Evaluating GANs for Medical Imagery Synthesis: A Literature Review

Shaylin Chetty
CHTSHA042@myuct.ac.za
University of Cape Town
Cape Town, South Africa

ABSTRACT

GANs have seen explosive growth since their inception in 2014. To date, numerous variations have been proposed that either address some of the shortcomings of the original GAN or improve on it for specialized applications. This literature review will evaluate the use of Vanilla GAN, WGAN and StyleGAN to improve the size and quality of limited datasets and will evaluate the impact of these generative methods on classification tasks. Overall, a review of related literature has shown promise for the use of GANs in pathology data generation and subsequence classification tasks. Albeit, further research is required on objective evaluation metrics.

Keywords: Generative Adversarial Networks, WGAN, WGAN-GP, StyleGAN, loss functions, self-supervised learning.

1 INTRODUCTION

Computer Vision is the study of using artificial intelligence and deep learning technologies to extract meaning from a scene (any visual input) [37], allowing a computer to interpret the world in the same manner as humans. Some applications include image segmentation, classification and object detection. Unfortunately, however, computer vision tasks in recent research are largely focused on supervised learning [18]. By nature, supervised learning algorithms require large volumes of labelled data, which, in certain domains, may not be easily accessible. Active research into this area has led to the availability of large-scale, accurate, diverse and structured datasets such as ImageNet [9] and the LSUN-Bedrooms dataset [56].

However, these are more general-purpose datasets that may not be suitable for some niche application domains that do not enjoy the same data availability. A lack of available data is usually due to domain constraints on the type of data that can be used. For example, some domains may require specific subject types, image quality and location while other domains may have privacy, intellectual property and other data protection issues that do not allow researchers to share and disseminate this data [55]. This hinders the progress made in deep learning applications in these domains. Capturing custom data would be the most obvious approach but most applications require $10^5 \sim 10^6$ images to create high-quality models. This poses a significant data capturing, cleaning and distribution challenge [24, 4]. Going ahead and using limited (small) datasets is not sufficient enough for robust models as these datasets are often not representative of the true population distribution of data [11], contain minority classes which can be dropped (this small imbalance has been shown to weaken classifier performance [33]), and may lead to model overfitting [24].

One proposed technique for addressing this issue is Data Augmentation. Data Augmentation artificially enhances the size of training data [11, 59] by performing geometric [24, 43] and photometric transformations on data [43, 25]. However, these techniques only make the trained model invariant to particular conditions and do not enhance the diversity of the data thereby still making the data unrepresentative of the population. That is, it increases the given sample spaces but doesn't explore the true sample space [22]. Additionally, the effectiveness of a data augmentation technique is highly correlated to the types of images it is applied to [4]. For example, XRays of the neck will not benefit from rotational transformations as these can only be analysed in one orientation.

To address this issue, Goodfellow et al. proposed a novel method for synthetic image generation called Generative Adversarial Networks (GANs) [14] that allows us to create entirely new images that confidently follow the same distribution as the original data, enhancing the size and quality of the training dataset leading to more robust models.

This paper reviews some of the current literature on Generative Adversarial Networks (GANs), the different GAN variants, and the applications of GANs in pathology and self-supervised frameworks.

2 SYNTHETIC IMAGE GENERATION USING GANS

2.1 Image Synthesis

Image synthesis is the process of artificially creating training data based on the statistical structure of the original data. GANs are the most robust and generic architectures for image synthesis and have been applied extensively to medicine [55]. Image synthesis can be broken down into three main categories: first, *unconditional synthesis* whereby the Generator of the GAN does not receive any auxiliary information and generates images relying solely on a random noise input and its probability approximation [55]. Unconditional synthesis of medical imagery typically uses StyleGAN [49] and WGAN [55] due to their increased stability over Vanilla GAN. Second, *Cross-modality synthesis* which involves "translating" images from one method of treatment (domain) to another. For example, Wolterink et al. trained a GAN to transform deep magnetic resonance images (MRI) into computed tomography (CT) scans [48]. Generally, CycleGANs are used. Lastly, *Conditional image synthesis*, which uses conditional GANs (cGAN) that pass information to the Generator which is used when creating data. For example, Mok et al. used cGANs to generate images for brain tumour segmentation by passing segmentation maps and generated brain MRIs to the Generator [34].

2.2 Generative Adversarial Networks (Vanilla GAN)

In their paper, "Generative Adversarial Networks", Goodfellow et al. proposed an implicit generative model to enhance the size and diversity of a dataset [14]. GANs do this by approximating the distribution of the training data to produce consistent samples. It consists of two neural networks (both MultiLayer Perceptrons (MLPs) [14]), namely a Generator and a Discriminator, that are simultaneously trained via an adversarial game. The Generator tries to model the underlying statistical structure of the training data and uses this to generate synthetic samples that are then used as input to the Discriminator which will classify the data as either synthetic or real. The Generator and Discriminator are brought together through the loss function, V (see Equation 1), which the Generator tries to minimize while the Discriminator tries to maximize. $G(z)$ is the Generators output upon receiving random noise sampled from a noise distribution $z \sim p(z)$, which is normally a uniform or Gaussian distribution. The noise is mapped to data space by a differentiable function representing a multilayer perceptron, $G(z, \theta_g)$ where θ_g is a hyperparameter. is the Discriminator's probability that its input is from the real data distribution and not from the Generator's approximated distribution. Lastly, and denotes the mean likelihood over all original and synthetic data respectively [14].

$$V(D, G) = \mathbb{E}_x[logD(x)] + \mathbb{E}_z[log(1 - D(G(z)))] \quad (1)$$

The Generator will receive noise as input and model it based on its approximated distribution, P_g, which is the Generator's approximation of the real data distribution, P_r. To learn P_g, the Generator uses a parametric family of density functions $(P_\theta)_{\theta \in d}$ [1] and finds the density function that maximizes the likelihood of our data.

