INVESTIGATING THE POTENTIAL OF AUXILIARY-CLASSIFIER GANS FOR IMAGE CLASSIFICATION IN LOW DATA REGIMES

Amil Dravid[*1], *Florian Schiffers*[1], *Yunan Wu*[2], *Oliver Cossairt*[1,2], *Aggelos K. Katsaggelos*[1,2]

[1]Department of Computer Science, Northwestern University, Evanston, IL 60208, USA
[2]Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL 60208, USA
[*]amildravid2023@u.northwestern.edu

ABSTRACT

Generative Adversarial Networks (GANs) have shown promise in augmenting datasets and boosting convolutional neural network (CNN) performance on image classification tasks. But they introduce more hyperparameters to tune as well as the need for additional time and computational power to train, supplementary to the CNN. In this work, we examine the potential for Auxiliary-Classifier GANs (AC-GANs) as a 'one-stop-shop' architecture for image classification, particularly in low data regimes. Additionally, we explore modifications to the typical AC-GAN framework, changing the generator's latent space sampling scheme and employing a Wasserstein loss with gradient penalty to stabilize the simultaneous training of image synthesis and classification. Through experiments on images of varying resolutions and complexity, we demonstrate that AC-GANs show promise in image classification, achieving competitive performance with standard CNNs. These methods can be employed as an 'all-in-one' framework with particular utility in the absence of large amounts of training data.

Index Terms— Generative Adversarial Networks, Image Classification, Convolutional Neural Networks, Data Augmentation, Deep Learning

1. INTRODUCTION

Convolutional Neural Networks (CNNs) are widely used for image classification in applications from natural image classification to computer-aided diagnosis of illnesses [1]. However, they rely on large training datasets in order to generalize well on validation or testing datasets [2]. Furthermore, obtaining more data can be costly and time prohibitive. Even in the case of models trained on large datasets, performance often drops when tested on similar data from other sources or collected in a different manner [3].

Generative Adversarial Networks (GANs) [4] provide one avenue to remedy the data scarcity problem. They show particular promise in data augmentation given their ability to produce new images mimicking a prior collection [5]. A GAN consists of a *discriminator* and a *generator* model that are trained in tandem. The generator G carries out *implicit density estimation*, whereby it learns a function to sample from an estimated probability distribution p_g and generate data $G(z)$ that mimics the true data distribution $p_{data}(x)$. Furthermore, the generator tries to map onto p_g a *latent* or *noise* vector z drawn from a simple distribution p_z, such as a normal distribution. The generator is trained to minimize the divergence between p_g and p_{data}, whereas the discriminator tries to maximize the divergence. This entire process can result in photorealistic images or accurate data of other modalities [1]. Thus, GANs can provide the much-needed additional images for training CNNs. However, this entails training two separate models: the CNN for classification and the GAN for augmentation.

The Auxiliary-Classifier GAN (AC-GAN) [6] builds upon the standard GAN. Inspired by the class-conditioned GAN (C-GAN) [7], the generator is fed class information as a condition via a one-hot-encoded vector concatenated to the noise vector. The discriminator then outputs both the source of the inputted image (real or fake) and a second label corresponding to the input's class. This enables the generator to synthesize images with greater fidelity to the class label. The AC-GAN follows a two-part objective function:

$$L_S = \mathbb{E}[\log P(S = real|X_{real}))] + \mathbb{E}[\log P(S = fake|X_{fake})] \quad (1)$$

$$L_C = \mathbb{E}[\log P(C = c|X_{real}))] + \mathbb{E}[\log P(C = c|X_{fake})], \quad (2)$$

where L_S denotes the log-likelihood of the discriminator assigning the correct source S, real or fake, and L_C denotes the log-likelihood of the discriminator assigning the correct class C given real and fake images. The discriminator seeks to maximize $L_S + L_C$, while the generator tries to maximize $L_C - L_S$.

