×32 RGB images across 10 categories, one for each digit [Netzer et al., 2011]. SVHN has 73.257 images in the training set and 26.032 images in the test set.

Radford *et al.* report 82.8% accuracy on CIFAR-10 and 77.52% accuracy on SVHN [Radford et al., 2015]. In this work, the entire training set of CIFAR-10 was used to fit the SVM classifier. Though SVHN has more training samples, an equal number of SVHN images was used to fit the classifier, due to memory limitations. This experiment also served as a robustness check of the base GANs.

4.4.2 Inception Score and Fréchet Inception Distance Calculations

In addition to Radford *et al.*'s classification experiment, this work also investigated the Inception Score and Fréchet Inception Distance of the base GANs, as well as of DCGANs with GANHacks added. These scores, developed after Radford *et al.*'s DCGAN paper, while quantitative, also act as a proxy for evaluation of the quality of a GAN's generated images. These scores are added in the evaluation of this work in an effort to comprehensively quantitatively evaluate the base GANs, as well as DCGANs with GANHacks added. This work calculated the IS (higher is better) and FID (lower is better) of each GAN, in a two-stage fashion. Each GAN, pre-trained on either Imagenet-1K or CIFAR-10, was used to produce 64.000 images to serve as the data set of GAN samples for the IS calculations. Inception scores for GANs

trained on MNIST were not calculated, as the use of IS on data sets other than CIFAR-10 and Imagenet is noted to produce misleading results and is not advised [Barratt and Sharma, 2018]. For the FID calculations, each GAN was pre-trained on either CIFAR-10, Imagenet-1K or MNIST. These data sets of GAN samples were compared with a subset of 64.000 images from the corresponding real data set for the FID calculations. FID calculations of the IWGAN on MNIST were not performed as the network's structure does not accommodate the single-channel (black-and-white) images of this data set as inputs.

4.5 HEART Pilot Study

A preliminary user study was conducted study was conducted to illustrate the use of HEART. The study is not intended to represent a comprehensive user study. The objectives of the this study were threefold:

1. Examine the emotional and cognitive impacts of human- and GAN-generated artworks.

2. Judge the visual quality of GANs' artworks overall, and through various characteristics of the artworks.

3. Judge the creativity of human- and GAN-generated artworks using the Turing test.

These objectives were carried out using an online survey, in which 20 university students (at both the undergraduate and post-graduate levels) answered the questions of three sections. Echoing Hagtvedt *et al.*, there was a deliberate focus on *untrained* users; no art scholars were selected as respondents. Each respondent was compensated ZAR 50 ($3,61 USD).

4.5.1 Section One

In the first section, the affect (emotional) and cognitive impacts of artwork were assessed, using the emotions and cognitive factor lists adapted from Hagtvedt *et al.* [Hagtvedt et al., 2008]. Two artworks, one human-generated and one GAN-generated, were chosen for this experiment: Norwegian Expressionist Edvard Munch's 1893 artwork *Der Schrei der Natur*, commonly known as *The Scream*, and a batch of 64 images generated by a conditional 64×64 WGANGP (CWGANGP) that was trained on the Augmented Wikiart data set. *The Scream* was selected as the human-generated artwork as it is known to be evocative, and it was hoped that it would be especially likely to induce emotions. The CWGANGP was chosen as the GAN-generated artwork, as it was felt to be the GAN that created the most diverse and high-quality images of all GANs trained on Wikiart. Rather than cherry-picking a single 64×64 image from a batch of samples, it was decided that a batch (hereafter referred to as a 'collage') of samples of the CWGANGP would be more likely to induce (some of) the wide range of emotions asked about in HEART. Moreover, the collage was similar in dimensions to the image of *The Scream*. For both *The Scream* and the CWGANGP collage, the respondents were asked to rate the extent to which the artwork induced each emotion on a scale from 'Strongly disagree' to 'Strongly agree'. The respondents were then asked to rate the extent to which they agreed

with statements corresponding to the perceived cognitive attributes, on the same scale as used in the emotion questions. Figure 4.1 shows *The Scream* and Figure 4.2 shows the CWGANGP collage.

Figure 4.1: *The Scream* - Edvard Munch

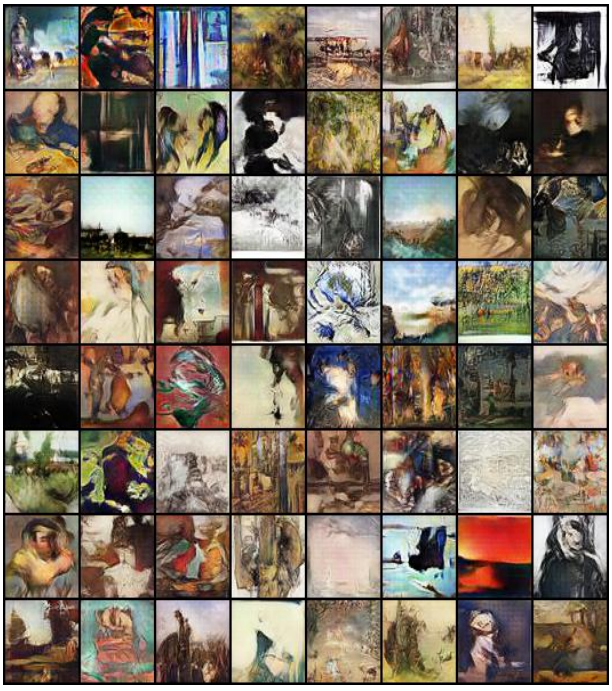


Figure 4.2: Collage of 64 samples of a CWGANGP

4.5.2 Section Two

The objective of the second section was to evaluate the visual quality (VQ) of two GANs. Collages of samples of two GANs trained on the Augmented Wikiart data

set were evaluated, through examination of the presence of characteristics commonly observed in GAN samples, such as blurriness. The two GANs evaluated were the same 64×64 CWGANGP as used in Section One, as well as a 64×64 DCGAN. The CWGANGP was chosen as it was felt to have the best-looking samples of all GANs trained on the Augmented Wikiart data set, and the DCGAN was chosen as a baseline for comparison. Figure 4.3 shows the DCGAN collage and Figure 4.2 the CWGANGP collage.

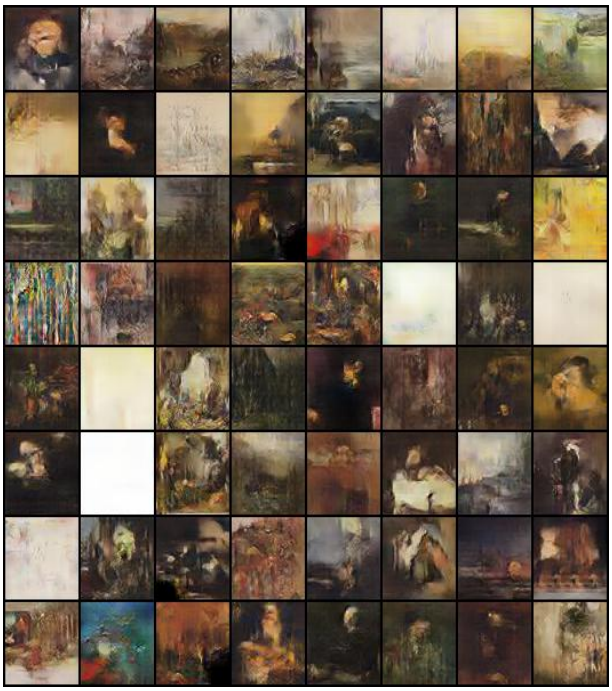


Figure 4.3: Collage of 64 samples of a DCGAN

Using a five point Likert scale, the respondents were asked the extent to which they agreed with the following statements:

1. There is blurriness/strange artifacts (e.g. speckled dots, weird lines) in the artworks of the collage.
2. The collage of artworks is diverse.
3. I can see structure in the artworks of the collage. (E.g. portraits have facial shapes)
4. The artworks in the collage are hallucinatory.

Finally, the respondents were asked to rate their overall judgment of the collages using a five point scale ranging from 'Extremely poor' to 'Extremely good'. Though Tan *et al.*'s make reference to the variety and presence of colour in their ArtGAN, the statement "There is varied/appropriate colour in the collage (E.g. portraits have skin tones)" was omitted following feedback during a trial run of the study. A respondent noted that some artworks, such as the portraits of Kazakh artist Vladimir Tretchikoff, do not feature colours typically associated with the subjects of their composition, but are nevertheless considered artworks. Figure 4.4 shows Tretchikoff's *Chinese Girl* (1952), a portrait with unusual colours.

Figure 4.4: Vladimir Tretchikoff's *Chinese Girl* (1952)

4.5.3 Section Three

In the final section, respondents were asked to perform the Turing Test on two images: a human-generated abstract painting, and an individual image from a batch of samples generated by the CWGANGP trained on Augmented Wikiart. The objective of this was to assess the creativity of the images. The respondents were asked to assign either the label 'Human-generated' or 'Computer-generated' to each image. To ensure a fair comparison, the human-generated abstract painting, Adolph Gottlieb's *Untitled* (1969), was downsized to the same resolution as the CWGANGP sample: 64×64. The particular single generated image of the CWGANGP was chosen as it was felt to resemble a modern abstract painting. Gottlieb's *Untitled* was chosen for its simple, abstract composition. Figure 4.5 shows a 64×64 version of *Untitled* (1969) and Figure 4.6 shows the sample from a 64×64 CWGANGP.

Figure 4.5: *Untitled* (1969) - Adolph Gottlieb (64×64)

Figure 4.6: Sample from a 64×64 CWGANGP

4.5.4 Limitations

It could be argued that the comparison of collages with non-collage artworks, such as *The Scream*, is an unfair one which may lead to biases. For example, an individual may have a distaste for collages, and thus be predisposed to prefer a non-collage artwork. The order of images presented to the respondents in section three was not randomised, which is another source of bias in this experiment. The experiment uses a small sample size of 20, which renders their responses less representative. Finally, only first moment statistics (average values) are calculated in this experiment, and the differences between artworks are not substantiated or refuted via statistical analysis by using a null hypothesis.

4.6 Summary

Three types of quantitative experiments were performed. Following Radford *et al.*, features from trained GAN discriminators were used to build a classifier to classify images from unseen data sets. The GANs were trained on Imagenet-1K and tested on CIFAR-10 and SVHN. The IS and FID scores of the GANs were calculated on CIFAR-10, MNIST and Imagenet-1K. A pilot study of HEART was performed with 20 undergraduate students. In this experiment, the emotional and cognitive impacts of a human-generated artwork and a set of samples outputs of a CWGANGP were examined. Two collages of a DCGAN and a CWGANGP were directly compared. Finally a Turing test was performed using a human-generated abstract artwork and a single image generated by a CWGANGP.

Chapter 5

Results

5.1 Quantitative Experiment Results

5.1.1 Classifying Unseen Data Using GAN Discriminators as Feature Extractors

Base GANs

Table 5.1 shows the classification accuracies of the base GANs on CIFAR-10, as well as that of Radford *et al.*

GAN	**CIFAR-10 Accuracy (%)**	**CIFAR-10 Optimised Accuracy (%)**
DCGAN	64.36	**73.64**
CDCGAN	64.8	71.52
WGAN	60.45	64.56
CWGAN	46.58	58.48
WGANGP	53.16	70.47
CWGANGP	42.88	67.87
IWGAN	34.75	71.59
Radford et al. [Radford et al., 2015]	**82.8**	-

Table 5.1: Base GAN CIFAR-10 Classification Accuracy

The highest accuracy achieved, apart from Radford *et al.*, was the optimised DCGAN (73.64%), though the optimised versions of CDCGAN and IWGAN achieved similar accuracy.

Table 5.2 shows the classification accuracies of the base GANs on SVHN, as well as that of Radford *et al.*

GAN	**SVHN Accuracy (%)**	**SVHN Optimised Accuracy (%)**
DCGAN	74.2	80.54
CDCGAN	65.10	71.11
WGAN	51.01	66.88
CWGAN	41.79	55.23
WGANGP	73.89	80.04
CWGANGP	48.2	**80.94**
IWGAN	23.37	65.22
Radford et al. [Radford et al., 2015]	**77.52**	-

Table 5.2: Base GAN SVHN Classification Accuracy

In contrast, while no GAN was able to achieve or surpass the accuracy of Radford *et al.* on CIFAR-10, their 77.52% accuracy on SVHN [Radford et al., 2015] was surpassed by the optimised versions of DCGAN, WGANGP and CWGANGP.

GANHacks

Table 5.3 shows the classification accuracies of 32×32 DCGANs, with and without GANHacks added, on CIFAR-10.

GANHack	**CIFAR-10 Accuracy (%)**	**CIFAR-10 Optimised Accuracy (%)**
No GANHack	60.76	**74.24**
Dropout (0.5) in G	55	66.35
`LReLu` in G	61.74	69.6
Flipped labels in D	65.36	61.93
Smoothed labels in D	59.66	54.59
FLND	46.06	61
FLN	44.29	57.96
FLD	48.25	48.38
FL	**65.56**	72.32

Table 5.3: GANHack CIFAR-10 Classification Accuracy

Interestingly, none of the GANHacks improved the classification performance of the optimised DCGAN without GANHacks on CIFAR-10, and in fact led to *worse* performance. However, some GANHacks did have higher classification accuracy than the DCGAN without any GANHacks when non-optimised.

Table 5.4 shows the classification accuracies of 32×32 DCGANs, with and without GANHacks added, on SVHN.

GANHack	**SVHN Accuracy (%)**	**SVHN Optimised Accuracy (%)**
No GANHack	59.78	**79.56**
Dropout (0.5) in G	50.46	52.08
`LReLu` in G	**71.98**	73.33
Flipped labels in D	68.59	78.68
Smoothed labels in D	65.76	77.69
FLND	34.47	56.39
FLN	40.82	71.6
FLD	47.01	71.3
FL	69.27	73.76

Table 5.4: GANHack SVHN Classification Accuracy

Similarly, none of the GANHacks improved the classification performance of the optimised DCGAN without GANHacks on SVHN. Again, some GANHacks did have higher classification accuracy than the DCGAN without any GANHacks when non-optimised.

5.1.2 Inception Scores and Fréchet Inception Distances

Base GANs

Table 5.5 shows the Inception Scores (higher is better) and Fréchet Inception Distances (lower is better) of the base GANs on the CIFAR-10 data set.

GAN	**Inception Score (IS)**	**Fréchet Inception Distance (FID)**
DCGAN	**2.00±0.02**	271.64
CDCGAN	1.57±0.03	**245.18**
WGAN	1.68±0.01	281.12
CWGAN	1.67±0.03	248.56
WGANGP	1.87±0.01	327.76
CWGANGP	1.65±0.03	247.79
IWGAN	1.66±0.01	273.36

Table 5.5: GAN Inception Scores and Fréchet Inception Distances - CIFAR-10

DCGAN achieved the highest IS, while CDCGAN achieved the lowest FID. Table 5.6 shows the scores of the base GANs on the Imagenet-1K data set.

GAN	Inception Score (IS)	Fréchet Inception Distance (FID)
DCGAN	**3.82±0.02**	146.94
CDCGAN	2.04±0.03	222.29
WGAN	2.22±0.02	279.05
CWGAN	1.79±0.03	265.61
WGANGP	3.42±0.04	150.01
CWGANGP	2.47±0.05	203.69
IWGAN	3.47±0.04	**103.7**

Table 5.6: GAN Inception Scores and Fréchet Inception Distances - Imagenet-1K

DCGAN achieved the highest Inception Score of the base GANs. However, the IWGAN achieved the lowest (best) FID.

Table 5.7 shows Fréchet Inception Distances of the base GANs on the MNIST data set.

GAN	Fréchet Inception Distance (FID)
DCGAN	96.77
CDCGAN	51.76
WGAN	59.65
CWGAN	**44.52**
WGANGP	122.45
CWGANGP	92.78

Table 5.7: GAN Fréchet Inception Distances - MNIST

CWGAN achieved the lowest (best) FID on MNIST.

GANHacks

Table 5.8 shows the Inception Scores and Fréchet Inception Distances of DCGANs with GANHacks added, as well as that of a DCGAN without any GANHacks, on the Imagenet-1K data set.