The Discriminator receives synthetic data and real data samples as illustrated in Figure 1. It is responsible for classifying images by estimating the probability that the data came from the original training set (therefore, a lower probability implies confidence that the data is synthetic). The original GAN used a sigmoid function [14][14], which limits the Discriminator's output to the range [0,1]. The Generator is optimized when the Discriminator's output is ½ [47] that is, the Discriminator is only 50% sure that the data is real $\implies$ 50% sure that the data is synthetic.

Thus, we are no longer sure whether the data is real or synthetic, successfully tricking the discriminator. The reader is referred to [47] for a mathematical proof.

2.3 (Vanilla) GAN Issues

Although GANs are a concrete data synthesis technique there are some challenges with training GANs that impact their stability:

2.3.1 Simultaneous Training Issues. The Generator and Discriminator are trained simultaneously, with the output of the Discriminator used to perform a gradient descent on the loss function of the Generator [47]. The aim of training a GAN is to get P_g to closely approximate P_r. To quantify the distance between both distributions, the original GAN used the *Jensen-Shannon divergence*, a divergence

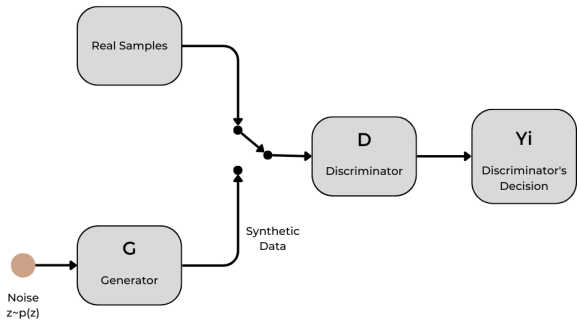

Figure 1: The architecture of Vanilla GAN

metric improving on the *Kullback-Leibler divergence*, as its loss function [14]. Due to the sigmoid-style curve of the Jensen-Shannon divergence, as the distance between the real and generated distribution increases, the gradient of the divergence approaches 0 meaning the gradient-descent algorithms don't learn much from the data [1]. Furthermore, as the distance between the distributions increases, the Discriminators' ability to classify the data correctly increases, possibly reaching a perfect Discriminator, that is $D(x) = 1, \forall x \in P_r$ and $D(x) = 0, \forall x \in P_g$. The loss function approaches 0 and we have less of a gradient to update the neural network weights between training epochs. This is referred to as the *vanishing gradient problem* which could lead to slow training and could result in the real distribution and approximated distribution diverging.

2.3.2 Mode Dropping and Mode Collapse. A robust Generator should be able to model the entire data distribution of the real data however there may be instances where the Generator may be poorly optimized and continuously map noise input to the same output [2] reducing the diversity [55] of synthetic images produced while maintaining image quality. This is referred to as *mode collapse*. Often, in complicated imagery, some hard-to-represent modes are disregarded by the Generator [2], producing realistic, yet less general results. This is referred to as *mode dropping*.

2.3.3 Leaky Augmentations. As mentioned, data augmentation is a common technique to prevent Generator overfitting. Training GANs with augmented data, however, leads to the Generator approximating the distribution of the augmented [24] data generating synthetic results that include the augmentations [58]. A common technique to prevent this is *adaptive data augmentation* [49] however for the purposes of this review, this will not be studied.

2.3.4 Instability. The aforementioned issues contribute to GAN training instability, which causes the Generator to produce non-representative samples (regardless of the visual quality of images). Proposed solutions to address these issues include using smaller learning rates, using gradient clipping [1], using a more robust loss function [1, 15], using better activation functions such as ReLU and LeakyReLU [36, 52], using batch normalization on all layers

excluding the Generator's output and the Discriminator's input layers [36] or using stridden convolutions for downsampling instead of pooling layers [36].

3 GANS IN MEDICINE

3.1 Generation

In medical literature, it is clear that there are 3 major data issues, namely data scarcity, patient privacy - which is a leading contributor to the lack of open-access databases - and a lack of annotated medical images [55]. Additionally, in pathology, due to the data collection constraints mentioned in section 1, there is often a lack of a sufficient number of positive cases of each pathology [22]. This makes supervised and unsupervised models inherently skewed by increasing the likelihood of finding patterns consistent with the data regardless of population distributions and overestimating the model's performance on the population [5]. This could also propagate the bias to downstream applications [31]. Applying GANs to medical imagery synthesis is a remedy to these issues as we are able to create imagery with similar properties to existing data. This has been done in unconditional synthesis, conditional synthesis and cross-modality synthesis [48] with most literature applying GANs to MRI and CT Scans of the brain and chest [22].

3.2 Discrimination

The discriminator of a well-trained GAN can be applied to other machine-learning tasks including image classification and segmentation [20].