The AC-GAN objective was constructed to encourage the generator to produce more class-discriminable images [6]. Yet, there is a gap in knowledge regarding the auxiliary classifier's performance. Our contribution lies in examining the effectiveness of the AC-GAN for image classification in multiple domains, from binary to multi-class classification, and small to large-scale images. Motivated by GANs' utility in augmenting data, we particularly investigate classification using small-sized datasets. Furthermore, we propose simple

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

modifications to the AC-GAN training scheme to facilitate classification. This can serve as an all-in-one framework that avoids the need to train a separate GAN and CNN.

2. RELATED WORKS

The literature on AC-GAN's potential as an image classifier is sparse. [8] employs AC-GANs for improving a 3-class CNN liver lesion classifier. One of the experiments entails using an isolated discriminator from an AC-GAN for classification, which realizes a $\sim 2\%$ decrease in performance in comparison to their CNN model. However, they employ the exact same original AC-GAN architecture presented in [6], whereas their CNN classifier is constructed with the explicit objective of accurate liver classification. Yet, all-in-one architectures that combine generation and classification have shown promise in the literature [9, 10, 11]. For instance, [11] proposes a domain-specific variant of the AC-GAN for effective spectrum classification in hyperspectral imaging, plant segmentation, as well as image classification. Additionally, the work in [12] applies an Auxiliary-Classifier Wasserstein GAN with gradient clipping to a specific signal classification task. These works provide motivation to understand, in general, how effective AC-GANs can be as an image classifier when compared to standard CNNs in a controlled setting (i.e., similar hyperparameters, adapted to the same domain/task, etc.). We explore methods to address this and further adapt AC-GANs to image classification, particularly in low data regimes.

3. METHOD

3.1. Wasserstein loss with gradient penalty

GAN training can often be unstable as the generator may ultimately only produce low-fidelity images or produce a single output that fools the discriminator, known as *mode collapse* [13]. Traditionally, the binary cross-entropy function is used as a loss function to approximate the *Jensen-Shannon* (JS) divergence between the generator's estimated density model p_g and the underlying data distribution p_{data} [4, 14]. The discriminator seeks to maximize this divergence. This may lead to saturated gradients, and little valuable feedback for the generator to improve [15, 16]. A *Wasserstein* loss [15] approximates the *Earth Mover's Distance* (EMD) between the two distributions, which has shown to mitigate these issues. However, the discriminator's gradients need to be *1-Lipschitz* continuous in order to have a valid approximation of the EMD. That is, the norm of the discriminator's gradients must be at most one. A gradient penalty scheme [17] encourages this constraint, improving upon clipping the weights as proposed in [15].

We employ the Wasserstein Loss with gradient penalty to model the loss for the discriminator assigning the correct source to an image: L_S. We keep Equation 2 as is, modeling the log-likelihood of the class. However, we modify L_S to become:

$$L_S = \mathbb{E}[D(x)] - \mathbb{E}[D(G(z))] \tag{3}$$

where D represents the discriminator's output for the source, x represents true data, and $G(z)$ denotes synthesized samples from the generator. The generator seeks to maximize $L_C - L_S$, while the discriminator tries to minimize $L_S - \omega L_C + \lambda\Phi$. The terms λ and Φ represent a gradient penalty coefficient and regularization term to encourage 1-Lipschitz Continuity. The method for calculating the regularization term is further explained in [17]. We also introduce a weight ω for the class loss component in order to balance the discriminator's focus between classifying images as well as identifying the source. Introducing a Wasserstein loss with gradient penalty allows us to train an AC-GAN for longer to optimize its classification without sacrificing the generator's training.