GANHack	Inception Score (IS)	Fréchet Inception Distance (FID)
No GANHack	**2.6±0.02**	192.65
Dropout (0.5) in G	$2.46\pm4.4 \times 10^{-16}$	349.67
`LReLu` in G	1.86±0	295.33
Flipped labels in D	2.46±0.02	**184.71**
Smoothed labels in D	2.46±0.02	187.01
FLND	1.88±0.01	244.71
FLN	1.74±0.01	252.79
FLD	1.49±0.03	196.56
FL	1.42±0.01	271.64

Table 5.8: GANHack IS and FID - Imagenet-1k

While the DCGAN without any GANHack achieved the highest (best) IS on Imagenet-1K, the GANHacks of flipped labels in D and smoothed labels in D achieved lower (better) FIDs than the DCGAN without any GANHacks.

Table 5.9 shows the Fréchet Inception Distances of DCGANs with GANHacks added, as well as a DCGAN with no GANHack, on the MNIST data set.

GANHack	**Fréchet Inception Distance (FID)**
No GANHack	96.77
Dropout (0.5) in G	196.2
LeakyReLu in G	189.07
Flipped labels in D	249.65
Smoothed labels in D	238.13
FLND	398.63
FLN	372.96
FLD	105.31
FL	**67.39**

Table 5.9: GANHack FID - MNIST

While the DCGAN without any GANHack achieved a lower FID than all but one of the GANHacks tested, the best FID score obtained by the FL-DCGAN was 44% better than that of the DCGAN without any GANHack.

5.2 HEART Pilot Study

5.2.1 Section One

In this section, respondents were asked the extent to which 15 emotion factors and 15 cognitive attributes were evoked by two artworks: Edvard Munch's *The Scream* and a collage of 64 images created by a conditional WGANGP trained on Augmented Wikiart. Respondents marked the extent to which these were evoked using a five-point Likert scale. The most popular responses to each factor and each attribute for both artworks are discussed, as well as means and standard deviations for each factor and each attribute.

Table 5.10 shows the most popular response to *The Scream* and the CWGANGP collage of the emotion factors.

Emotion Factor	***The Scream***	**CWGANGP Collage**
Unease	Agree (75%)	Agree (45%)
Anxiety	Agree (50%)	Agree (45%)
Uncertainty	Agree (50%)	Agree (60%)
Disquiet	Agree (40%)	Agree (46%)
Sadness	Agree (35%)	Disagree (55%)
Despair	Neutral (25%)	Disagree (40%)
Gloom	Agree (35%)	Disagree (40%)
Loneliness	Disagree (30%)	Disagree (50%)
Excitement	Strongly disagree (45%)	Disagree (45%)
Enthusiasm	Strongly disagree (50%)	Disagree (45%)
Thrill	Strongly disagree (40%)	Disagree (40%)
Happiness	Strongly disagree (60%)	Disagree (40%)
Joy	Strongly disagree (55%)	Disagree (45%)
Gladness	Strongly disagree (60%)	Disagree (55%)
Serenity	Strongly disagree (55%)	Disagree (55%)

Table 5.10: Most Popular Response (with proportion of total) for Emotion Factors

The majority of respondents stated that *The Scream* induced negative emotions, namely unease, anxiety and uncertainty, with 'Agree' being the most popular response to these emotion factors. In contrast, the more positive emotion factors, such as gladness, joy and serenity, were *not* felt by the respondents on viewing *The Scream*. This can be seen in 'Strongly disagree' being the most popular response for all positive emotions.

The respondents viewed the CWGANGP collage in a similar fashion; the negative emotion factors of unease, anxiety, uncertainty as well as disquiet were most commonly responded to with 'Agree'. The most popular responses to this artwork were more uniform than to *The Scream*, with all other emotion factors receiving the 'Disagree' response, including both positive and negative emotion factors.

Table 5.11 shows the means and standard deviations of responses to *The Scream* and the CWGANGP collage of the emotion factors.

Emotion Factor	***The Scream***	**CWGANGP Collage**
Unease	3.85±0.79±	3.15±1.11
Anxiety	3.5±0.97	3.05±0.97
Uncertainty	3.3±1.05	3.45±1.2
Disquiet	3.75±1.04	3.45±0.97
Sadness	2.5±1.07	2.25±0.83
Despair	2.85±1.28	2.15±0.85
Gloom	3.25±1.37	2.2±0.93
Loneliness	2.85±1.24	2.1±0.89
Excitement	1.95±1.07	2.25±0.89
Enthusiasm	1.65±0.73	2.15±0.91
Thrill	2±1.05	2.45±1.12
Happiness	1.5±0.67	2.1±0.77
Joy	1.6±0.73	2.15±0.73
Gladness	1.45±0.59	1.95±0.67
Serenity	1.55±0.67	1.95±0.67

Table 5.11: Means and standard deviations for Emotion Factors

The Scream achieved a higher mean score than the CWGANGP collage in all 8 negative emotions except 'uncertainty', and the CWGANGP collage achieved a higher mean score than *The Scream* on all positive emotions. HEART results were thus as expected given *The Scream*'s noted unsettling nature.

Table 5.12 shows the most popular response to *The Scream* and the CWGANGP collage through the cognitive attributes.

Attribute	***The Scream***	**CWGANGP Collage**
Interesting	Agree (45%)	Agree (40%)
Arouses curiosity	Agree (70%)	Agree (45%)
Fascinating	Agree (55%)	Neutral (35%)
Intellectually stimulating	Agree (40%)	Neutral (40%)
Aesthetically appealing	Neutral (35%)	Neutral (35%)
Appealing to the senses	Agree (50%)	Neutral (40%)
Original	Agree (55%)	Agree (45%)
Distinct	Strongly agree (45%)	Neutral (35%)
Creative	Agree (60%)	Neutral (45%)
Inventive	Agree (55%)	Neutral (45%)
Of excellent workmanship	Agree (65%)	Neutral (40%)
Well-crafted	Agree (80%)	Agree (35%)
Skillfully-made	Agree (75%)	Agree (35%)
Compelling	Agree (60%)	Neutral (45%)
Shows intent	Agree (40%)	Neutral (40%)

Table 5.12: Most Popular Response (with proportion of total) for Cognitive Factors

With regard to the cognitive factors, *The Scream* was largely well received. With the exception of the attribute 'aesthetically appealing', the most popular response to the remaining attributes were of agreement. A large majority of respondents judged *The Scream* as arousing curiosity, well-crafted and skillfully-made. Interestingly, only 40% of respondents believed *The Scream* shows intent, despite their assertion that it is well-crafted.

While *The Scream* received mainly agreements as was expected, the CWGANGP collage did well by scoring a similarly low proportion of 'Strong Disagree' (5.6%) to that famously evocative artwork (3.33%). It also had only about 1 in 3 of the responses given as neutral, and there was a spread of opinions in the answers, rather than consistent disagreement.

Table 5.13 shows the means and standard deviations of responses to *The Scream* and the CWGANGP collage for the cognitive attributes.

Attribute	***The Scream***	**CWGANGP Collage**
Interesting	3.7±1	3.8±0.98
Arouses Curiosity	3.75±0.83	4.05±0.97
Fascinating	3.65±0.85	3.35±1.06
Intellectually Stimulating	3.4±0.97	3.05±1.02
Aesthetically appealing	3.15±1.24	2.85±1.11
Appealing to the senses	3.4±1.11	3.15±1.01
Original	4±0.77	3.25±0.94
Distinct	4.2±0.98	2.9±1.04
Creative	4.15±0.73	3.3±0.95
Inventive	3.5±0.87	2.9±0.83
Of excellent workmanship	3.85±0.57	2.95±0.86
Well-crafted	4±0.45	3.15±0.96
Skillfully-made	3.95±0.5	3.15±0.96
Compelling	4±0.63	3±0.84
Shows intent	4.15±0.85	2.95±1.16

Table 5.13: Means and standard deviations for cognitive attributes

With the exception of 'interesting' and 'arouses curiosity', *The Scream* achieved a higher mean score for all of the cognitive attributes.

5.2.2 Section Two

In this section, the respondents were asked the extent to which four qualities (blurriness and noise, structure, diversity and hallucinatory) were evoked in collages of samples of two GANs: the same CWGANGP as in Section One, and a DCGAN. Finally, the respondents were asked for their overall judgment of the quality of each collage. Figure 5.1 shows the ratings of blurriness and noise assigned to the collages.

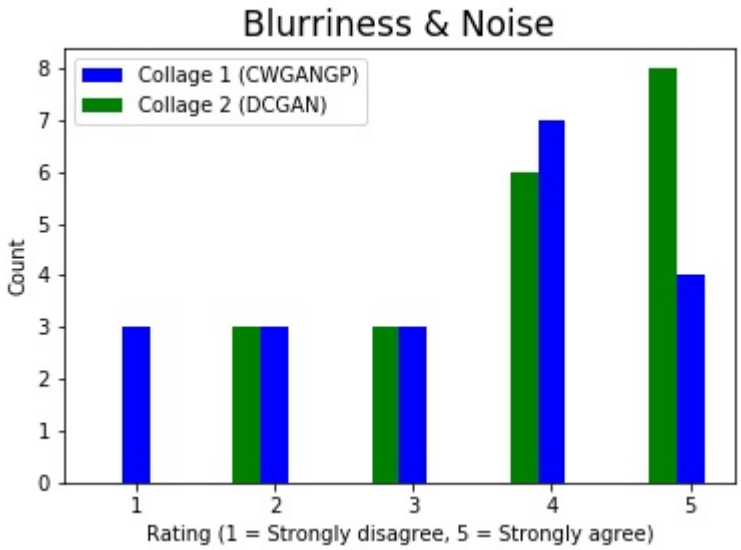


Figure 5.1: Blurriness and noise in the collages

While both GAN collages were judged as having blur and other noisy artifacts (70% for DCGAN and 55% for CWGANGP), the DCGAN collage was judged as more noisy and blurry, with 40% of respondents responding 'Strongly agree'. Figure 5.2 shows the ratings of diversity assigned to the collages.

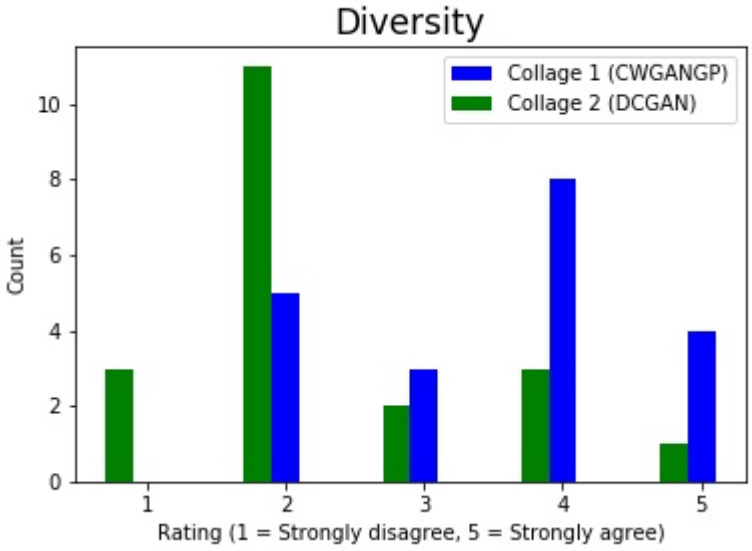


Figure 5.2: Diversity of the collages

In addition, most respondents (60%) agreed that the CWGANGP collage was diverse, while most (65%) disagreed that the DCGAN collage was diverse. Figure 5.3 shows the ratings of structure assigned to the collages.

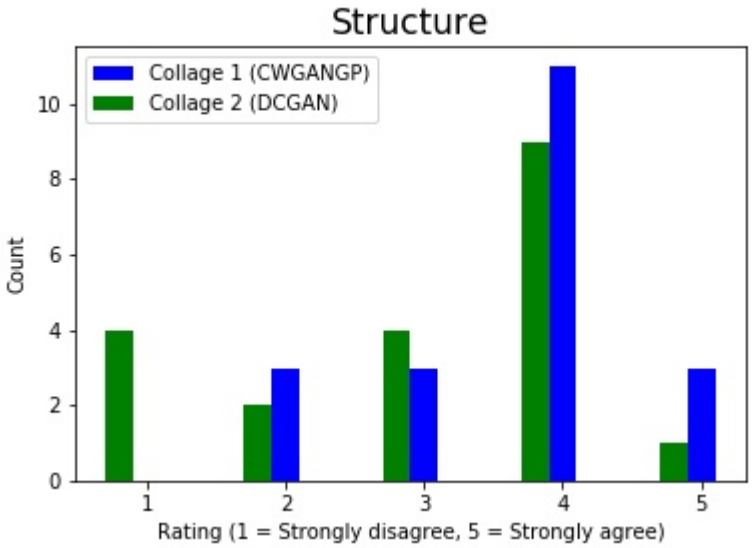


Figure 5.3: Structure in the collages

Both GAN collages were viewed as having structure by the respondents. However, 70% of respondents agreed that the CWGANGP collage showed structure in its images, while only 50% of respondents agreed that the DCGAN collage showed structure. Figure 5.4 shows the ratings for 'hallucinatory' assigned to the collages.

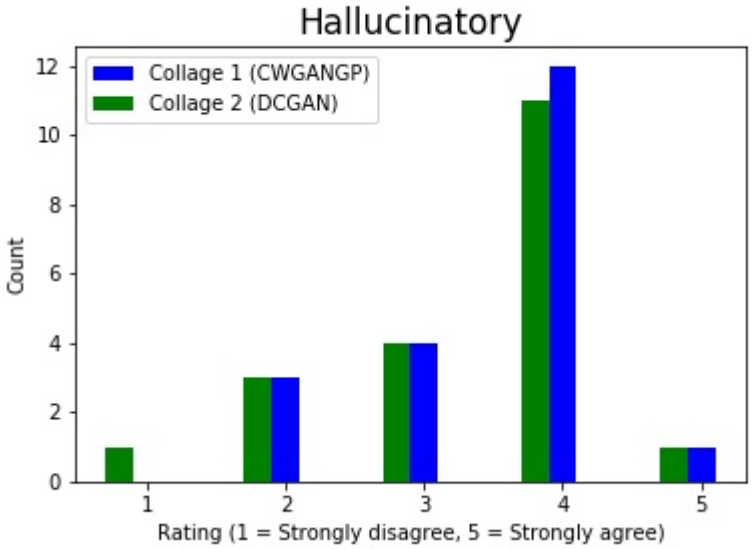


Figure 5.4: Hallucinatory elements in the collages

The respondents agreed that both collages were hallucinatory. Interestingly, though a more diverse collage, a greater proportion of respondents viewed the CWGANGP collage as more hallucinatory than the DCGAN collage. Figure 5.5 shows the overall ratings assigned to the collages.

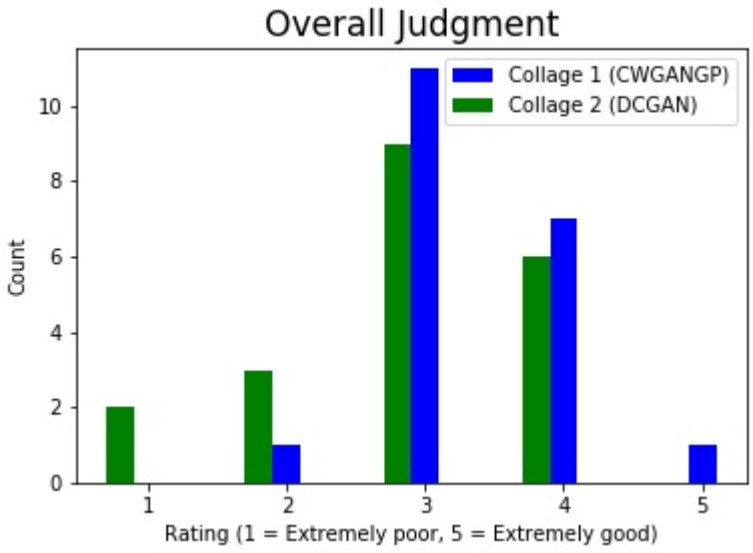


Figure 5.5: Overall Judgment

While both collages were mostly viewed as halfway between 'Extremely poor' and 'Extremely good', the CWGANGP collage was more positively received than the DCGAN collage, with 40% of respondents judging the former as 'Good' or 'Extremely good'. In contrast, the latter was judged as 'Good' by only 30% of respondents.