Although GANs show potential in medical imagery research, the adoption of the technology is prevented by training complexity, computational costs and the lack of reliable and efficient evaluation methods [49].

4 VANILLA GAN ALTERNATIVES

Extensive research has led to many of the aforementioned stability-inducing techniques being adopted in GANs. GAN variants can be broadly categorized into three board sections:

4.0.1 Discriminator Changes. Some models have proposed stabilising the Discriminator by replacing the GAN loss function with a more suitable alternative. Examples include Wasserstein GAN (WGAN) (evaluated later in this literature review) [1, 15] and Least-Squares GAN [32].

4.0.2 Generator Changes. Vanilla GAN uses a decoder network to transform noise into a sample, $G(z)$ however, this has since been improved on via the use of variational autoencoder networks [29] or by passing auxiliary information to the generator to improve its performance (usually for image-to-image translations). For example, CycleGAN [61].

4.0.3 Architecture Changes. Recent research has considered changing the architecture of the Discriminator and Generator. Most notably, the multilayer perceptrons have been replaced with deep convolutional layers (DCGAN) [36, 28] which have proved to have more stability than Vanilla GAN. StyleGAN has changed the architecture of the Generator to take in input noise at each convolution leading to an incremental refinement of images.

4.1 Wasserstein Generative Adversarial Networks (WGANs)

WGAN uses the Wasserstein distance to improve the Vanilla GAN value function [15] improving stability. Firstly, the *Wasserstein Distance* (also known as the Earth Mover Distance) is the minimum cost of transporting data in converting one distribution to another that is. it measures the distance between any two probability distributions. WGANs replace the loss function of Vanilla GAN with the Wasserstein Distance such that we have a new loss function (see equation 2) where f_w is the Discriminator. This has been shown to improve the behaviour of the gradient compared to other distance measures [1].

$$max_{w \in W} \mathbb{E}_{x\sim,P_r}[f_w(x)] - \mathbb{E}_{z\sim p(z)}[f_w g_\theta(z)] \tag{2}$$

While in Vanilla GAN, the sigmoid activation function limited the Discriminator's values to the range [0,1], WGAN generates a range of values. These values can be considered a measurement of how real the images are [47]. To limit the weights, however, gradient clipping is applied [1] limiting the Discriminator's values to the range [-c, c] where c is a hyperparameter. The choice of the clipping hyperparameter, c, is extremely important as a value too large leads to an exploding gradient and a diverging model. Gulrajani et al. [15] proposed a gradient penalty instead of using weight clipping in WGAN-GP. The idea is to penalize the model if the gradient norm differs too much from the target value of 1. This ensures the model converges (stabilises training), prevents vanishing gradient issues and requires less hyperparameter tuning.

The most significant contribution made by WGANs is the introduction of a meaningful loss metric. The Jensen-Shannon Divergence has been shown [1] to have no significant correlation between the quality of the generated image but rather measures how well the Generator can fool the Discriminator. The WGAN-GP loss function, however, has been shown to have a significant correlation to the visual quality of the image [1] providing an objective image quality evaluation metric.

The use of WGANs is recurrent in medical literature, for example, it has been used to generate MR imagery [23] or to assist in cancer detection by correcting the imbalance in cancer gene expression data [35].

4.2 StyleGAN

The traditional GAN architecture's generator does not allow one to control how images are generated. Its results are close to the original data and follow the same underlying distribution however we have no control over changing stochastic features (such as finer details of images that could influence classification). StyleGAN [23] is the state-of-the-art proposed solution that splits key attributes and adds stochastic variation to generated images by introducing noise (from a Gaussian distribution) at each convolution within the Generator as opposed to just the Generator input as is the case of Vanilla GAN. Instead, the Generator starts with a learnt constant tensor and adjusts the image at each convolution based on the *style* applied at that convolution. Noise is inserted at each convolution to add stochastic variation to small details while keeping the main structure of the image the same. The rest of the GAN structure

remains unchanged besides the use of the W-Distance as a loss function (along with Saturating Losses).

This makes StyleGAN suitable to applications where fine control over the image generation process is needed. It can also be applied to applications that have aspects of images that are variant as we can use style mixing to enhance results and capture these smaller variations. This has proved to be extremely successful with initial tests [23] increasing the FID score of facial images by close to 20% compared to Vanilla GAN.

5 SELF-SUPERVISED TECHNIQUES

Clearly, access to (balanced) data is one of the leading issues slowing down the adoption of deep learning techniques in some domains. Research into this area has led to significant contributions to reducing the need for large training datasets. Transfer learning [35] is a good alternative but given the intrinsic variation of pathology data (The size, location, and orientation of tumours [8, 27]), this is unlikely to work and would still require annotated medical data [4]. *Self-Supervised Learning* is another research area that focuses on creating models with limited data [57]. Under this regime, we use a small amount of labelled data to apply labels on labelled data through a process known as *pseudo-labelling*.

In literature, convolutional deep residual networks (for example, ResNet) are quite common in self-supervised vision tasks [7]. Recent work has investigated the use of *transformers* [46]. These, unlike CNNs, are attentional making them usable in applications where the relations between data points are important. Traditionally applied to NLP, Dosovitskiy et al. [12] proposed Vision Transformers suitable for computer vision tasks however these are still aimed at datasets with relational features. As such, ResNets will be expanded upon in this section.