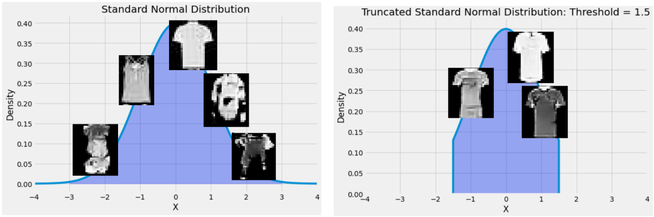
3.2. Sampling the latent vector z

Feeding in poorly generated examples into a classifier may harm classification performance, so it is ideal to augment with the most realistic possible data [11, 18]. To that end, in training the AC-GAN's discriminator for image classification, we propose using the so-called *truncation trick* presented in [19]. Oftentimes, images from a generator are produced by sampling the latent vector z from the spherical normal distribution $\mathcal{N}(0, \boldsymbol{I})$. The truncation trick entails sampling closer to the mode of the distribution (also the mean in $\mathcal{N}(0, \boldsymbol{I})$), resulting in images with greater realism, but low diversity. On the contrary, sampling further away from the mode results in lower fidelity images but with greater diversity [19]. Figure 1 provides an example to illustrate this concept.

We suggest to feed in generated images sampled by a truncated normal distribution when optimizing L_C for the discriminator. That is, sampling within a specific domain that is closer to the mode. Note that when optimizing the discriminator's parameters with respect to L_C as well as all the generator's parameters, we sample from a standard $\mathcal{N}(0, \boldsymbol{I})$. This facilitates a standard GAN training scheme that will not result in significant trade-off between diversity and fidelity in the generator. However, the truncation sampling technique is applied to optimize L_C with respect to the discriminator's objective, which will feed in higher fidelity images to improve the discriminator's image classification task. Our experiments prove this to be a useful and novel adaptation of the truncation trick, which has previously only been used after training for visualizing the generator's performance.

4. EXPERIMENTS

To examine the effectiveness of AC-GANs in classification, we conducted experiments on three datasets of different complexity. We utilized the Fashion-MNIST dataset, which consists of 28x28 grayscale images of clothing items from 10 classes [20]. Additionally, we used the CIFAR10 dataset,



(a) Full distribution

(b) Truncated distribution

Fig. 1: Drawing the generator's latent vector z from a truncated normal distribution allows us to feed in the most realistic generated samples for classification.

a collection of 32x32x3 natural images of 10 classes [21]. To address higher dimensional spaces, we used the COVID-19 Radiography Database [22], resizing the chest X-rays to 128x128x3 for COVID-positive versus negative classification.

We first examined the utility of the modified AC-GAN scheme on different training set sizes in comparison to a standard CNN and a standard AC-GAN. The AC-GAN is modified with a gradient penalty Wasserstein loss and truncation, as detailed in Section 3. From here on out, we refer to the modified AC-GAN as WAC-GAN-GPT (Wasserstein AC-GAN with Gradient Penalty and Truncation). We trained a standard CNN classifier on Fashion-MNIST using training set sizes of 500, 2500, 10000, 20000, and 40000 training examples with random horizontal and vertical flips for augmentation. For each trained model, we used 5000 images for validation and evaluated on a held-out test set of 10000 images. Next, we constructed a standard AC-GAN. To maintain a controlled setting for a fair comparison between the CNN and AC-GAN for classification, we maintained the same feature extraction and classification layers. We fed in a generator into the baseline CNN architecture, and to facilitate the generator's training, we added a layer to the CNN for discriminating between real and fake images. The AC-GAN was similarly trained on the incremental training sets, but without any traditional augmentation, and evaluated on the same test set. Lastly, we modified the AC-GAN framework using the methods we proposed in Section 3. Each modification, the truncation trick and Wasserstein loss with gradient penalty, is independently evaluated through ablation.

To examine the ability for the AC-GAN to generalize from classifying on one data distribution to a slightly shifted distribution, we used 40000 training and 10000 validation images from the CIFAR10 dataset. We then used the so-called CIFAR 10.1v6 dataset for testing, which consists of 2000 images from the Tiny Images dataset [23] with equal balance for each of the ten classes corresponding to CIFAR10. The authors of [3] constructed this dataset and empirically show that it follows a slightly varied distribution from CIFAR10. We trained an AlexNet architecture [24] on the CIFAR10 training set as [3] notes the greatest drop in performance on CIFAR10 to CIFAR 10.1v6 for this architecture. We then used the AlexNet architecture as the discriminator for our AC-GAN scheme.