Table 5.14 shows, for each characteristic, how many respondents rated both collages the same, how many rated the DCGAN collage higher, and how many rated the CWGANGP collage higher.

	Not blurry	**Diverse**	**Structured**	**Not hallucinatory**	**Overall**
Rated equally	10	4	8	12	9
CWGANGP preferred	9	13	9	3	9
DCGAN preferred	1	3	3	5	2

Table 5.14: Preferences of respondents in Section 2

When rating the CWGANGP and DCGAN collages overall, nine of the 20 respondents rated them equally, nine rated CWGANGP higher and only two rated DCGAN higher. Of the 10 respondents who rated the collages differently in terms of blurriness, all but one considered DCGAN more blurry than CWGANGP. This tendency was also apparent, but less pronounced, with regard to structure: of the 12 who rated the collages differently in terms of structure, all but three felt DCGAN showed less structure than CWGANGP. Only four rated the diversity of the two collages equally, with 13 finding CWGANGP more diverse than DCGAN. The only case where the DCGAN collage fared better than CWGANGP was where five found

CWGANGP more hallucinatory, compared to three finding DCGAN more hallucinatory. Thus, Section 2 showed a clear respondent preference for the CWGANGP collage. Table 5.15 shows the means and standard deviations for the qualities of the GAN collages.

Quality	**CWGANGP Collage**	**DCGAN Collage**
Blurriness and noise	3.3±1.35	3.95±1.07
Diversity	3.55±1.07	2.4±1.07
Structure	3.7±0.9	3.05±1.24
Hallucinatory	3.55±0.8	3.4±0.97
Overall	3.4±0.66	2.95±0.92

Table 5.15: Means and standard deviations for qualities of GAN collages

In 5.15 we can see a relationship between the scores for each of the four qualities and the overall score given to each collage. The CWGANGP collage achieved a higher mean score for 'diversity', 'structure' than the DCGAN collage, and a lower score for 'blurriness and noise' and 'hallucinatory'. As the CWGANGP collage was thus seen as less blurry, more structured and more diverse, it achieved a higher overall mean score than the DCGAN collage.

5.2.3 Section Three

In the final section, respondents were asked whether two images, Adolph Gottlieb's *Untitled* and a sample from the same CWGANGP as in the previous sections, were human-generated or computer-generated in a Turing test. Table 5.16 shows the results of this Turing test.

Painting	**Human-generated Votes (%)**	**Computer-generated Votes (%)**
Untitled	50	50
CWGANGP Sample	40	60

Table 5.16: Turing Test Results

Interestingly, while the majority of respondents classed the sample from the CWGANGP as computer-generated, the respondents were evenly split on the origin of Gottlieb's *Untitled*. Table 5.17 shows a breakdown of the responses in the Turing test.

Response	Number of respondents
Both answers incorrect	6
Both images human-generated	2
Both images computer-generated	4
Both answers correct	8

Table 5.17: Turing Test Responses

The correct response was the first artwork was human-created and the second was computer-generated. While 8 had both correct, 6 had both incorrect. The others said both were computer-generated (4 respondents) or both human-generated (2 respondents).

5.3 Visual Quality of Samples

In this work, unless otherwise stated, all samples shown are those produced in the last training epoch. A selection of samples produced by the GANs in this work across all data sets can be found in the appendices.

Table 5.18 shows a subjective ranking of VQ of the base GANs on the benchmark data sets.

	Data Set		
GAN	MNIST	CIFAR-10	Imagenet-1k
DCGAN	2	**1**	4
CDCGAN	**1**	2	5
WGAN	5	3	6
CWGAN	4	5	7
WGANGP	6	4	2
CWGANGP	3	6	3
IWGAN	-	7	**1**

Table 5.18: VQ Rankings of Base GANs on Benchmark Data Sets

The visual quality of samples (VQ) differed considerably across GANs. On the smaller benchmark data sets, MNIST and CIFAR-10, both of which have 50,000 training samples, the non-conditional and conditional DCGAN produced the samples of the highest VQ. On the much larger data set, Imagenet-1K, the IWGAN produced the samples of the highest quality.

Table 5.19 shows HEART Section Two ratings of the base GANs trained on Augmented Wikiart (1 = Strongly disagree, 5 = Strongly agree). This table informed the ranking of VQ of these GANs.

GAN	Blurriness and noise	Diversity	Structure	Hallucinatory	Overall
DCGAN	5	2	1	2	1
CDCGAN	3	3	2	3	3
WGAN	3	4	2	3	3
CWGAN	5	2	1	1	2
WGANGP	1	4	3	4	4
CWGANGP	1	5	3	4	4
IWGAN	5	1	1	3	1

Table 5.19: HEART Section Two Ratings of Base GANs

Table 5.20 shows the ranking of VQ of the base GANs on the Augmented Wikiart data set.

GAN	Augmented Wikiart VQ Rank
CWGANGP	**1**
WGANGP	2
WGAN	3
CDCGAN	4
CWGAN	5
DCGAN	6
IWGAN	7

Table 5.20: VQ Rankings of Base GANs on Augmented Wikiart

With the exception of the IWGAN and the CWGAN, the Wasserstein GANs produced samples of high quality.

5.4 GAN Training Time

As mentioned previously, all GAN models were trained using a Nvidia Tesla V100-SXM2-16GB. Table 5.21 shows the training time of the base GANs on Imagenet-1K.

GAN	Training Time
DCGAN	3h28m
CDCGAN	2h30m
WGAN	3h11m
CWGAN	**1h38m**
WGANGP	5h27m
CWGANGP	2h6m
IWGAN	26h23m

Table 5.21: Base GAN Training Time - Imagenet-1K

Table 5.22 shows the training time of the base GANs on CIFAR-10.

GAN	Training Time
DCGAN	**6m40s**
CDCGAN	11m32s
WGAN	7m15s
CWGAN	6m52s
WGANGP	6m47s
CWGANGP	10m9s
IWGAN	2h8m

Table 5.22: Base GAN Training Time - CIFAR-10

Table 5.23 shows the training time of 32×32 DCGANs with GANHacks, as well as a 32×32 DCGAN without GANHacks, on Imagenet-1K.

GAN	Training Time
DCGAN without GANHack	2h27m
Dropout (0.5) in G	2h28m
`LReLU` in G	2h28m
Flipped labels in D	**2h23m**
Smoothed labels in D	2h26m
FLND	2h35m
FLN	2h28m
FLD	2h37m
FL	2h48m

Table 5.23: GANHacks Training Time - Imagenet-1K

5.5 Relationship between Quantitative Performance and VQ

Table 5.24 shows the relationship between IS, FID and VQ of the base GANs on Imagenet-1K.

GAN	Inception Score (IS)	Fréchet Inception Distance (FID)	VQ Rank
IWGAN	3.47±0.04	**103.7**	**1**
WGANGP	3.42±0.04	150.01	2
CWGANGP	2.47±0.05	203.69	3
DCGAN	**3.82±0.02**	146.94	4
CDCGAN	2.04±0.03	222.29	5
WGAN	2.22±0.02	279.05	6
CWGAN	1.79±0.03	265.61	7

Table 5.24: Comparison of IS, FID and VQ of Base GANs on Imagenet-1k

On Imagenet-1k, we can see a relationship between quantitative scores and VQ. The base GANs that ranked highest in VQ (IWGAN, WGAN, CWGANGP and DCGAN respectively) achieved the four best Inception Scores *and* the four best FID scores. The lowest ranking GANs by VQ had the worst FID scores as well as

the worst IS. Table 5.25 shows the relationship between IS, FID and VQ of the base GANs on CIFAR-10.

GAN	Inception Score (IS)	Fréchet Inception Distance (FID)	VQ Rank
DCGAN	**2.00±0.02**	271.64	**1**
CDCGAN	1.57±0.03	**245.18**	2
WGAN	1.68±0.01	281.12	3
WGANGP	1.87±0.01	327.76	4
CWGAN	1.67±0.03	248.56	5
CWGANGP	1.65±0.03	247.79	6
IWGAN	1.66±0.01	273.36	7

Table 5.25: Comparison of IS, FID and VQ of Base GANs on CIFAR-10

On CIFAR-10, we can see a weaker relationship between quantitative scores and VQ than on Imagenet-1k. The base GANs that ranked highest in VQ (DCGAN, CDCGAN, WGAN and WGANGP respectively) achieved the best Inception Scores of the base GANs, with the exception of CDCGAN. However, while CDCGAN achieved the best FID, the remaining high-ranking base GANs with respect to VQ largely achieved worse FID scores than the base GANs who ranked lower in VQ. The GANS that ranked lowest in VQ achieved poor IS and FID scores. Table 5.26 shows the relationship between FID and VQ of the base GANs on MNIST.

GAN	Fréchet Inception Distance (FID)	VQ Rank
CDCGAN	51.76	**1**
DCGAN	96.77	2
CWGANGP	92.78	3
CWGAN	**44.52**	4
WGAN	59.65	5
WGANGP	122.45	6

Table 5.26: Comparison of FID and VQ of Base GANs on MNIST

On MNIST, we can see a weak relationship between quantitative score (FID) and VQ. Though the base GAN that ranked first in VQ, CDCGAN, achieved the second-best FID score, the remaining base GANs that ranked highest in VQ (DCGAN and CWGANGP) achieve worse FID scores than WGAN, which ranked lower in VQ. However, the GAN that ranked lowest in VQ also had the worst FID score.

5.6 Summary

The majority of GANs did not surpass Radford *et al.*'s accuracy in the classification of unseen data experiment. DCGAN achieved the best Inception Score (S) on both CIFAR-10 and Imagenet-1K, though it did not achieve the best Fréchet Inception Distance (FID) score on any of the data sets tested. Though DCGAN produced the samples of the highest visual quality (VQ) on MNIST and CIFAR-10, most Wasserstein GANs and some GANHacks produced samples of a higher VQ than DCGAN on the larger Imagenet-1K data set and in particular on all art data sets.

In the HEART pilot study, it was found that the CWGANGP collage evoked emotional responses and was judged favourably, but to a lesser extent than *The Scream*, as expected. Overall judgments of GAN collages were found to be influenced by judgements of qualities such as diversity, and it was shown that in one instance, a GAN-produced image could pass the Turing Test. IWGAN produced the samples of the highest VQ on Imagenet-1K, but had a dramatically longer training time than all other GANs. Typically, the GANs that performed well quantitatively (by IS and FID) also performed well qualitatively, especially so on Imagenet-1K.

Chapter 6

Discussion

6.1 Research Questions

6.1.1 *To what extent does a GAN's quantitative performance align with its qualitative performance on benchmark data sets?*

The relationship between quantitative performance and qualitative performance of the GANs was especially seen on Imagenet-1K. The top four base GANs in terms of IS and FID achieved the four best VQ ranks on Imagenet-1K, though the best IS achieved (DCGAN) ranked fourth by VQ. The relationship was weaker on CIFAR-10 and MNIST. Typically, base GANs that performed well on CIFAR-10 did so in both IS and VQ. Similarly, the base GANs that achieved better FID scores tended to rank higher with respect to VQ. Thus, a GAN's quantitative evaluation aligns well with its qualitative evaluation on benchmark data sets. It must be noted, however, that it was not the case that a GAN which achieved better quantitative scores than another would *always* achieve higher VQ.

6.1.2 *Does the use of GANHacks improve qualitative, quantitative or runtime performance of DCGAN?*

The GANHacks of `LReLU` in G and flipped labels in D, and the combination thereof (FL-DCGAN), achieved a higher un-optimised CIFAR-10 classification accuracy than DCGAN. In some cases GANHacks led to better non-optimised classification accuracy on SVHN. However, in all other cases, the DCGAN without GANHacks achieved higher accuracy. No GANHack nor combination thereof achieved a better IS than DCGAN on Imagenet-1K. However, some GANHacks did result in a slightly better FID than the DCGAN without GANHacks on Imagenet-1K, and the FL-DCGAN did achieve a considerably better FID than DCGAN on MNIST. Thus, the use of GANHacks largely did *not* improve quantitative performance of DCGAN.

The VQ of samples trained on Imagenet-1K with the flipped labels in D, smoothed labels in D, `LReLU` in G and the FL-DCGAN were all higher than that of the DCGAN samples. The GANHacks produced more colourful, brighter and more structured samples than DCGAN on this data set. On the Augmented Wikiart data set, the smoothed labels in D and flipped labels GANHacks exhibited improved qualitative performance over DCGAN. It must be noted that while the FL GANHack combination and the GANHack of using `LReLu` in G collapsed in the last three epochs of training, and thus produced samples of lower quality than DCGAN, their samples produced earlier in training were of higher quality than those of DCGAN at the same point of training. Prior to their collapsing, both of these GANs improved

qualitative performance of DCGAN. Thus, if sample visual quality of an art GAN is prioritised, and one produces samples during training, a semi-trained DCGAN with GANHacks will have better qualitative performance than a semi-trained DCGAN without GANHacks. Thus, the use of some GANHacks did improve qualitative performance of DCGAN.

Only two GANHacks, the use of flipped labels in D and the use of smoothed labels in D, improved training time of DCGAN on Imagenet-1K, and only by a few minutes. The use of GANHacks thus largely did not improve runtime performance of DCGAN.

6.1.3 *Does the use of Wasserstein GANs improve quantitative, qualitative or runtime performance of DCGAN?*

Several Wasserstein GANs achieved better FID scores than DCGAN. On Imagenet-1K, IWGAN achieved a better FID than DCGAN, and on CIFAR-10, CWGAN and CWGANGP both achieved better a better FID than DCGAN. Finally, with the exception of WGANGP, all Wasserstein GANs achieved a better FID than DCGAN on MNIST. However, no Wasserstein GAN achieved a higher Inception Score than DCGAN. Thus, the use of Wasserstein GANs did improve quantitative performance of DCGAN under the FID metric, though not under the IS metric.

Wasserstein GANs produced samples of a higher VQ than DCGAN on Imagenet-1K and all art data sets. These data sets are the larger of the data sets used in this work. The art-trained CWGANGP's samples were better received in the HEART study than the art-trained DCGAN. DCGAN produced higher quality samples than the Wasserstein GANs on CIFAR-10 and MNIST. The use of Wasserstein GANs therefore did improve qualitative performance of DCGAN on some benchmark data sets and on all art data sets. It must be noted that WGANGP performed better than WGAN both quantitatively and qualitatively. WGANGP achieved better IS, FID and produced samples of higher VQ than WGAN.

While WGAN had slightly faster training times than DCGAN on Imagenet-1K and CIFAR-10, the remaining Wasserstein GANs had longer training times than DCGAN. IWGAN required 600% longer to train than DCGAN on Imagenet-1K, and was the only GAN to require hours, rather than under ten minutes, to train on CIFAR-10. Therefore, the use of *some* Wasserstein GANs improved runtime performance of DCGAN. The use of IWGAN is thus not recommended for use on modest budgets, due to its far longer training times than other Wasserstein GANs.

6.1.4 *Does improved qualitative performance of GANs on benchmark data sets translate to better qualitative performance on art data sets?*

DCGAN had superior qualitative performance on MNIST and CIFAR-10 to the other base GANs, but its qualitative performance on the art data sets was lower than other base GANs. In contrast, all Wasserstein GANs except IWGAN had higher VQ than DCGAN on all art data sets. The GANs that performed better qualitatively on Imagenet-1K, the larger of the benchmark data sets, also performed better on art data sets, with the exception of IWGAN. However, the GANs that had higher VQ on the smaller benchmark data sets tended not to have better qualitative

performance on art data sets. Therefore, improved qualitative performance of GANs on benchnark data sets does not translate to better qualitative performance on art data sets.

6.1.5 *Can the proposed qualitative evaluation method successfully be used to evaluate the emotional impact, cognitive impact, visual quality and creativity of the creations of art GANs?*

Respondents were able to complete the HEART survey and gave interesting results. Both GAN-generated and human-generated art were evaluated cognitively and with regard to emotional impact. The GAN-generated collage was well-received compared to Munch's *The Scream*, as the latter scored better where expected, but not by too great a margin. HEART also allowed for evaluating *and* substantiating VQ judgments and preferences among GANs with the examination of qualities of the art works such as blur, an examination which informed overall judgment of sample quality. Finally, in the Turing Test section of HEART, the majority of respondents failed to distinguish GAN-generated art from human-generated art given one instance of each, showing that in some cases GAN-generated art can be as creative as human-generated art. Thus, the pilot survey of HEART was successful.