Firstly, a common debate in computer vision is the question of the effect of stacking more layers on network learning [17]. Recent evidence [40, 41] shows that deeper convolutional networks perform considerably better on the ImageNet dataset however vanishing or exploding gradients have become a problem [17]. Normalized initialization [13] has been used ensuring networks converge when stochastic gradient descent with back-propagation is used. However, as the layers converge we have a new *degradation* problem whereby the accuracy gets saturated as the network depth increases [17]. This was experimentally proven by He et al. [16], who showed that network depth reduces accuracy even if filter sizes are unchanged, furthermore, this was tested under unconstrained time complexity. ResNet [17] fixes this by introducing shortcut connections to the neural network. These shortcut connections are identity functions that recast the original mapping to $F(x) + x$ where $F(x)$ is the output of the stacked layers and x is the input. The idea is that a deeper model with identity shortcuts should have a training error at least the same as the shallower version as the identity shortcuts provide an additional path for the gradient to flow solving the vanishing gradient problem. Adding these identity shortcuts adds no extra parameters or computational complexity to the method.

In their experiment, He et al. [17] showed that normal networks do suffer from degradation for reasons other than vanishing gradients however adding shortcut connections has eliminated this issue and deepening the network did indeed increase accuracy and speed up convergence time. For example, Kolesnikov et al.) concluded that images do not degrade in quality while traversing the network [26]. ResNet-50 was introduced to speed up training time (over the original ResNet18) by having three layers (a 3x3 convolution between two 1x1 convolutions) between shortcut connections giving us a 50-layer network using projection shortcuts (projecting a sample to a higher dimension) to increase dimensions if needed and using identity shortcuts elsewhere.

To improve on the performance of ResNet, many applications have introduced contrastive learning [18, 6], with promising results. Recently, this is done using momentum contrast [18]. This uses a large, dynamic, queue-implemented dictionary with a moving-average encoder. There are two encoders (originally, ResNets), F_q and F_k with output vectors q and k, which can be images or patches. q behaves like a query and the goal is to find the corresponding key such that the query is similar to its key but dissimilar to other keys. Chen et al. [7] improved on this by abandoning the memory queue and introducing a symmetrized loss and introducing an extra prediction head.

Further alternatives have been created specifically for pathology, notably *Self-Path* [27]. Self-Path is a self-supervised CNN for domain invariant applications in pathology that takes advantage of the semantic and contextual features of pathology images. The feature-extraction processes are managed by ResNet.

6 EVALUATING THE PERFORMANCE OF GANS

Clearly, there is an opportunity to viably apply GANs to pathology data. However, the choice of GAN is crucial. This task is made increasingly difficult due to the lack of objective loss function in GANs which makes it difficult to compare models [38]. To aid in this, numerous qualitative and quantitative measures have been suggested in literature.

6.1 Manual Inspection

Manual inspection is the process of visually evaluating a generated image and is commonly used to assess the output of generative models [10, 44]. This is an expensive and cumbersome technique [3] whose results are subjective, variant and biased to models that overfit [3]. Additionally, since GAN outputs seem realistic, these methods will fail to detect mode dropping.

6.2 Qualitative GAN Evaluation

Preference Judgement uses a rank-based approach by asking reviewers to rank images based on how real they appear [49]. *Rapid Scene Categorization* is similar to the above however each image is only shown for a short period upon which reviews categorize the image as real or synthetic, often using an interface [10]. Together, these two methods can be used to rank how realistic the generated images appear exposing the flaws in poor-performing generated images [38]. The results can then be summarized using False (or True) Recognition Rates [8].

6.3 Quantitative GAN Evaluation

Vanilla GAN uses a method called the *Maximum Log-Likelihood* for training the model. Essentially, this method measures the likelihood of the original data under the approximated distribution [14]. However, this has been proven to favour trivial models and has a fairly weak relationship to sampling quality [44, 3, 21].

6.3.1 Inception Scores. . Inception Scores, proposed in [38] are the most widely used GAN evaluation method [3]. Inception Scores use a pre-trained deep-learning neuronal network called Inception Net [42], which has been trained on ImageNet [9], a hierarchical image database of over 3.2 million images originally developed for improving image search algorithms. The premise is on estimating the probability that the image belongs to each class defined in ImageNet. The probabilities are summarized to determine how close an image is to the rest of the images in the given class and scored accordingly. A higher inception score supposedly implies better-quality imagery [38]. However, Inception Scores are unable to detect overfitting [54] and have been shown to favour models that produce a diverse range of images over good quality images [60, 53].

6.3.2 FID Scores. . As an improvement, Fréchet Inception Distances were proposed [19] to improve the efficiency, robustness and discriminability power of Inception Scores [3]. It is based on the same Inception Net that Inception Scores use, however the last pooling layer of Inception Net is adjusted to capture specific features of an inputted image. Unlike Inception Scores, FID values can detect mode dropping and measure both image diversity and quality [30].

6.3.3 A Note for Medicine. While Inception Scores and FID values are both well-used metrics, their applicability to medical applications has been questioned [45]. This is because both metrics are based on Inception Net, trained on ImageNet, which contains no medical imagery [49]. However, Woodland et al. [49] proved the negative correlation between FID scores and human perceptual evaluation of medical images hence these metrics will be assumed valid for this research. Recent developments [39, 50, 51] have proposed that GAN-generated output be applied to classification tasks to assess their effectiveness. As such, the accuracy and sensitivity of classification tasks are useful in objectively describing the effectiveness of GANs.

7 DISCUSSIONS

Generative Adversarial Networks (GANs) have proven to be a powerful tool for synthetic image generation that adds to the size and diversity of small datasets and equalises imbalances in skewed data. This is in contrast to traditional data augmentation methods that make models invariant to geometric and photometric transformations, thereby increasing the sample space. GANs allow researchers to explore the true sample space by adding realistic images to existing datasets while maintaining the underlying statistical distribution of pixels within the original images. This allows us to add data to under-represented classes, helping to prevent model overfitting and the dropping of seemingly insignificant modes. This ability makes GANs extremely attractive in data-constrained environments, such as the field of medical imagery, where the number of positive cases for each pathology may be low.

The basic structure of a GAN consists of two neural networks (either MultiLayer Perceptrons or Convolutional Neural Networks) that compete in an adversarial game. One network, the Generator, tries to learn the distribution of the original data by applying its current approximated distribution to random noise input. The output is then evaluated by the second network, the Discriminator, which will try to discern real data from synthetic data. The response from the Discriminator allows the Generator to update its approximated distribution, improving its ability to generate realistic samples. In essence, the GAN takes original data and creates unique copies of the data with slight variations.

The baseline GAN, Vanilla GAN, has been shown to perform well and is relatively simple and cheap to implement however the drawbacks, such as mode dropping, vanishing gradients and simultaneous training issues do threaten the stability of training and could produce images that do not capture the true representation of the data albeit realistic. Compounding this issue, Vanilla GAN contains no objective loss function meaning there is no way to objectively measure how well the Generator captures the data distribution. WGAN provides a new loss function, the Wasserstein Distance, that improves the stability of training, prevents mode collapse and provides a loss metric that correlates with sample quality. WGAN was improved by adding in gradient penalties (WGAN-GP). Some applications, though, do require GANs that produce extremely high-resolution images, even at the expense of computation time. This is where StyleGAN2 comes in. StyleGAN2 enables control over the image generation process and helps us capture small variations in images that WGAN-GP may not be able to do. These small (stochastic) variations are useful when small details could affect the class an item may have to be classified in downstream applications.

A common drawback of each of the discussed GANs is more specific to medical applications: The common quantitative metrics to evaluate a GAN, Inception Score and FID, are based on non-medical data. Regardless of how realistic the images seem, the ability of the GAN to capture the underlying structure can only be captured through human evaluation. This is a time-consuming and ineffective technique however as recent evidence [49] suggests, the FID metric is suitable for medical data however more robust and medical-oriented metrics would be needed to improve the confidence of GANs in pathology.

While GANs are useful in pathology, their uptake in the field has two major obstacles. Firstly, medical data is often more unbalanced and less annotated than other niche domains - even though GANs have the ability to generate invariant samples, their ability to do this reduces as the size of the training dataset reduces. Secondly, the computational costs and the training and implementation complexity mean that the effectiveness of the GAN is heavily dependent on the choice of GAN, the use case and the skills of the medical practitioner. GANs that are simpler to implement and that could provide high-quality outcomes without the added complexity would be highly beneficial to medical practitioners, for example, Vanilla GAN and WGAN-GP. These models could further be improved to match the quality of advanced models (for example, DCGAN [36]) through collective research into more stable activation functions,

loss functions and optimization functions to ensure the convergence of distributions whilst retaining the simplicity of existing models

8 CONCLUSION

As this literature review has discussed, GANs are powerful techniques that add to the quality of small datasets. Their ability to generate data based on the statistical distribution of training samples makes GANs the preferred technique in data-constrained environments. While Vanilla GANs offer a relatively simple and good basis for generative models, the improvements made by WGAN-GP make it a better, yet more complicated model. StyleGAN was evaluated for the purpose of generating high-resolution images with small variations, suitable in applications where these small variations could influence classification. Overall, GANs should be applied in limited data environments, such as pathology tasks, to improve the balance, size and quality of limited datasets leading to more reliable training of classifiers in downstream tasks. The use of GANs should also be broadened to include discriminatory tasks (by using the GANs Discriminator) taking advantage of reduced datasets and limited training time. Lastly, further research on medicine-inclined evaluation metrics is encouraged.

REFERENCES

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. *Wasserstein GAN*. en. arXiv:1701.07875 [cs, stat]. Dec. 2017. URL: http://arxiv.org/abs/1701.07875 (visited on 03/03/2023).

[2] Sanjeev Arora and Yi Zhang. *Do GANs actually learn the distribution? An empirical study*. arXiv:1706.08224 [cs]. June 2017. URL: http://arxiv.org/abs/1706.08224 (visited on 03/07/2023).

[3] Ali Borji. *Pros and Cons of GAN Evaluation Measures*. arXiv:1802.03446 [cs]. Oct. 2018. URL: http://arxiv.org/abs/1802.03446 (visited on 03/05/2023).

[4] Lorenzo Brigato and Luca Iocchi. "A Close Look at Deep Learning with Small Data". In: *2020 25th International Conference on Pattern Recognition (ICPR)*. ISSN: 1051-4651. Jan. 2021, pp. 2490–2497. DOI: 10.1109/ICPR48806.2021.9412492.