6.2 Key Findings

1. GANHacks do not improve quantitative performance of DCGAN, though some improve VQ.

2. Typically, a GAN's quantitative performance aligns well with its qualitative performance on benchmark data sets.

3. Wasserstein GANs achieved higher VQ than DCGAN on Imagenet-1K and all art data sets.

4. Good qualitative performance on benchmark data sets does not necessarily translate to good qualitative performance on art data sets.

5. HEART can be used to evaluate the emotional and cognitive impacts, as well as important characteristics of GAN-produced images.

6.3 GAN Training Behaviour

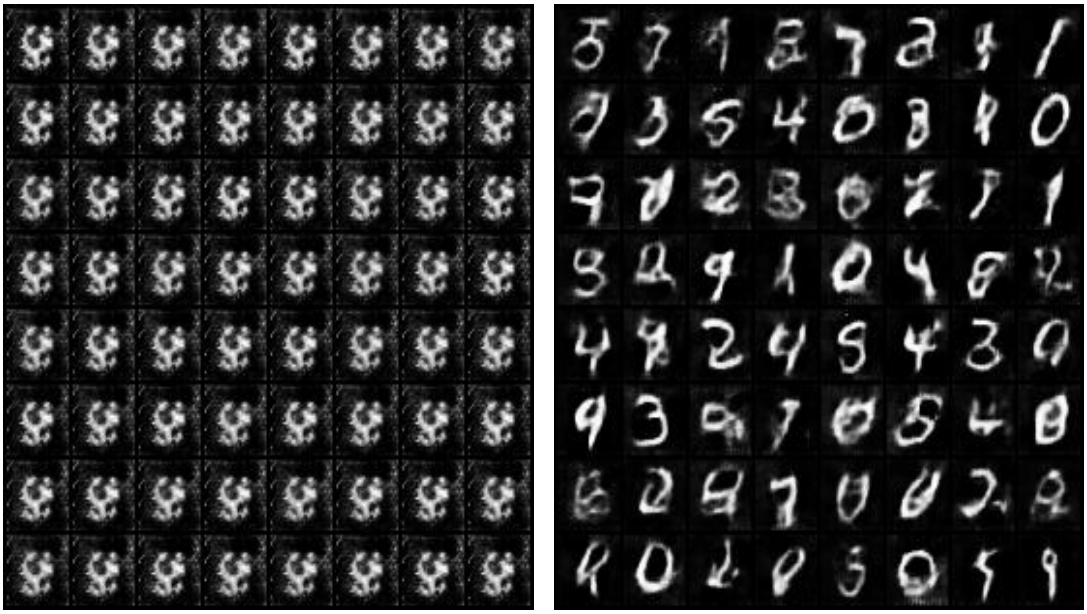
Perhaps the most salient of observations across the hundreds of GAN models trained in this work is the inherent erratic behaviour of GANs. On multiple occasions, the exact same GAN model would collapse on one run but learns to produce high-quality images on another. This erratic behaviour may have been exacerbated by the use of a random seed - the *de facto* approach of official implementations of the base GANs. GANs are noted to be sensitive to random seeds [Lucic et al., 2017].

GANs could collapse at any point during training. Figure 6.1 shows a GAN, trained on the Augmented Wikiart, which collapsed in its eighth epoch of training.

GAN: Epoch 8 GAN: Epoch 9

Figure 6.1: Collapsed GAN

Some models also failed when trained on simple data sets such as MNIST. Figure 6.2 shows such a collapsed CWGANGP trained on MNIST.

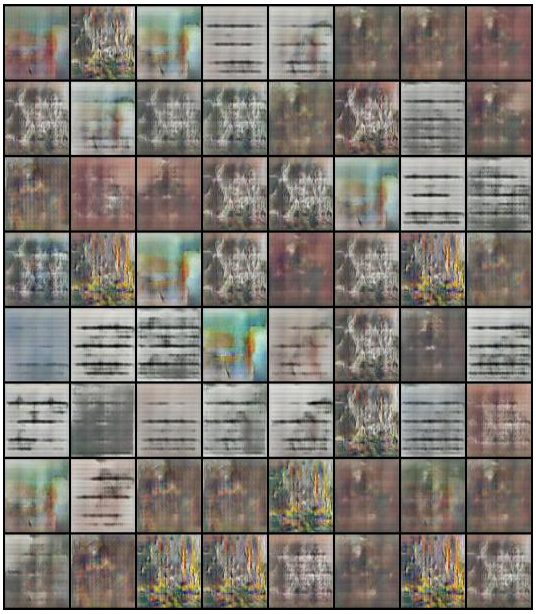


Collapsed CWGANGP trained on MNIST CWGANGP trained on MNIST

Figure 6.2: Collapsed CWGANGP trained on MNIST

At times, the GANs did not collapse, but rather produced *far* inferior samples on different training runs. Figure 6.3 shows the difference in VQ of a DCGAN trained on Augmented Wikiart across different runs.

DCGAN trained on Augmented Wikiart



DCGAN trained on Augmented Wikiart

Figure 6.3: Difference of VQ across runs of DCGAN

6.4 GAN Classification Performance

Despite faithfully following the training guidelines of DCGAN, the classification accuracy achieved by Radford *et al.* could not be replicated on CIFAR-10. As mentioned previously, Radford *et al.* report that the linear model built from the DCGAN discriminator achieves 82.8% accuracy on CIFAR-10 [Radford et al., 2015]. None of the GANs achieved or surpassed this accuracy - even when optimised using *hyperopt-sklearn*. In contrast, while no GAN was able to achieve or surpass the accuracy of Radford *et al.* on CIFAR-10, their 77.52% accuracy on SVHN [Radford et al., 2015] was surpassed by the optimised versions of DCGAN, WGANGP and CWGANGP.

The additions of GANHacks meant to advantage the discriminator (the inclusion of dropout and Gaussian noise in G) did not lead to improved classification accuracy. Though counter-intuitive, a stronger discriminator thus does *not* necessarily lead to improved classification accuracy. It may be that generators in the GANHack-modified GANs were dominated by their discriminators throughout training and thus could not provide useful gradients to their corresponding strengthened discriminators.

In their original papers, Wasserstein GANs are not evaluated by discriminator classification accuracy. In addition to its noted improvements [Gulrajani et al., 2017] to sample quality over WGAN, WGANGP achieved notably better classification accuracies than WGAN on both CIFAR-10 and SVHN.

6.5 Inception Scores

Though not calculated in the Radford *et al.* paper, DCGAN is reported to achieve an IS of 6.5 [Wang and Liu, 2016, Bang and Shim, 2018]. As with the discriminator classification experiment, none of the GANs in this work achieved or surpassed this

IS. The best IS achieved in this work was 3.82, from the DCGAN without GANHacks. It must be noted that the reported IS for DCGAN [Wang and Liu, 2016, Bang and Shim, 2018] was obtained via the training methodology of the *TensorFlow* implementation of DCGAN, in which G is trained twice per training iteration of D, as well as the official *TensorFlow* IS implementation [Bang and Shim, 2018]. The *PyTorch* implementation of the IS [Salimans et al., 2016] notes that *TensorFlow* does not correctly implement the IS algorithm as described in the original IS paper [Salimans et al., 2016], and that the correct *PyTorch* version achieves lower scores than those reported using the *TensorFlow* implementation. It is thus unclear if the Inception Scores found in this work, calculated using the *PyTorch* version of the IS, are indeed lower. Nevertheless, the DCGAN without GANHacks achieved the highest IS of all GANs tested on both CIFAR-10 and Imagenet-1K. WGANGP achieved a superior IS to WGAN, though IWGAN's IS was higher than WGANGP on Imagenet-1K.

Interestingly, all individual GANHacks produced nearly identical Inception Scores, except for a dramatically lower IS when `LReLU` was used in G. Since using `LReLU` in G should lead to a better generator, this should produce superior samples and thus achieve higher Inception Scores. The FLD GAN, which featured flipped labels in D as well as dropout and `LReLU` in G, achieved a considerably higher IS than all other GANHack combination GANs.

6.6 Fréchet Inception Distances

Typically, the GANs tested achieved FIDs in line with their Inception Scores; those that obtained higher Inception Scores obtained lower FIDs. While again DCGAN achieved a better (lower) FID than other GANs, IWGAN trained on Imagenet-1K achieved the best FID of all GANs tested; and one considerably better than DCGAN. While WGANGP obtained a better FID on Imagenet-1K, it achieved the worst FID of any of the base GANs on CIFAR-10. Interestingly, the FIDs on CIFAR-10 were worse for all base GANs. With the exception of WGAN, the FIDs of the base GANs were *considerably* worse on the smaller CIFAR-10. Mirroring the classification and IS experiments, the unmodified DCGAN obtained a better FID than the GANHack-modified DCGAN, with the exception of the DCGAN with flipped labels in D. As with its IS, the FLD GAN obtained a considerably better FID than the other GANHack combinations, though its FID was worse than the DCGAN with flipped labels in D.

6.7 GAN Sample Visual Quality and Data Sets

6.7.1 Benchmark Data Sets

The size and complexity of data set influenced the sample VQ of all GANs. GANs trained on the smaller data sets, MNIST and CIFAR-10, produced samples of a lower quality than ones trained on the larger Imagenet-1K data set. Both MNIST and CIFAR-10 consist of 50000 training samples. The samples produced by GANs trained on MNIST had higher VQ than those trained on CIFAR-10, as MNIST is a less complex data set. Though both MNIST and CIFAR-10 have ten classes, the images of MNIST are black-and-white, as opposed to the CIFAR-10, and the visual differences between the classes of MNIST, handwritten digits, are smaller than those of the classes of CIFAR-10, which include animals and vehicles.

As Imagenet-1K has 1000 classes, it is far more complex than both MNIST and CIFAR-10. However, it has 14 million training samples. Thus, despite its complexity, the GANs, particularly IWGAN, were able to create images of high VQ, albeit at a cost of a much longer training time.

6.7.2 Art Data Sets

As with the benchmark data sets, the size and complexity of the art data sets influence the sample VQ of all GANs. As Wikiart contains 27 art styles but only 80000 training images, GANs trained on the native Wikiart were tasked with creating complex, varied images using limited data. Thus, GANs such as DCGAN were only able to produce noisy images when trained on Wikiart. Through augmenting the data ten-fold to produce Augmented Wikiart, the GANs were able to produce images of much higher VQ. Despite only having one class, the portraits, landscapes and cityscapes art data sets were of similar size to MNIST and CIFAR-10 and thus the images produced by GANs trained on them were of lower VQ than those produced by GANs trained on Augmented Wikiart.

6.8 HEART

6.8.1 Section One - Cognitive and Emotional Impact

Both *The Scream* and the CWGANGP collage evoked emotions, particularly negative emotions, in the respondents. That the GAN-produced art (the CWGANGP collage) did arouse emotion is especially significant, as its direct competitor, *The Scream*, is famed for its arousing unease and similar emotions. The response to the CWGANGP collage was more neutral than to *The Scream*. Similarly, both *The Scream* and the CWGANGP collage were well-received, with regard to the cognitive attributes. Though again *The Scream* received a more positive response, the CWGANGP collage was thought of as 'interesting', 'original' and 'fascinating'. Thus, despite its direct competition being a famous artwork, cognitively, the GAN-generated art was judged favourably. Together, this suggests that due to the CWGANGP collage's emotional and cognitive impacts, the respondents viewed the GAN-generated art as good.

The respondents of the pilot study gave interesting feedback on this section, which may impact further implementations of HEART. One respondent noted that it may not be suitable for those who are not first-language English speakers, due to its use of uncommon words such as 'disquiet'. A common point made by respondents was that some emotion factors were too similar to others, such as 'joy' and 'gladness', and the section felt more laborious and repetitive as a result. The respondents also noted this problem in the attribute list, such as 'creative' and 'inventive', though this was a less-reported point. The respondents found the Likert scale easy to understand, and did not remark that a finer-grained scale (such as the nine-point one used by Hagtvedt *et al.*) would be more useful. In fact, five options for each emotion and each cognitive attribute was sufficient, as there was much variety in the responses, which led to spread-out results. Many respondents felt it was difficult to judge *multiple* artworks, such as the collages produced by the CWGANGP, as while some of the images of the collage may induce certain emotions, others may not, rendering it unclear how to assign an *overall* judgment. While a *single* GAN-generated

image would avoid this issue, it was felt that a single, small GAN image could not compete with a full-size artwork such as the *The Scream*.

6.8.2 Section Two - GAN Comparison

Though half of the respondents judged the DCGAN collage and CWGANGP equally overall, the CWGANGP collage was better received than the DCGAN collage in terms of three criteria of blurriness and noise, diversity, and structure. The CWGANGP collage was viewed as more diverse and more structured than the DCGAN by the majority of respondents, and fewer respondents viewed it as the more hallucinatory collage. The CWGANGP collage was also the only collage to be judged 'extremely good' overall. This, coupled with nine of the remaining ten respondents' viewing it as the superior collage, indicate that there is a relationship between these criteria and the overall judgment of the visual quality of GAN samples.

As with the first section, many respondents noted the difficulty of assigning characteristics, such as structure, to the *whole* collage, given that not all images of the collage may feature said characteristics. However, it was felt that to comprehensively judge a GAN, multiple samples were needed. Some respondents remarked that the supplementary text in the first statement '(e.g. speckled dots, weird lines)' clarified what was meant by noise, and that without this additional text, they would not have understood the statement. Unsurprisingly, given GANs' erratic training and noted presence of noise, the majority of respondents observed noise and blurriness in both collages. It was hoped that the collage from the more sophisticated GAN, the CWGANGP, would *not* be viewed as noisy, but just over half of the respondents thought it was. Despite the majority of the images of the CWGANGP collage being abstract, the collage was viewed as both having structure and diverse. Unfortunately, according to the majority of respondents, both GAN collage fell victim to the noted hallucinatory tendencies of computer-generated art [Elgammal et al., 2017]. As with section one, respondents remarked that HEART may pose difficult for those who are not first-language English speakers, as 'hallucinatory' is not a commonly-used word. Moreover, some respondents required clarification on the meaning of 'hallucinatory'. Surprisingly, though featuring many highly dark and highly bright, plain-looking images, the DCGAN collage was not judged much worse than the CWGANGP collage, as seen in Tables 3.5 and 5.11.

6.8.3 Section Three - Turing Test

As only two cherry-picked images were used in this Turing Test, no definitive conclusions can be drawn, but it was noteworthy that neither image had a clear origin to the respondents.

6.8.4 Comparison with Existing Evaluation Approaches

HEART builds upon existing evaluation approaches to GAN-produced artworks, as it examines both the cognitive impact and the emotional impact of the artworks, while other approaches, such as that of the Creative Adversarial Network (CAN), examine only the cognitive impact. HEART also considers qualities noted to be common to artificially-generated artworks, such as noise, as part of its evaluation of GAN-produced artworks. However, unlike CAN's evaluation, HEART does not make use of statistical tests such as the t-test to supplement judgments of the difference of responses to artworks of different GANs.

Chapter 7

Conclusions and Future Work

7.1 Summary

This work compared DCGAN [Radford et al., 2015] with the Wasserstein GANs [Arjovsky et al., 2017, Gulrajani et al., 2017], and investigated the incorporation of GANHacks [Chintala et al., 2016] in DCGAN, on a limited computational budget, to investigate whether Wasserstein GANs and GANHacks can improve upon DCGAN-based fine art generation. DCGAN and the Wasserstein GANs were evaluated quantitatively on the MNIST [LeCun et al., 1998], Imagenet-1K [Deng et al., 2009] and CIFAR-10 [Krizhevsky, 2009] data sets using the Inception Score and Fréchet Inception Distance. The classification accuracy of their discriminators on the CIFAR-10 and SVHN [Netzer et al., 2011] data sets, as well as of DCGANs with GANHacks added, was calculated. The qualitative performance of these GANs on these data sets and on the Wikiart fine art data set was examined. The Holistic Evaluation of Art (HEART) tool for qualitative evaluation of art GANs was proposed and tested.