[5] John E. Burkhardt et al. "Recommendations for the Evaluation of Pathology Data in Nonclinical Safety Biomarker Qualification Studies". en. In: *Toxicologic Pathology* 39.7 (Dec. 2011), pp. 1129–1137. ISSN: 0192-6233, 1533-1601. DOI: 10.1177/0192623311422082. URL: http://journals.sagepub.com/doi/10.1177/0192623311422082 (visited on 03/16/2023).

[6] Ting Chen et al. *A Simple Framework for Contrastive Learning of Visual Representations*. en. arXiv:2002.05709 [cs, stat]. June 2020. URL: http://arxiv.org/abs/2002.05709 (visited on 03/19/2023).

[7] Xinlei Chen, Saining Xie, and Kaiming He. *An Empirical Study of Training Self-Supervised Vision Transformers*. en. arXiv:2104.02057 [cs]. Aug. 2021. URL: http://arxiv.org/abs/2104.02057 (visited on 03/03/2023).

[8] Maria J. M. Chuquicusma et al. *How to Fool Radiologists with Generative Adversarial Networks? A Visual Turing Test for Lung Cancer Diagnosis*. en. arXiv:1710.09762 [cs, q-bio]. Jan. 2018. URL: http://arxiv.org/abs/1710.09762 (visited on 03/20/2023).

[9] Jia Deng et al. "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. ISSN: 1063-6919. June 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[10] Emily Denton et al. *Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks*. arXiv:1506.05751 [cs]. June 2015. URL: http://arxiv.org/abs/1506.05751 (visited on 03/16/2023).

[11] Serge Dolgikh. *Analysis and Augmentation of Small Datasets with Unsupervised Machine Learning*. en. preprint. Health Informatics, Apr. 2021. DOI: 10.1101/2021.04.21.21254796. URL: http://medrxiv.org/lookup/doi/10.1101/2021.04.21.21254796 (visited on 03/15/2023).

[12] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. arXiv:2010.11929 [cs]. June 2021. URL: http://arxiv.org/abs/2010.11929 (visited on 03/05/2023).

[13] Xavier Glorot and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks". en. In: ().

[14] Ian J. Goodfellow et al. *Generative Adversarial Networks*. arXiv:1406.2661 [cs, stat]. June 2014. URL: http://arxiv.org/abs/1406.2661 (visited on 03/05/2023).

[15] Ishaan Gulrajani et al. *Improved Training of Wasserstein GANs*. arXiv:1704.00028 [cs, stat]. Dec. 2017. URL: http://arxiv.org/abs/1704.00028 (visited on 03/09/2023).

[16] Kaiming He and Jian Sun. *Convolutional Neural Networks at Constrained Time Cost*. en. arXiv:1412.1710 [cs]. Dec. 2014. URL: http://arxiv.org/abs/1412.1710 (visited on 03/20/2023).

[17] Kaiming He et al. *Deep Residual Learning for Image Recognition*. en. arXiv:1512.03385 [cs]. Dec. 2015. URL: http://arxiv.org/abs/1512.03385 (visited on 03/03/2023).

[18] Kaiming He et al. *Momentum Contrast for Unsupervised Visual Representation Learning*. en. arXiv:1911.05722 [cs]. Mar. 2020. URL: http://arxiv.org/abs/1911.05722 (visited on 03/19/2023).

[19] Martin Heusel et al. *GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium*. arXiv:1706.08500 [cs, stat]. Jan. 2018. URL: http://arxiv.org/abs/1706.08500 (visited on 03/05/2023).

[20] Bo Hu et al. "Unsupervised Learning for Cell-level Visual Representation in Histopathology Images with Generative Adversarial Networks". en. In: *IEEE Journal of Biomedical and Health Informatics* 23.3 (May 2019). arXiv:1711.11317 [cs], pp. 1316–1328. ISSN: 2168-2194, 2168-2208. DOI: 10.1109/JBHI.2018.2852639. URL: http://arxiv.org/abs/1711.11317 (visited on 03/17/2023).

[21] Ferenc Huszár. *How (not) to Train your Generative Model: Scheduled Sampling, Likelihood, Adversary?* arXiv:1511.05101 [cs, math, stat]. Nov. 2015. URL: http://arxiv.org/abs/1511.05101 (visited on 03/16/2023).

[22] Jiwoong J. Jeong et al. "Systematic Review of Generative Adversarial Networks (GANs) for Medical Image Classification and Segmentation". en. In: *Journal of Digital Imaging* 35.2 (Apr. 2022), pp. 137–152. ISSN: 0897-1889, 1618-727X. DOI: 10.1007/s10278-021-00556-w. URL: https://link.springer.com/10.1007/s10278-021-00556-w (visited on 03/20/2023).

[23] Tero Karras, Samuli Laine, and Timo Aila. *A Style-Based Generator Architecture for Generative Adversarial Networks*. arXiv:1812.04948 [cs, stat]. Mar. 2019. URL: http://arxiv.org/abs/1812.04948 (visited on 03/05/2023).

[24] Tero Karras et al. *Training Generative Adversarial Networks with Limited Data*. en. arXiv:2006.06676 [cs, stat]. Oct. 2020. URL: http://arxiv.org/abs/2006.06676 (visited on 03/03/2023).

[25] Parvinder Kaur, Baljit Singh Khehra, and Er. Bhupinder Singh Mavi. "Data Augmentation for Object Detection: A Review". en. In: *2021 IEEE International Midwest Symposium on Circuits and Systems (MWSCAS)*. Lansing, MI, USA: IEEE, Aug. 2021, pp. 537–543. ISBN: 978-1-66542-461-5. DOI: 10.1109/MWSCAS47672.2021.9531849. URL: https://ieeexplore.ieee.org/document/9531849/ (visited on 03/15/2023).

[26] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. *Revisiting Self-Supervised Visual Representation Learning*. en. arXiv:1901.09005 [cs]. Jan. 2019. URL: http://arxiv.org/abs/1901.09005 (visited on 03/23/2023).

[27] Navid Alemi Koohbanani et al. "Self-Path: Self-Supervision for Classification of Pathology Images With Limited Annotations". en. In: *IEEE Transactions on Medical Imaging* 40.10 (Oct. 2021), pp. 2845–2856. ISSN: 0278-0062, 1558-254X. DOI: 10.1109/TMI.2021.3056023. URL: https://ieeexplore.ieee.org/document/9343323/ (visited on 03/21/2023).

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Vol. 25. Curran Associates, Inc., 2012. URL: https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html (visited on 03/06/2023).

[29] Anders Boesen Lindbo Larsen et al. *Autoencoding beyond pixels using a learned similarity metric*. arXiv:1512.09300 [cs, stat]. Feb. 2016. URL: http://arxiv.org/abs/1512.09300 (visited on 03/09/2023).

[30] Mario Lucic et al. *Are GANs Created Equal? A Large-Scale Study*. arXiv:1711.10337 [cs, stat]. Oct. 2018. URL: http://arxiv.org/abs/1711.10337 (visited on 03/05/2023).

[31] Vongani H. Maluleke et al. "Studying Bias in GANs Through the Lens of Race". en. In: *Computer Vision – ECCV 2022*. Ed. by Shai Avidan et al. Vol. 13673. Series Title: Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022, pp. 344–360. ISBN: 978-3-031-19777-2 978-3-031-19778-9. DOI: 10.1007/978-3-031-19778-9_20. URL: https://link.springer.com/10.1007/978-3-031-19778-9_20 (visited on 03/16/2023).

[32] Xudong Mao et al. *Least Squares Generative Adversarial Networks*. arXiv:1611.04076 [cs]. Apr. 2017. URL: http://arxiv.org/abs/1611.04076 (visited on 03/07/2023).

[33] Maciej A. Mazurowski et al. "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance". en. In: *Neural Networks* 21.2-3 (Mar. 2008), pp. 427–436. ISSN: 08936080. DOI: 10.1016/j.neunet.2007.12.031. URL: https://linkinghub.elsevier.com/retrieve/pii/S0893608007002407 (visited on 03/20/2023).

[34] Tony C. W. Mok and Albert C. S. Chung. "Learning Data Augmentation for Brain Tumor Segmentation with Coarse-to-Fine Generative Adversarial Networks". In: vol. 11383. arXiv:1805.11291 [cs]. 2019, pp. 70–80. DOI: 10.1007/978-3-030-11723-8_7. URL: http://arxiv.org/abs/1805.11291 (visited on 03/15/2023).

[35] Sinno Jialin Pan and Qiang Yang. "A Survey on Transfer Learning". en. In: *IEEE Transactions on Knowledge and Data Engineering* 22.10 (Oct. 2010), pp. 1345–1359. ISSN: 1041-4347. DOI: 10.1109/TKDE.2009.191. URL: http://ieeexplore.ieee.org/document/5288526/ (visited on 03/16/2023).

[36] Alec Radford, Luke Metz, and Soumith Chintala. *Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks*. arXiv:1511.06434 [cs]. Jan. 2016. URL: http://arxiv.org/abs/1511.06434 (visited on 03/12/2023).

[37] A. Rosenfeld. "Computer vision: basic principles". en. In: *Proceedings of the IEEE* 76.8 (Aug. 1988), pp. 863–868. ISSN: 00189219. DOI: 10.1109/5.5961. URL: http://ieeexplore.ieee.org/document/5961/ (visited on 03/22/2023).

[38] Tim Salimans et al. *Improved Techniques for Training GANs*. arXiv:1606.03498 [cs]. June 2016. URL: http://arxiv.org/abs/1606.03498 (visited on 03/10/2023).

[39] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. *How good is my GAN?* en. arXiv:1807.09499 [cs]. July 2018. URL: http://arxiv.org/abs/1807.09499 (visited on 04/14/2023).

[40] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. en. arXiv:1409.1556 [cs]. Apr. 2015. URL: http://arxiv.org/abs/1409.1556 (visited on 03/20/2023).

[41] Christian Szegedy et al. *Going Deeper with Convolutions*. en. arXiv:1409.4842 [cs]. Sept. 2014. URL: http://arxiv.org/abs/1409.4842 (visited on 03/20/2023).

[42] Christian Szegedy et al. *Rethinking the Inception Architecture for Computer Vision*. arXiv:1512.00567 [cs]. Dec. 2015. URL: http://arxiv.org/abs/1512.00567 (visited on 03/16/2023).

[43] Luke Taylor and Geoff Nitschke. "Improving Deep Learning with Generic Data Augmentation". en. In: *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*. Bangalore, India: IEEE, Nov. 2018, pp. 1542–1547. ISBN: 978-1-5386-9276-9. DOI: 10.1109/SSCI.2018.8628742. URL: https://ieeexplore.ieee.org/document/8628742/ (visited on 03/15/2023).