7.2 Findings

A GAN's quantitative performance was found to typically align well with its qualitative performance on benchmark data sets, and especially so on the large Imagenet-1K data set. Wasserstein GANs produced the samples of the best visual quality on art data sets as well as Imagenet-1K. GANs that performed well qualitatively on benchmark data sets did not necessarily also perform well qualitatively on art data sets. The flipped labels and smoothed labels in D and `LReLU` in G GANHacks were found to improve the visual quality of DCGAN's samples. HEART was successfully used by 20 students to evaluate the emotional and cognitive impacts, as well as important characteristics of GAN-generated art.

7.3 Limitations

Due to a limited computational budget, for each GAN, data set and quantitative measure, one score was calculated as each GAN model was trained only once. Due to this limited budget, the size of GAN samples produced in this work were constrained to a maximum of 64 pixels. As the samples produced by the Creative Adversarial Network (CAN) are 256 pixels [Elgammal et al., 2017], a fair visual quality comparison between the samples of the GANs in this work and those of prominent fine art GANs such as CAN was not possible.

The results of the pilot study of HEART are limited by the relatively small scale

of the experiment. Only 20 respondents participated. Moreover, these participants evaluated a single human-generated artwork and the output of only two GANs, instead of a range of GANs and various artworks. The Turing test section of the HEART pilot study also used a cherry-picked GAN sample and a human-generated artwork specifically selected because of its abstract nature.

7.4 Conclusions

While Wasserstein GANs produced the highest quality samples on art data sets and the large Imagenet-1K data set, the qualitative performances of GANs on benchmark data sets do not necessarily mirror their qualitative performance on art data sets. However, quantitative performance of GANs typically mirrors qualitative performance on benchmark data sets. While limited, HEART can be used to compare GAN-produced artworks, and to compare a GAN's artwork with human-authored art.

7.5 Future Work

Varying the hyper-parameters of the GANs tested in this work, and obtaining quantitative scores for each configuration of these GANs to provide *averaged* scores for each GAN and data set, is needed to confirm the results of the IS and FID experiments of this work.

GANs that generate large images, such as StackGAN, could be used to generate larger art images [Huang et al., 2016]. Such images would allow for fairer comparisons with human-generated art. Training the GANs on different visual art data sets, such as the *Behance Artistic Media* digital art data set [Wilber et al., 2017], would allow for the investigation into the ability of GANs to create digital art - an investigation which has not yet occurred. Similarly, the use of a non-Western fine art data set would allow for a more comprehensive investigation into the ability of GANs to produce visual art as a whole; art of various styles and origins.

Comprehensive implementations of HEART would provide evidence to support the motivation by this work that HEART is indeed a suitable and appropriate method for qualitative evaluation of visual art. Comprehensive implementations would entail surveying users of varying skill and knowledge of art appreciation, surveying both large and small-scale groups, and comparing various GANs' samples with various human-generated works of fine art.

Appendices

Model Architectures

Generator $G(z)$				
Layer	Kernel Size	Batch Normalisation	Activation Function	Output Shape
z	-	No	-	100
ConvTranspose2D_1	[4 x 4]	Yes	ReLU	1024 x 4 x 4
ConvTranspose2D_2	[4 x 4]	Yes	ReLU	512 x 4 x 4
ConvTranspose2D_3	[4 x 4]	Yes	ReLU	256 x 8 x 8
ConvTranspose2D_4	[4 x 4]	Yes	ReLU	128 x 16 x 16
ConvTranspose2D_5	[4 x 4]	Yes	ReLU	64 x 32 x 32
ConvTranspose2D_6	[4 x 4]	No	Tanh	3 x 64 x 64

Table 1: DCGAN 64x64 Generator Architecture

Discriminator $D(x)$				
Layer	Kernel Size	Batch Normalisation	Activation Function	Output Shape
Conv2D_1	[4 x 4]	No	LeakyReLU	3 x 64 x 64
Conv2D_2	[4 x 4]	Yes	LeakyReLU	64 x 32 x 32
Conv2D_3	[4 x 4]	Yes	LeakyReLU	128 x 16 x 16
Conv2D_4	[4 x 4]	Yes	LeakyReLU	256 x 8 x 8
Conv2D_5	[4 x 4]	Yes	LeakyReLU	512 x 4 x 4
Conv2D_6	[4 x 4]	No	Sigmoid	1024 x 4 x 4

Table 2: DCGAN 64x64 Discriminator Architecture

Discriminator$D(x)$				
Layer	Kernel Size	Batch Normalisation	Activation Function	Output Shape
Conv2D_1	[4 x 4]	Yes	LeakyReLU	3 x 32 x 32
Conv2D_2	[4 x 4]	Yes	LeakyReLU	32 x 16 x 16
Conv2D_3	[4 x 4]	Yes	LeakyReLU	64 x 8 x 8
Conv2D_4	[4 x 4]	Yes	LeakyReLU	128 x 4 x 4
Conv2D_5	[4 x 4]	Yes	LeakyReLU	256 x 4 x 4
Linear	[4 x 4]	No	Sigmoid	512 x 1
Cond	[4 x 4]	No	Softmax	512 x 10

Table 3: 64x64 Conditional DCGAN Discriminator

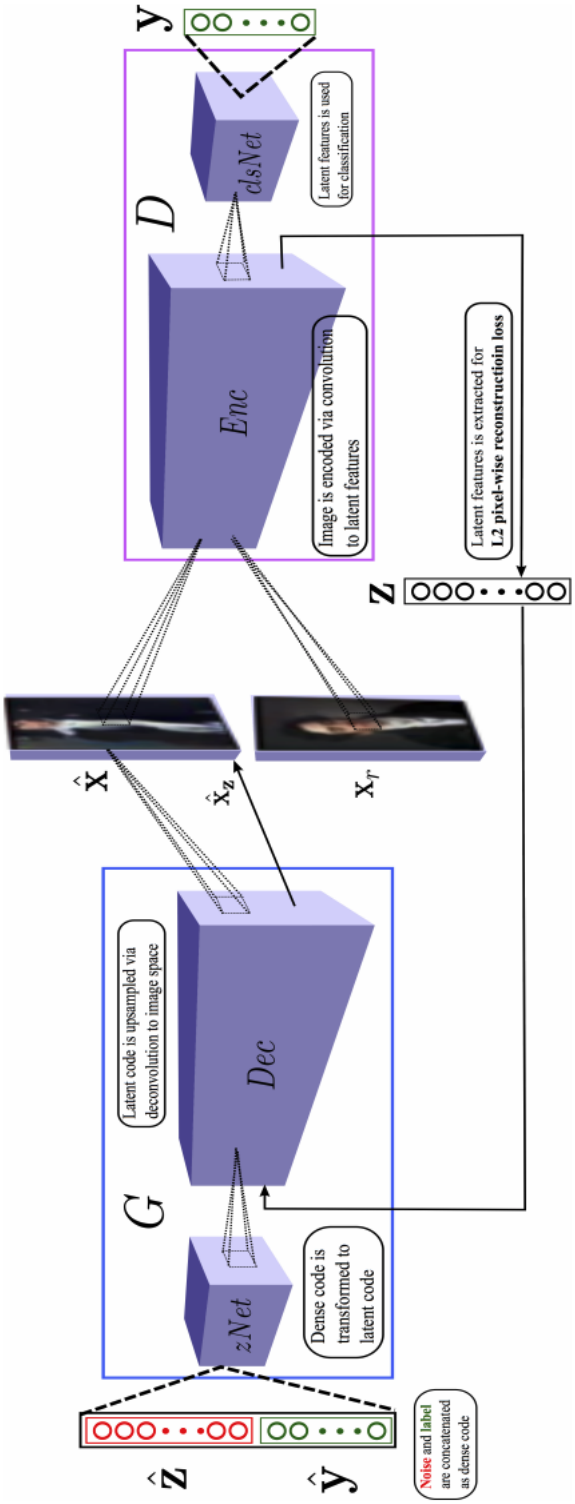


Figure 1: Structure of the ArtGAN

Generator $G(z)$				
Layer	Kernel Size	Batch Normalisation	Activation Function	Output Shape
z	-	No	-	100
ConvTranspose2D_1	[4 x 4]	Yes	ReLU	512 x 4 x 4
ConvTranspose2D_2	[4 x 4]	Yes	ReLU	256 x 8 x 8
ConvTranspose2D_3	[4 x 4]	Yes	ReLU	128 x 16 x 16
ConvTranspose2D_4	[4 x 4]	No	TanH	3 x 32 x 32

Table 4: DCGAN 32x32 Generator Architecture

Discriminator $D(x)$				
Layer	Kernel Size	Batch Normalisation	Activation Function	Output Shape
Conv2D_1	[4 x 4]	No	LeakyReLU	3 x 64 x 64
Conv2D_2	[4 x 4]	Yes	LeakyReLU	64 x 32 x 32
Conv2D_3	[4 x 4]	Yes	LeakyReLU	128 x 16 x 16
Conv2D_4	[4 x 4]	Yes	LeakyReLU	256 x 8 x 8
Conv2D_5	[4 x 4]	No	Sigmoid	512 x 4 x 4

Table 5: DCGAN 32x32 Discriminator Architecture

Discriminator $D(x)$				
Layer	Kernel Size	Batch Normalisation	Activation Function	Output Shape
Conv2D_1	[4 x 4]	No	LeakyReLU	3 x 32 x 32
Conv2D_2	[4 x 4]	Yes	LeakyReLU	32 x 64 x 64
Conv2D_3	[4 x 4]	Yes	LeakyReLU	64 x 32 x 32
Conv2D_4	[4 x 4]	Yes	LeakyReLU	128 x 16 x 16
Linear	[4 x 4]	No	Sigmoid	512 x 1
Cond	[4 x 4]	No	Softmax	512 x 10

Table 6: 32x32 Conditional DCGAN Discriminator

Critic $D(x)$				
Layer	Kernel Size	Batch Normalisation	Activation Function	Output Shape
Conv2D_1	[4 x 4]	No	LeakyReLU	3 x 64 x 64
Conv2D_2	[4 x 4]	Yes	LeakyReLU	64 x 32 x 32
Conv2D_3	[4 x 4]	Yes	LeakyReLU	128 x 16 x 16
Conv2D_4	[4 x 4]	Yes	LeakyReLU	256 x 8 x 8
Conv2D_5	[4 x 4]	No	LeakyReLU	512 x 4 x 4

Table 7: WGAN 64x64 Critic Architecture

Critic $D(x)$				
Layer	Kernel Size	Batch Normalisation	Activation Function	Output Shape
Conv2D_1	[4 x 4]	No	LeakyReLU	3 x 32 x 32
Conv2D_2	[4 x 4]	Yes	LeakyReLU	32 x 64 x 64
Conv2D_3	[4 x 4]	Yes	LeakyReLU	64 x 128 x 128
Conv2D_4	[4 x 4]	Yes	LeakyReLU	128 x 128 x 128
Linear	[4 x 4]	No	Sigmoid	512 x 1
Cond	[4 x 4]	No	Softmax	512 x 10

Table 8: 32x32 Conditional WGAN Discriminator

Critic $D(x)$				
Layer	Kernel Size	Batch Normalisation	Activation Function	Output Shape
Conv2D_1	[4 x 4]	No	LeakyReLU	3 x 64 x 64
Conv2D_2	[4 x 4]	No	LeakyReLU	64 x 32 x 32
Conv2D_3	[4 x 4]	No	LeakyReLU	128 x 16 x 16
Conv2D_4	[4 x 4]	No	LeakyReLU	256 x 8 x 8
Conv2D_5	[4 x 4]	No	LeakyReLU	512 x 4 x 4

Table 9: WGANGP 64x64 Critic Architecture

Critic $D(x)$				
Layer	Kernel Size	Batch Normalisation	Activation Function	Output Shape
Conv2D_1	[4 x 4]	No	LeakyReLU	3 x 32 x 32
Conv2D_2	[4 x 4]	No	LeakyReLU	32 x 16 x 16
Conv2D_3	[4 x 4]	No	LeakyReLU	64 x 8 x 8
Conv2D_4	[4 x 4]	No	LeakyReLU	128 x 4 x 4
Conv2D_5	[4 x 4]	No	LeakyReLU	256 x 4 x 4
Linear	[4 x 4]	No	Sigmoid	512 x 1
Cond	[4 x 4]	No	Softmax	512 x 10

Table 10: 64x64 Conditional WGANGP Discriminator

Generator G(z)			
Layer	Kernel Size	Resample	Output Shape
z	-	-	128
Linear	-	-	128 x 4 x 4
Residual Block	[3 x 3] x 2	Up	128 x 8 x 8
Residual Block	[3 x 3] x 2	Up	128 x 16 x 16
Residual Block	[3 x 3] x 2	Up	128 x 32 x 32
Conv, tanh	[3 x 3]	-	3 x 32 x 32

Table 11: Improved WGAN Generator Architecture

Critic $D(x)$			
Layer	Kernel Size	Resample	Output Shape
Residual Block	[3 x 3] x 2	Down	128 x 16 x 16
Residual Block	[3 x 3] x 2	Down	128 x 8 x 8
Residual Block	[3 x 3] x 2	-	128 x 8 x 8
Residual Block	[3 x 3] x 2	-	128 x 8 x 8
ReLU, mean pool	-	-	128
Linear	-	-	1