[44] Lucas Theis, Aäron van den Oord, and Matthias Bethge. *A note on the evaluation of generative models*. arXiv:1511.01844 [cs, stat]. Apr. 2016. URL: http://arxiv.org/abs/1511.01844 (visited on 03/16/2023).

[45] Lorenzo Tronchin et al. "Evaluating GANs in Medical Imaging". en. In: *Deep Generative Models, and Data Augmentation, Labelling, and Imperfections*. Ed. by Sandy Engelhardt et al. Vol. 13003. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021, pp. 112–121. ISBN: 978-3-030-88209-9 978-3-030-88210-5. DOI: 10.1007/978-3-030-88210-5_10. URL: https://link.springer.com/10.1007/978-3-030-88210-5_10 (visited on 03/21/2023).

[46] Ashish Vaswani et al. *Attention Is All You Need*. en. arXiv:1706.03762 [cs]. Dec. 2017. URL: http://arxiv.org/abs/1706.03762 (visited on 03/20/2023).

[47] Lilian Weng. *From GAN to WGAN*. arXiv:1904.08994 [cs, stat]. Apr. 2019. URL: http://arxiv.org/abs/1904.08994 (visited on 03/11/2023).

[48] Jelmer M. Wolterink et al. *Deep MR to CT Synthesis using Unpaired Data*. arXiv:1708.01155 [cs]. Aug. 2017. URL: http://arxiv.org/abs/1708.01155 (visited on 03/15/2023).

[49] McKell Woodland et al. "Evaluating the Performance of StyleGAN2-ADA on Medical Images". en. In: vol. 13570. arXiv:2210.03786 [cs, eess]. 2022, pp. 142–153. DOI: 10.1007/978-3-031-16980-9_14. URL: http://arxiv.org/abs/2210.03786 (visited on 03/03/2023).

[50] Yong Xia, Wenyi Wang, and Kuanquan Wang. "ECG signal generation based on conditional generative models". en. In: *Biomedical Signal Processing and Control* 82 (Apr. 2023), p. 104587. ISSN: 17468094. DOI: 10.1016/j.bspc.2023.104587. URL: https://linkinghub.elsevier.com/retrieve/pii/S1746809423000204 (visited on 04/15/2023).

[51] Yawen Xiao, Jun Wu, and Zongli Lin. "Cancer diagnosis using generative adversarial networks based on deep learning from imbalanced data". en. In: *Computers in Biology and Medicine* 135 (Aug. 2021), p. 104540. ISSN: 00104825. DOI: 10.1016/j.compbiomed.2021.104540. URL: https://linkinghub.elsevier.com/retrieve/pii/S0010482521003346 (visited on 03/20/2023).

[52] Bing Xu et al. *Empirical Evaluation of Rectified Activations in Convolutional Network*. arXiv:1505.00853 [cs, stat]. Nov. 2015. URL: http://arxiv.org/abs/1505.00853 (visited on 03/12/2023).

[53] Qiantong Xu et al. *An empirical study on evaluation metrics of generative adversarial networks*. arXiv:1806.07755 [cs, stat]. Aug. 2018. URL: http://arxiv.org/abs/1806.07755 (visited on 03/16/2023).

[54] Jianwei Yang et al. *LR-GAN: Layered Recursive Generative Adversarial Networks for Image Generation*. arXiv:1703.01560 [cs]. Aug. 2017. URL: http://arxiv.org/abs/1703.01560 (visited on 03/16/2023).

[55] Xin Yi, Ekta Walia, and Paul Babyn. "Generative Adversarial Network in Medical Imaging: A Review". en. In: *Medical Image Analysis* 58 (Dec. 2019). arXiv:1809.07294 [cs], p. 101552. ISSN: 13618415. DOI: 10.1016/j.media.2019.101552. URL: http://arxiv.org/abs/1809.07294 (visited on 03/05/2023).

[56] Fisher Yu et al. *LSUN: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop*. arXiv:1506.03365 [cs]. June 2016. URL: http://arxiv.org/abs/1506.03365 (visited on 03/15/2023).

[57] Xiaohua Zhai et al. "S4L: Self-Supervised Semi-Supervised Learning". en. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul, Korea (South): IEEE, Oct. 2019, pp. 1476–1485. ISBN: 978-1-72814-803-8. DOI: 10.1109/ICCV.2019.00156. URL: https://ieeexplore.ieee.org/document/9010283/ (visited on 03/23/2023).

[58] Han Zhang et al. *Consistency Regularization for Generative Adversarial Networks*. en. arXiv:1910.12027 [cs, stat]. Feb. 2020. URL: http://arxiv.org/abs/1910.12027 (visited on 03/16/2023).

[59] Yu Zheng et al. *Deep AutoAugment*. en. arXiv:2203.06172 [cs]. Mar. 2022. URL: http://arxiv.org/abs/2203.06172 (visited on 03/03/2023).

[60] Zhiming Zhou et al. "ACTIVATION MAXIMIZATION GENERATIVE ADVERSARIAL NETS". en. In: (2018).

[61] Jun-Yan Zhu et al. *Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks*. en. arXiv:1703.10593 [cs]. Aug. 2020. URL: http://arxiv.org/abs/1703.10593 (visited on 03/17/2023).