Table 12: Improved WGAN Critic Architecture

Generated Samples

MNIST-trained GANs

DCGAN

Conditional DCGAN

WGAN

Conditional WGAN

WGANGP

Conditional WGANGP

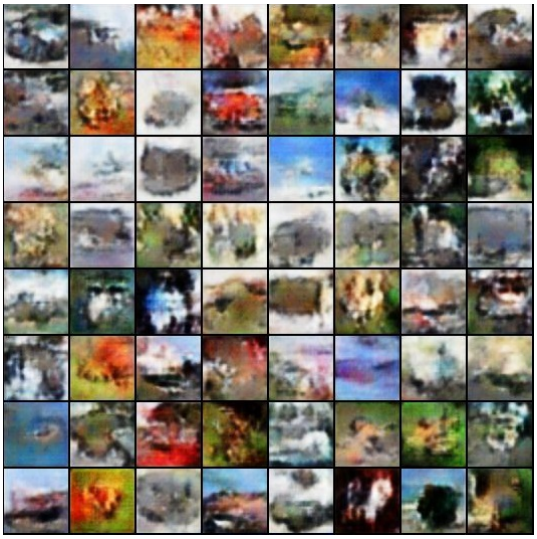
CIFAR-10 Trained GANs

DCGAN

Conditional DCGAN

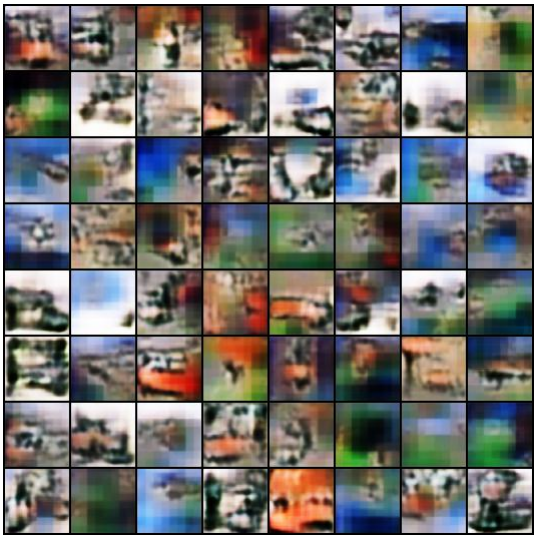
WGAN

Conditional WGAN



WGANGP

Conditional WGANGP



IWGAN

Flipped Labels in D

`LeakyReLu` in G

Dropout in G

Smoothed Labels in D

FLND

FLN

FLD

FL

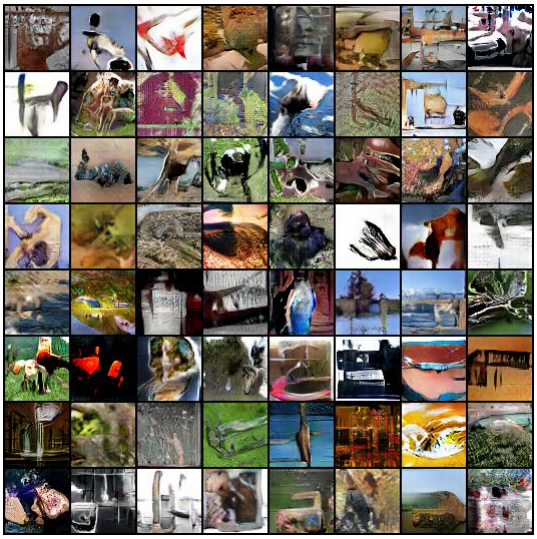
Imagenet-1k Trained GANs

DCGAN

Conditional DCGAN

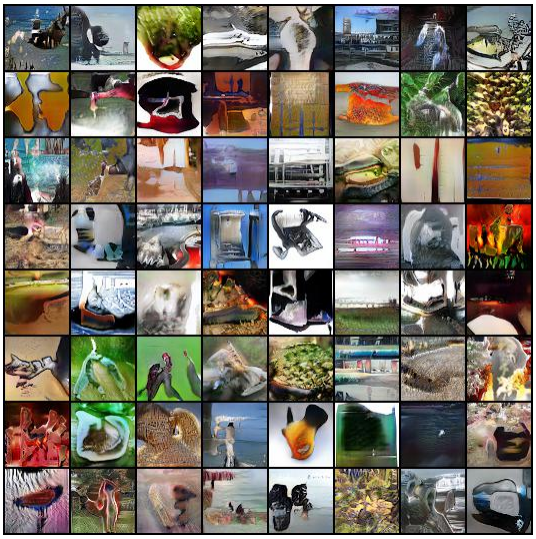
WGAN

Conditional WGAN



WGANGP

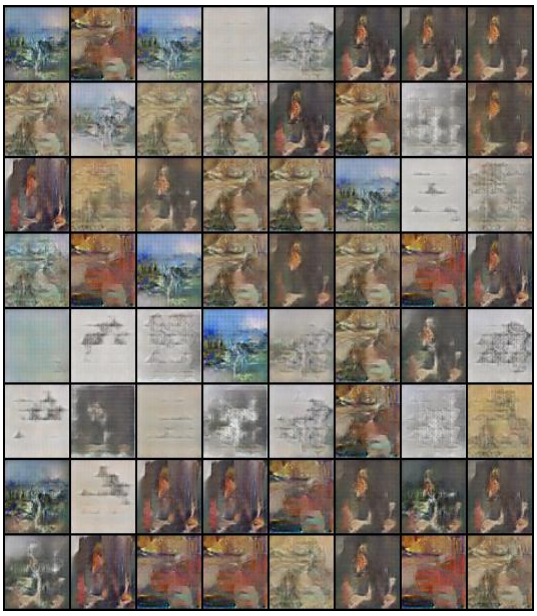
Conditional WGANGP



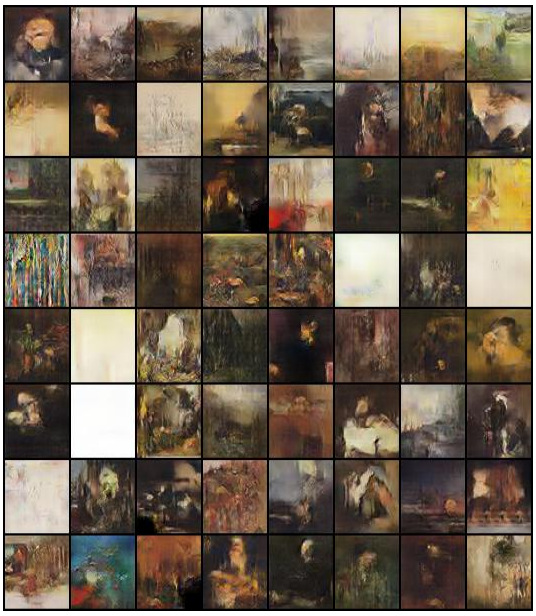
IWGAN

Wikiart Trained GANs

Base GANs



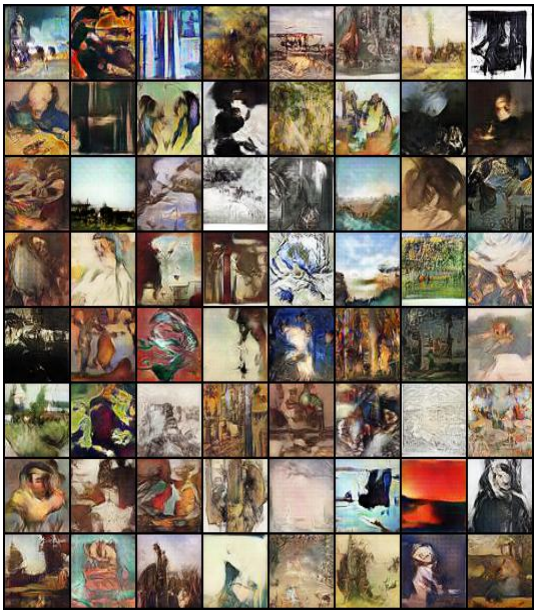
DCGAN



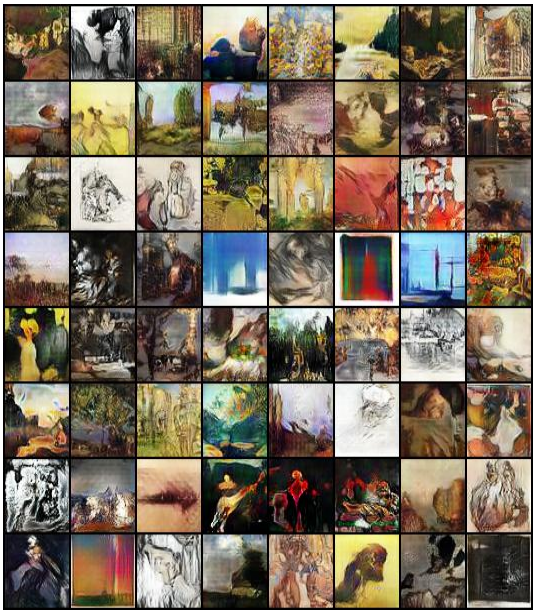
Conditional DCGAN

WGAN

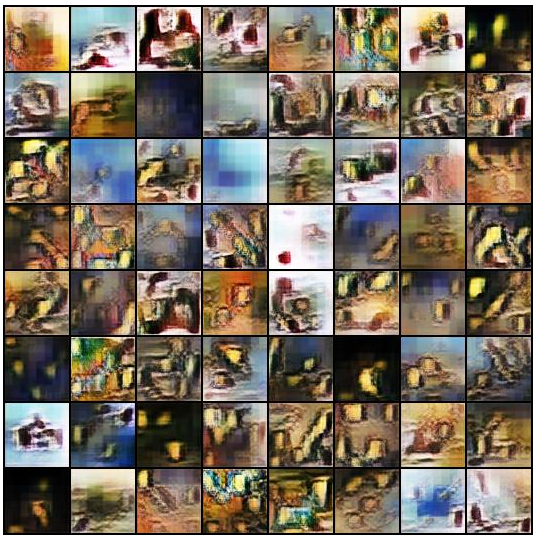
Conditional WGAN



WGANGP

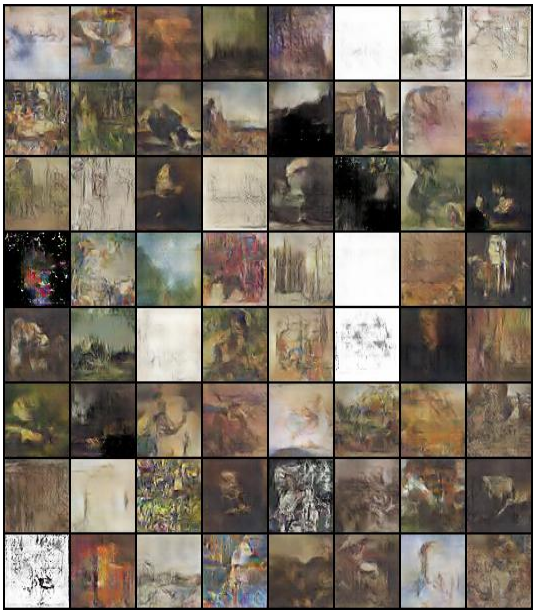


Conditional WGANGP



IWGAN

DCGAN: 5th epoch of training



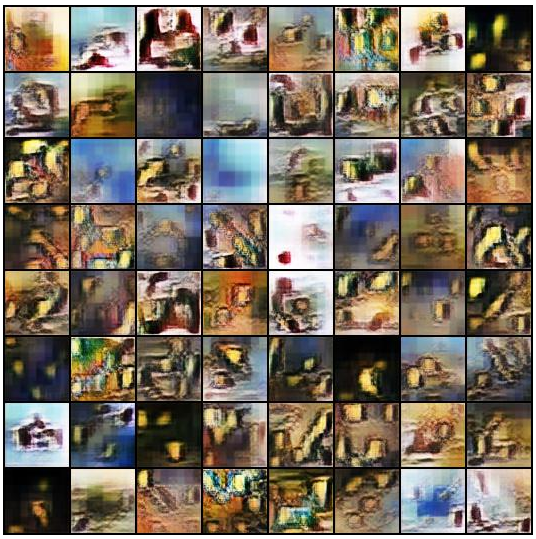
Conditional DCGAN: 5th epoch of training

WGAN: 5th epoch of training

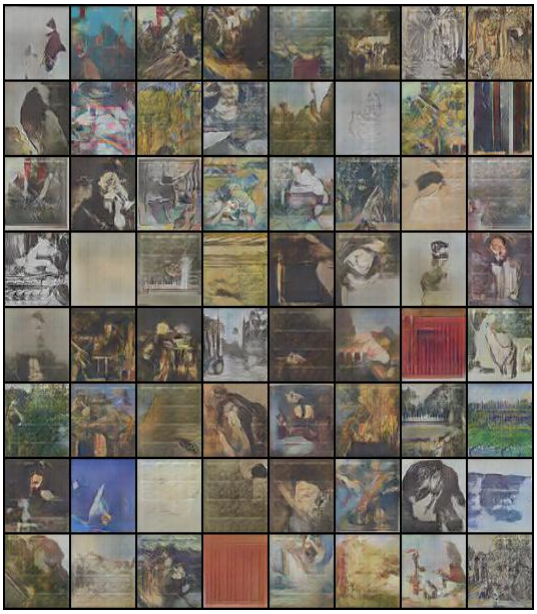
Conditional WGAN: 5th epoch of training

WGANGP: 5th epoch of training

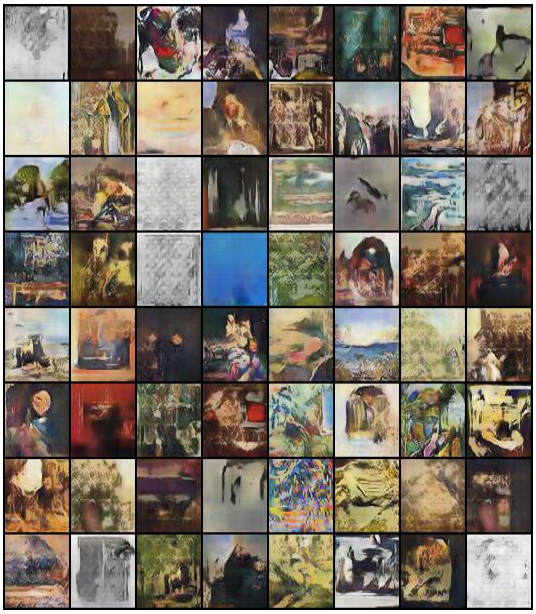
Conditional WGANGP: 5th epoch of training



IWGAN: 5th epoch of training



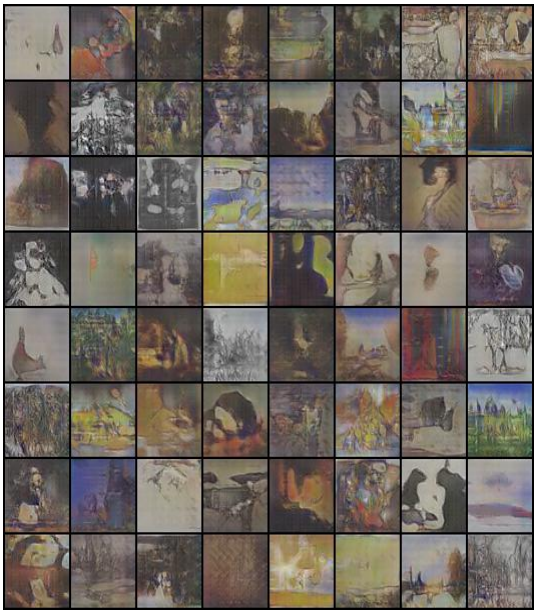
Flipped labels in D



Label smoothing

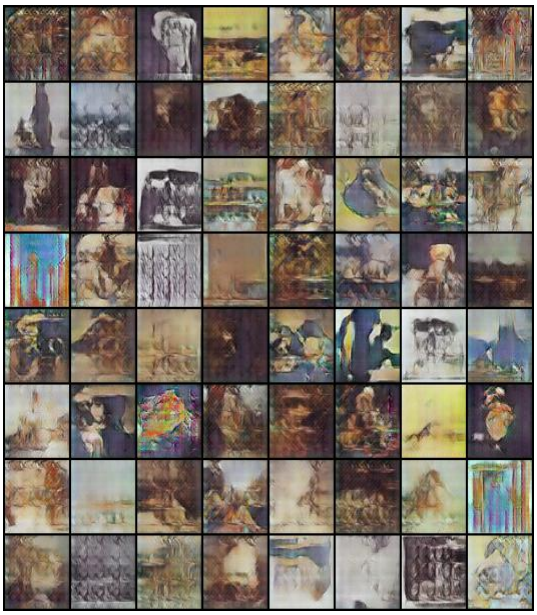
`LReLU` in G

FL-DCGAN

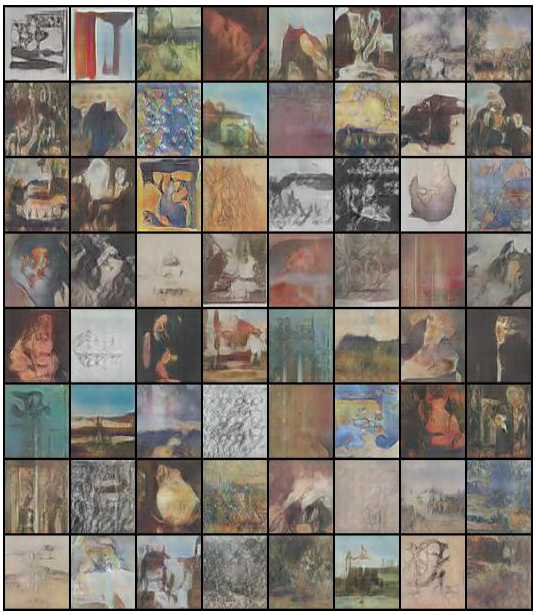


Flipped labels in D

Label smoothing



`LReLU` in G

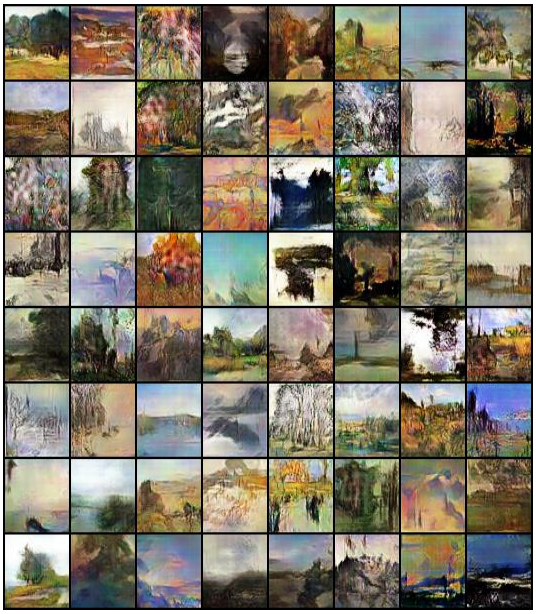


FL-DCGAN

Genre Data sets

DCGAN Landscapes

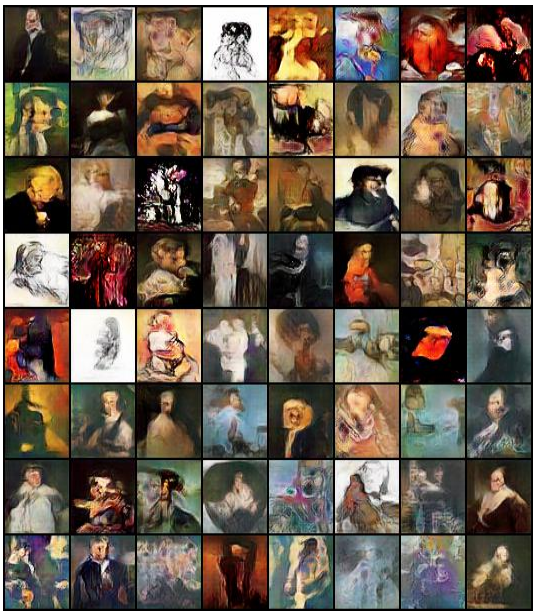
WGANGP Landscapes



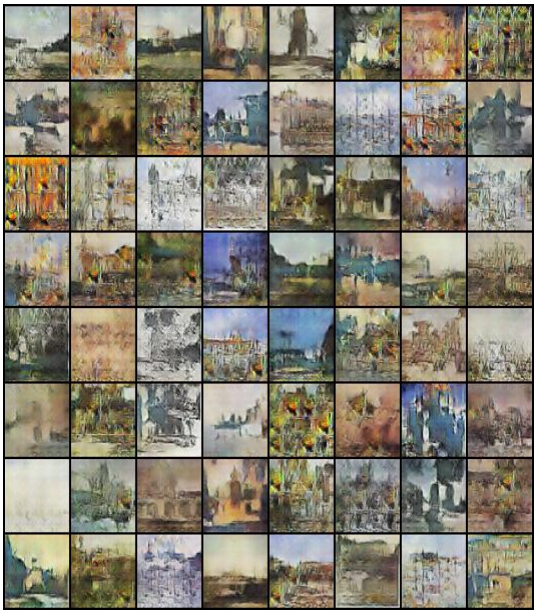
CWGANGP Landscapes

DCGAN Portraits

WGANGP Portraits



CWGANGP Portraits



DCGAN Cityscapes

WGANGP Cityscapes

HEART Pilot Study Survey

HEART - The Holistic Evaluation of Art

This the pilot run of HEART, a new framework for evaluating art (specifically visual art, both human- and computer-generated). You will be asked a few questions about how art images make you feel, and what you think of them. Please answer all of the questions.

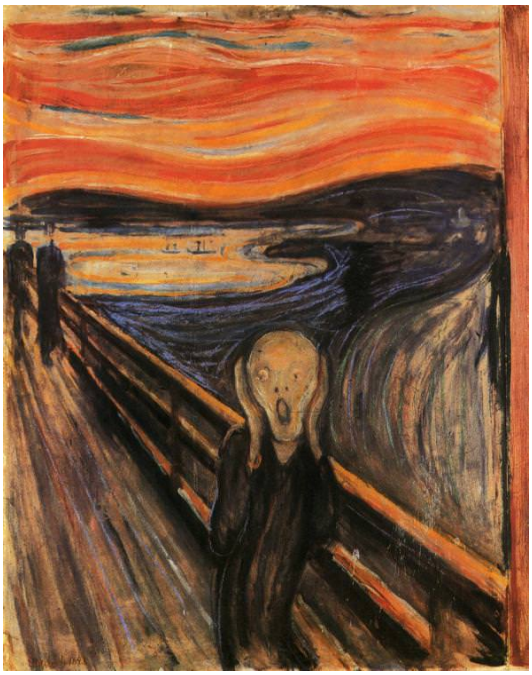
Thank you very much for participating in this research. Your responses will be kept confidential and anonymous.

* Required

1. **Email address** *

Section 1.1: Affect

The Scream - Edvard Munch



2. **The above image makes me feel:** *
Mark only one oval per row.

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
Unease					
Anxiety					
Uncertainty					
Disquiet					
Sadness					
Despair					
Gloom					
Loneliness					
Excitement					
Enthusiasm					
Thrill					
Happiness					
Joy					
Gladness					
Serenity					

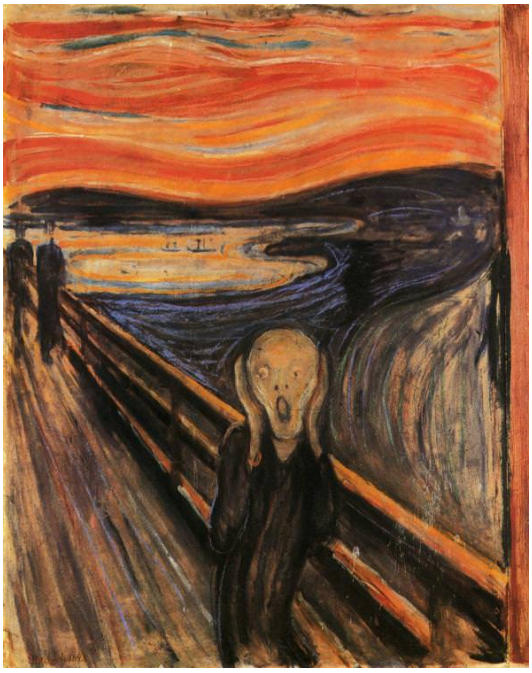
Collection of Paintings

3. **The above image makes me feel:** *
Mark only one oval per row.

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
Unease					
Anxiety					
Uncertainty					
Disquiet					
Sadness					
Despair					
Gloom					
Loneliness					
Excitement					
Enthusiasm					
Thrill					
Happiness					
Joy					
Gladness					
Serenity					

Section 1.2: Cognition

The Scream - Edvard Munch



4. **I think the above image is/shows:** *
Mark only one oval per row.

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
Interesting					
Arouses curiosity					
Fascinating					
Intellectually stimulating					
Aesthetically appealing					
Appealing to the senses					
Original					
Distinct					
Creative					
Inventive					
Of excellent workmanship					
Well-crafted					
Skillfully-made					
Compelling					
Shows intent					

Collection of Paintings

5. **I think the above image is/shows:** *
Mark only one oval per row.

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
Interesting					
Arouses curiosity					
Fascinating					
Intellectually stimulating					
Aesthetically appealing					
Appealing to the senses					
Original					
Distinct					
Creative					
Inventive					
Of excellent workmanship					
Well-crafted					
Skillfully-made					
Compelling					
Shows intent					

Section 2:

For each collage, answer the questions that follow it.

Collage 1

6. **There is blurriness/strange artifacts (e.g. speckled dots, weird lines) in the artworks of the collage.** *
Mark only one oval.

	1	2	3	4	5	
Strongly Disagree	◯	◯	◯	◯	◯	Strongly Agree

7. **The collage of artworks is diverse.** *
Mark only one oval.

	1	2	3	4	5	
Strongly Disagree	◯	◯	◯	◯	◯	Strongly Agree

8. **I can see structure in the artworks of the collage. (E.g. portraits have facial shapes)** *
Mark only one oval.

	1	2	3	4	5	
Strongly Disagree	◯	◯	◯	◯	◯	Strongly Agree

9. **The artworks in the collage are hallucinatory.** *
Mark only one oval.

	1	2	3	4	5	
Strongly Disagree	◯	◯	◯	◯	◯	Strongly Agree

10. **My overall judgment of the collage:** *
Mark only one oval.

	1	2	3	4	5	
Extremely Poor	◯	◯	◯	◯	◯	Extremely Good

Collage 2

11. **There is blurriness/strange artifacts (e.g. speckled dots, weird lines) in the artworks of the collage.** *
Mark only one oval.

	1	2	3	4	5	
Strongly Disagree	◯	◯	◯	◯	◯	Strongly Agree

12. **The collage of artworks is diverse.** *
Mark only one oval.

	1	2	3	4	5	
Strongly Disagree	◯	◯	◯	◯	◯	Strongly Agree

13. **I can see structure in the artworks of the collage. (E.g. portraits have facial shapes)** *
Mark only one oval.

	1	2	3	4	5	
Strongly Disagree	◯	◯	◯	◯	◯	Strongly Agree

14. **The artworks in the collage are hallucinatory.** *
Mark only one oval.

	1	2	3	4	5	
Strongly Disagree	◯	◯	◯	◯	◯	Strongly Agree

15. **My overall judgment of the collage:** *
Mark only one oval.

	1	2	3	4	5	
Extremely Poor	◯	◯	◯	◯	◯	Extremely Good

Section 3:

16. **I feel the above image is:** *
Mark only one oval.

- ◯ Human-generated
- ◯ Computer-generated

17. **I feel the above image is:** *
Mark only one oval.

- ◯ Human-generated
- ◯ Computer-generated

A copy of your responses will be emailed to the address you provided

Powered by
Google Forms

Bibliography

[Abadi et al., 2015] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

[Achlioptas et al., 2017] Achlioptas, P., Diamanti, O., Mitliagkas, I., and Guibas, L. J. (2017). Representation learning and adversarial generation of 3d point clouds. *CoRR*, abs/1707.02392.

[Antipov et al., 2017] Antipov, G., Baccouche, M., and Dugelay, J. (2017). Face aging with conditional generative adversarial networks. *CoRR*, abs/1702.01983.

[Arjovsky et al., 2017] Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein gan. *arXiv preprint arXiv:1701.07875*.

[Bang and Shim, 2018] Bang, D. and Shim, H. (2018). Improved training of generative adversarial networks using representative features. *arXiv preprint arXiv:1801.09195*.

[Barratt and Sharma, 2018] Barratt, S. and Sharma, R. (2018). A note on the inception score. *arXiv preprint arXiv:1801.01973*.

[Barsoum et al., 2017] Barsoum, E., Kender, J., and Liu, Z. (2017). HP-GAN: probabilistic 3d human motion prediction via GAN. *CoRR*, abs/1711.09561.

[Bellemare et al., 2017] Bellemare, M. G., Danihelka, I., Dabney, W., Mohamed, S., Lakshminarayanan, B., Hoyer, S., and Munos, R. (2017). The cramer distance as a solution to biased wasserstein gradients. *arXiv preprint arXiv:1705.10743*.

[Bergmann et al., 2017] Bergmann, U., Jetchev, N., and Vollgraf, R. (2017). Learning texture manifolds with the periodic spatial GAN. *CoRR*, abs/1705.06566.

[Berthelot et al., 2017] Berthelot, D., Schumm, T., and Metz, L. (2017). BEGAN: boundary equilibrium generative adversarial networks. *CoRR*, abs/1703.10717.

[Boden, 2009] Boden, M. A. (2009). Computer models of creativity. *AI Magazine*, 30(3):23.

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

[Brock et al., 2016] Brock, A., Lim, T., Ritchie, J. M., and Weston, N. (2016). Neural photo editing with introspective adversarial networks. *CoRR*, abs/1609.07093.

[Cao et al., 2017] Cao, G., Yang, Y., Lei, J., Jin, C., Liu, Y., and Song, M. (2017). Tripletgan: Training generative model with triplet loss. *CoRR*, abs/1711.05084.

[Cardoso et al., 2009] Cardoso, A., Veale, T., and Wiggins, G. A. (2009). Converging on the divergent: The history (and future) of the international joint workshops in computational creativity. *AI Magazine*, 30(3):15.

[Chavdarova and Fleuret, 2017] Chavdarova, T. and Fleuret, F. (2017). Sgan: An alternative training of generative adversarial networks. *arXiv preprint arXiv:1712.02330*.

[Che et al., 2016] Che, T., Li, Y., Jacob, A. P., Bengio, Y., and Li, W. (2016). Mode regularized generative adversarial networks. *CoRR*, abs/1612.02136.

[Chen et al., 2017] Chen, Z., Wu, C., Lu, Y., Lerch, A., and Lu, C. (2017). Learning to fuse music genres with generative adversarial dual learning. *CoRR*, abs/1712.01456.

[Chintala et al., 2016] Chintala, S., Denton, E., Arjovsky, M., and Mathieu, M. (2016). How to train a gan? tips and tricks to make gans work.

[Choi et al., 2017] Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F., and Sun, J. (2017). Generating multi-label discrete electronic health records using generative adversarial networks. *CoRR*, abs/1703.06490.

[Colton, 2008] Colton, S. (2008). Creativity versus the perception of creativity in computational systems.

[Colton et al.,] Colton, S., Charnley, J. W., and Pease, A. Computational creativity theory: The face and idea descriptive models.

[Colton et al., 2009] Colton, S., de Mántaras, R. L., and Stock, O. (2009). Computational creativity: Coming of age. *AI Magazine*, 30(3):11.

[Colton and Wiggins, 2012] Colton, S. and Wiggins, G. A. (2012). Computational creativity: The final frontier? In *Proceedings of the 20th European conference on artificial intelligence*, pages 21–26. IOS Press.

[Creswell et al., 2017] Creswell, A., Bharath, A. A., and Sengupta, B. (2017). Conditional autoencoders with adversarial information factorization. *CoRR*, abs/1711.05175.

[Curto et al., 2017] Curto, J. D., Zarza, I. C., Torre, F. D. L., King, I., and Lyu, M. R. (2017). High-resolution deep convolutional generative adversarial networks. *CoRR*, abs/1711.06491.

[Deng et al., 2009] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

[Denton et al., 2016] Denton, E., Gross, S., and Fergus, R. (2016). Semi-supervised learning with context-conditional generative adversarial networks. *arXiv preprint arXiv:1611.06430*.

[Di et al., 2017] Di, X., Sindagi, V. A., and Patel, V. M. (2017). GP-GAN: gender preserving GAN for synthesizing faces from landmarks. *CoRR*, abs/1710.00962.

[Dickson, 2018] Dickson, B. (2018). What is gan, the ai technique that makes computers creative?

[Dolhansky and Canton-Ferrer, 2017] Dolhansky, B. and Canton-Ferrer, C. (2017). Eye in-painting with exemplar generative adversarial networks. *CoRR*, abs/1712.03999.

[Donahue and McAuley, 2017] Donahue, C. and McAuley, J. (2017). Disentangled representations of style and content for visual art with generative adversarial networks.

[Donahue et al., 2018] Donahue, C., McAuley, J., and Puckette, M. (2018). Synthesizing audio with generative adversarial networks. *CoRR*, abs/1802.04208.

[Dong et al., 2017] Dong, H.-W., Hsiao, W.-Y., Yang, L.-C., and Yang, Y.-H. (2017). Musegan: Symbolic-domain music generation and accompaniment with multi-track sequential generative adversarial networks. *arXiv preprint arXiv:1709.06298*.

[Durugkar et al., 2016] Durugkar, I. P., Gemp, I., and Mahadevan, S. (2016). Generative multi-adversarial networks. *CoRR*, abs/1611.01673.

[Eghbal-zadeh and Widmer, 2017] Eghbal-zadeh, H. and Widmer, G. (2017). Probabilistic generative adversarial networks. *CoRR*, abs/1708.01886.

[Eiben et al.,] Eiben, A. E., Smith, J. E., et al. *Introduction to evolutionary computing*, volume 53. Springer.

[Elgammal et al., 2017] Elgammal, A., Liu, B., Elhoseiny, M., and Mazzone, M. (2017). Can: Creative adversarial networks, generating" art" by learning about styles and deviating from style norms. *arXiv preprint arXiv:1706.07068*.

[Elgammal and Saleh, 2015] Elgammal, A. and Saleh, B. (2015). Quantifying creativity in art networks. *arXiv preprint arXiv:1506.00711*.

[Fabbri et al., 2018] Fabbri, C., Islam, M. J., and Sattar, J. (2018). Enhancing underwater imagery using generative adversarial networks. *CoRR*, abs/1801.04011.

[Gervás, 2009] Gervás, P. (2009). Computational approaches to storytelling and creativity. *AI Magazine*, 30(3):49.

[Goodfellow et al., 2014] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

[Goodfellow, 2017] Goodfellow, I. J. (2017). NIPS 2016 tutorial: Generative adversarial networks. *CoRR*, abs/1701.00160.

[Gormley, 2017] Gormley, M. (2017). Machine learning.

[Grinblat et al., 2017] Grinblat, G. L., Uzal, L. C., and Granitto, P. M. (2017). Class-splitting generative adversarial networks. *arXiv preprint arXiv:1709.07359*.

[Grnarova et al., 2017] Grnarova, P., Levy, K. Y., Lucchi, A., Hofmann, T., and Krause, A. (2017). An online learning approach to generative adversarial networks. *CoRR*, abs/1706.03269.

[Gu et al., 2017] Gu, G., Kim, S. T., Kim, K. H., Baddar, W. J., and Ro, Y. M. (2017). Differential generative adversarial networks: Synthesizing non-linear facial variations with limited number of training data. *CoRR*, abs/1711.10267.

[Gulrajani et al., 2017] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. (2017). Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5769–5779.

[Hagtvedt et al., 2008] Hagtvedt, H., Patrick, V. M., and Hagtvedt, R. (2008). The perception and evaluation of visual art. *Empirical studies of the arts*, 26(2):197–218.

[Heusel et al., 2017] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Klambauer, G., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a nash equilibrium. *CoRR*, abs/1706.08500.

[Hindupur, 2017] Hindupur, A. (2017). The gan zoo. Accessed: 2018-01-23.

[Hjelm et al., 2017] Hjelm, R. D., Jacob, A. P., Che, T., Trischler, A., Cho, K., and Bengio, Y. (2017). Boundary-seeking generative adversarial networks. *arXiv preprint arXiv:1702.08431*.

[Hoang et al., 2017] Hoang, Q., Nguyen, T. D., Le, T., and Phung, D. Q. (2017). Multi-generator generative adversarial nets. *CoRR*, abs/1708.02556.

[Hochreiter, 1998] Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116.

[Hofstadter, 2002] Hofstadter, D. (2002). Staring emmy straight in the eye—and doing my best not to flinch. *Creativity, Cognition, and Knowledge: An Interaction*, page 67.

[Hou et al., 2017] Hou, M., Zhao, Q., Li, C., and Chaib-draa, B. (2017). A generative adversarial framework for positive-unlabeled classification. *CoRR*, abs/1711.08054.

[Huang et al., 2016] Huang, X., Li, Y., Poursaeed, O., Hopcroft, J., and Belongie, S. (2016). Stacked generative adversarial networks. *arXiv preprint arXiv:1612.04357*.

[Im et al., 2016] Im, D. J., Kim, C. D., Jiang, H., and Memisevic, R. (2016). Generating images with recurrent adversarial networks. *CoRR*, abs/1602.05110.

[Jaiswal et al., 2018] Jaiswal, A., AbdAlmageed, W., and Natarajan, P. (2018). Capsulegan: Generative adversarial capsule network. *arXiv preprint arXiv:1802.06167*.

[Johnson et al., 2016a] Johnson, J., Alahi, A., and Fei-Fei, L. (2016a). Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer.

[Johnson et al., 2016b] Johnson, J., Alahi, A., and Li, F. (2016b). Perceptual losses for real-time style transfer and super-resolution. *CoRR*, abs/1603.08155.

[Johnson and Zhang, 2018] Johnson, R. and Zhang, T. (2018). Composite functional gradient learning of generative adversarial models. *arXiv preprint arXiv:1801.06309*.

[Jones, 2017] Jones, K. (2017). Gangogh: Creating art with gans. `https://towardsdatascience.com/gangogh-creating-art-with-gans-8d087d8f74a1`. Accessed: 2018-01-23.

[Juefei-Xu et al., 2017] Juefei-Xu, F., Boddeti, V. N., and Savvides, M. (2017). Gang of gans: Generative adversarial networks with maximum margin ranking. *CoRR*, abs/1704.04865.

[Karras et al., 2017] Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196.

[Kim et al., 2018] Kim, Y., Kim, M., and Kim, G. (2018). Memorization precedes generation: Learning unsupervised gans with memory networks. *CoRR*, abs/1803.01500.

[Kingma and Ba, 2014] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[Kodali et al., 2017] Kodali, N., Abernethy, J. D., Hays, J., and Kira, Z. (2017). How to train your DRAGAN. *CoRR*, abs/1705.07215.

[Komer et al., 2019] Komer, B., Bergstra, J., and Eliasmith, C. (2019). Hyperopt-sklearn. In *Automated Machine Learning*, pages 97–111. Springer.

[Krizhevsky, 2009] Krizhevsky, A. (2009). Learning multiple layers of features from tiny images.

[Kupyn et al., 2017] Kupyn, O., Budzan, V., Mykhailych, M., Mishkin, D., and Matas, J. (2017). Deblurgan: Blind motion deblurring using conditional adversarial networks. *CoRR*, abs/1711.07064.

[Larsen et al., 2015] Larsen, A. B. L., Sønderby, S. K., and Winther, O. (2015). Autoencoding beyond pixels using a learned similarity metric. *CoRR*, abs/1512.09300.

[Le et al., 2017] Le, T., Nguyen, T. D., and Phung, D. Q. (2017). KGAN: how to break the minimax game in GAN. *CoRR*, abs/1711.01744.

[LeCun et al., 1998] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

[Ledig et al., 2016] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al. (2016). Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802*.

[Li et al., 2017a] Li, C., Xu, K., Zhu, J., and Zhang, B. (2017a). Triple generative adversarial nets. *CoRR*, abs/1703.02291.

[Li et al., 2017b] Li, J., Skinner, K. A., Eustice, R. M., and Johnson-Roberson, M. (2017b). Watergan: Unsupervised generative network to enable real-time color correction of monocular underwater images. *CoRR*, abs/1702.07392.

[Lin, 1991] Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

[Lin, 2017] Lin, M. (2017). Softmax GAN. *CoRR*, abs/1704.06191.

[Liu et al., 2015] Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

[Lucic et al., 2017] Lucic, M., Kurach, K., Michalski, M., Gelly, S., and Bousquet, O. (2017). Are gans created equal? a large-scale study. *arXiv preprint arXiv:1711.10337*.

[Ma et al., 2017] Ma, D., Liu, B., Kang, Z., Zhu, J., and Xu, Z. (2017). Two birds with one stone: Iteratively learn facial attributes with gans. *CoRR*, abs/1711.06078.

[Mao et al., 2016] Mao, X., Li, Q., Xie, H., Lau, R. Y. K., and Wang, Z. (2016). Multi-class generative adversarial networks with the L2 loss function. *CoRR*, abs/1611.04076.

[Mariani et al., 2018] Mariani, G., Scheidegger, F., Istrate, R., Bekas, C., and Malossi, A. C. I. (2018). BAGAN: data augmentation with balancing GAN. *CoRR*, abs/1803.09655.

[Metz et al., 2016] Metz, L., Poole, B., Pfau, D., and Sohl-Dickstein, J. (2016). Unrolled generative adversarial networks. *CoRR*, abs/1611.02163.

[Mirza and Osindero, 2014] Mirza, M. and Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

[Mirzaei et al., 2017] Mirzaei, A., Srivastava, N., Lee, K., and Xu, S. (2017). Generative adversarial networks, and applications.

[Miyato et al., 2018] Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*.

[Mogren, 2016] Mogren, O. (2016). C-rnn-gan: Continuous recurrent neural networks with adversarial training. *arXiv preprint arXiv:1611.09904*.

[Mroueh and Sercu, 2017] Mroueh, Y. and Sercu, T. (2017). Fisher GAN. *CoRR*, abs/1705.09675.

[Netzer et al., 2011] Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. (2011). Reading digits in natural images with unsupervised feature learning.

[Neyshabur et al., 2017] Neyshabur, B., Bhojanapalli, S., and Chakrabarti, A. (2017). Stabilizing GAN training with multiple random projections. *CoRR*, abs/1705.07831.

[Nowozin et al., 2016] Nowozin, S., Cseke, B., and Tomioka, R. (2016). f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, pages 271–279.

[Odena et al., 2016] Odena, A., Olah, C., and Shlens, J. (2016). Conditional image synthesis with auxiliary classifier gans. *arXiv preprint arXiv:1610.09585*.

[Oliehoek et al., 2017] Oliehoek, F. A., Savani, R., Gallego-Posada, J., van der Pol, E., de Jong, E. D., and Groß, R. (2017). Gangs: Generative adversarial network games. *arXiv preprint arXiv:1712.00679*.

[Olut et al., 2018] Olut, S., Sahin, Y. H., Demir, U., and Ünal, G. B. (2018). Generative adversarial training for MRA image synthesis using multi-contrast MRI. *CoRR*, abs/1804.04366.

[Pardo, 2007] Pardo, B. (2007). Machine learning: Topic 15: Reinforcement learning.

[Paszke et al., 2019] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., dAlché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

[Qi, 2017] Qi, G. (2017). Loss-sensitive generative adversarial networks on lipschitz densities. *CoRR*, abs/1701.06264.

[Radford et al., 2015] Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

[Ren et al., 2017] Ren, H., Chen, D., and Wang, Y. (2017). RAN4IQA: restorative adversarial nets for no-reference image quality assessment. *CoRR*, abs/1712.05444.

[Rosca et al., 2017] Rosca, M., Lakshminarayanan, B., Warde-Farley, D., and Mohamed, S. (2017). Variational approaches for auto-encoding generative adversarial networks. *arXiv preprint arXiv:1706.04987*.

[Russo et al., 2017] Russo, P., Carlucci, F. M., Tommasi, T., and Caputo, B. (2017). From source to target and back: symmetric bi-directional adaptive GAN. *CoRR*, abs/1705.08824.

[Salimans et al., 2016] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242.

[Salimans et al., 2018] Salimans, T., Zhang, H., Radford, A., and Metaxas, D. N. (2018). Improving gans using optimal transport. *CoRR*, abs/1803.05573.

[Shaikh, 2017] Shaikh, F. (2017). Introductory guide to generative adversarial networks (gans) and their promise!

[Shang et al., 2017] Shang, W., Sohn, K., Akata, Z., and Tian, Y. (2017). Channel-recurrent variational autoencoders. *CoRR*, abs/1706.03729.

[Shrivastava et al., 2016] Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., and Webb, R. (2016). Learning from simulated and unsupervised images through adversarial training. *CoRR*, abs/1612.07828.

[Spampinato et al., 2018] Spampinato, C., Palazzo, S., D'Oro, S., Murabito, F., Giordano, D., and Shah, M. (2018). VOS-GAN: adversarial learning of visual-temporal dynamics for unsupervised dense prediction in videos. *CoRR*, abs/1803.09092.

[Sun et al., 2017] Sun, Z., Ozay, M., and Okatani, T. (2017). Linear discriminant generative adversarial networks. *arXiv preprint arXiv:1707.07831*.

[Tan et al., 2017a] Tan, W. R., Chan, C. S., Aguirre, H., and Tanaka, K. (2017a). Artgan: Artwork synthesis with conditional categorial gans. *arXiv preprint arXiv:1702.03410*.

[Tan et al., 2017b] Tan, W. R., Chan, C. S., Aguirre, H., and Tanaka, K. (2017b). Learning a generative adversarial network for high resolution artwork synthesis. *arXiv preprint arXiv:1708.09533*.

[Tran et al., 2018] Tran, N., Bui, T., and Cheung, N. (2018). Generative adversarial autoencoder networks. *CoRR*, abs/1803.08887.

[Tripathy et al., 2017] Tripathy, A., Wang, Y., and Ishwar, P. (2017). Privacy-preserving adversarial networks. *CoRR*, abs/1712.07008.

[Ulyanov et al., 2017] Ulyanov, D., Vedaldi, A., and Lempitsky, V. S. (2017). Adversarial generator-encoder networks. *CoRR*, abs/1704.02304.

[van Amersfoort et al., 2017] van Amersfoort, J. R., Shi, W., Acosta, A., Massa, F., Totz, J., Wang, Z., and Caballero, J. (2017). Frame interpolation with multi-scale deep loss functions and generative adversarial networks. *CoRR*, abs/1711.06045.

[Wang et al., 2017a] Wang, C., Xu, C., Wang, C., and Tao, D. (2017a). Perceptual adversarial networks for image-to-image transformation. *CoRR*, abs/1706.09138.

[Wang and Liu, 2016] Wang, D. and Liu, Q. (2016). Learning to draw samples: With application to amortized mle for generative adversarial learning. *arXiv preprint arXiv:1611.01722*.

[Wang et al., 2017b] Wang, G., Chen, Y., and Chen, Y. (2017b). Chinese painting generation using generative adversarial networks. `http://cs231n.stanford.edu/reports/2017/pdfs/311.pdf`. Accessed: 2018-01-23.

[Wang et al., 2017c] Wang, L., Sindagi, V. A., and Patel, V. M. (2017c). High-quality facial photo-sketch synthesis using multi-adversarial networks. *CoRR*, abs/1710.10182.

[Wang et al., 2017d] Wang, T., Liu, M., Zhu, J., Tao, A., Kautz, J., and Catanzaro, B. (2017d). High-resolution image synthesis and semantic manipulation with conditional gans. *CoRR*, abs/1711.11585.

[Wang and Gupta, 2016] Wang, X. and Gupta, A. (2016). Generative image modeling using style and structure adversarial networks. In *European Conference on Computer Vision*, pages 318–335. Springer.

[Wang et al., 2018] Wang, Y., Perazzi, F., McWilliams, B., Sorkine-Hornung, A., Sorkine-Hornung, O., and Schroers, C. (2018). A fully progressive approach to single-image super-resolution. *CoRR*, abs/1804.02900.

[Wilber et al., 2017] Wilber, M. J., Fang, C., Jin, H., Hertzmann, A., Collomosse, J., and Belongie, S. J. (2017). Bam! the behance artistic media dataset for recognition beyond photography. *CoRR*, abs/1704.08614.

[Wu et al., 2017a] Wu, B., Duan, H., Liu, Z., and Sun, G. (2017a). SRPGAN: perceptual generative adversarial network for single image super resolution. *CoRR*, abs/1712.05927.

[Wu et al., 2017b] Wu, H., Zheng, S., Zhang, J., and Huang, K. (2017b). Gp-gan: Towards realistic high-resolution image blending. *arXiv preprint arXiv:1703.07195*.

[Wu et al., 2017c] Wu, J., Huang, Z., Thoma, J., and Gool, L. V. (2017c). Energy-relaxed wasserstein gans(energywgan): Towards more stable and high resolution image generation. *CoRR*, abs/1712.01026.

[Xian et al., 2017] Xian, Y., Lorenz, T., Schiele, B., and Akata, Z. (2017). Feature generating networks for zero-shot learning. *CoRR*, abs/1712.00981.

[Xie et al., 2018] Xie, Y., Franz, E., Chu, M., and Thuerey, N. (2018). tempogan: A temporally coherent, volumetric GAN for super-resolution fluid flow. *CoRR*, abs/1801.09710.

[Yi et al., 2017] Yi, Z., Zhang, H., Tan, P., and Gong, M. (2017). Dualgan: Unsupervised dual learning for image-to-image translation. *CoRR*, abs/1704.02510.

[Yin et al., 2017] Yin, X., Yu, X., Sohn, K., Liu, X., and Chandraker, M. (2017). Towards large-pose face frontalization in the wild. *CoRR*, abs/1704.06244.

[Zhang et al., 2017] Zhang, C., Ouyang, X., and Patras, P. (2017). Zipnet-gan: Inferring fine-grained mobile traffic patterns via a generative adversarial neural network. *CoRR*, abs/1711.02413.

[Zhao et al., 2018] Zhao, B., Chang, B., Jie, Z., and Sigal, L. (2018). Modular generative adversarial networks. *CoRR*, abs/1804.03343.

[Zhao et al., 2016] Zhao, J. J., Mathieu, M., and LeCun, Y. (2016). Energy-based generative adversarial network. *CoRR*, abs/1609.03126.

[Zhou et al., 2017] Zhou, Z., Rong, S., Cai, H., Zhang, W., Yu, Y., and Wang, J. (2017). Generative adversarial nets with labeled data by activation maximization. *CoRR*, abs/1703.02000.

[Zhu et al., 2017a] Zhu, J., Park, T., Isola, P., and Efros, A. A. (2017a). Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, abs/1703.10593.

[Zhu et al., 2017b] Zhu, J., Zhang, R., Pathak, D., Darrell, T., Efros, A. A., Wang, O., and Shechtman, E. (2017b). Toward multimodal image-to-image translation. *CoRR*, abs/1711.11586.