Finally, we trained a CNN (with traditional data augmentation), standard AC-GAN, and WAC-GAN-GPT on a sample of the COVID-19 Radiography database at a training size of 800 with even class split, then validated on 1000 images. All common hyperparameters and initialization schemes were shared between the trained CNNs and AC-GANs for all experiments. Architecture and training details are further specified at https://github.com/avdravid/AC-GANS-FOR-IMAGE-CLASSIFICATION.

5. RESULTS AND DISCUSSION

On the Fashion-MNIST dataset, the WAC-GAN-GPT outperforms both the baseline CNN model and standard AC-GAN by $\sim 1-5\%$ based on all training set sizes, as detailed in Table 1 and Figure 2. The CNN was trained using traditional affine transformations for augmentation, but the classifier/discriminator in the AC-GAN frameworks were not. This demonstrates that the generated images provide more meaningful additional data to classify. Both the AC-GAN with truncation and AC-GAN with a gradient-penalty Wasserstein loss (WAC-GAN-GP) show slight improvement over the CNN and standard AC-GAN, with the truncation providing a greater contribution to performance gain. The combination of the two (WAC-GAN-GPT) results in the most competitive results of the five tested frameworks. The standard AC-GAN has comparable performance but does not outperform the baseline CNN at all training set sizes. GANs need sufficient data to synthesize decent quality images [25]. If there is greater stochasticity in sampling the generator's images, especially when the AC-GAN is trained on a small dataset, the generated images fed into the discriminator may be of poor quality, which can hurt the classification training process.

Train Size	Baseline CNN	AC-GAN	WAC-GAN-GP	AC-GAN with Truncation	WAC-GAN-GPT
500	$77.5\% \pm 1.5$	$77.6\% \pm 1.7$	$77.9\% \pm 1.5$	$78.8\% \pm 1.5$	$\mathbf{79}.8\% \pm 1.5$
2500	$83.5\% \pm 1.0$	$81.2\% \pm 2.1$	$84.4\% \pm 1.5$	$84.8\% \pm 1.1$	$\mathbf{86}.0\% \pm 1.2$
10000	$86.4\% \pm 1.5$	$87.3\% \pm 1.3$	$87.6\% \pm 0.9$	$87.8\% \pm 0.7$	$\mathbf{88}.4\% \pm 1.1$
20000	$87.5\% \pm 1.3$	$88.6\% \pm 1.6$	$88.1\% \pm 1.2$	$89.1\% \pm 0.5$	$\mathbf{89}.8\% \pm 0.9$
40000	$90.3\% \pm 0.8$	$90.9\% \pm 0.8$	$91.0\% \pm 0.4$	$90.7\% \pm 0.8$	$\mathbf{91}.3\% \pm 0.7$

Table 1: 95% confidence intervals for accuracy on Fashion MNIST test set as a function of training set size. Best results in bold.

To examine the distributional relationships between the AC-GANs and CNN at the low data regime, we conduct t-SNE analysis [26] using the CNN, AC-GAN, and WAC-GAN-GPT trained on just 500 Fashion MNIST samples. We sample 300 real images, 300 generated images from the standard AC-GAN and 300 images drawn using a truncated standard normal distribution from the WAC-GAN-GPT. We then

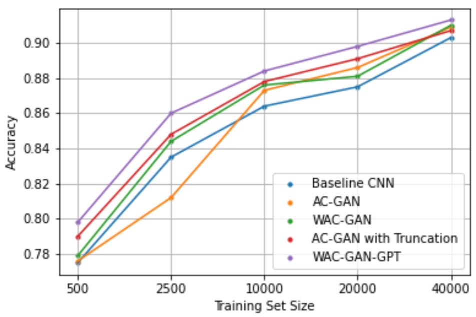
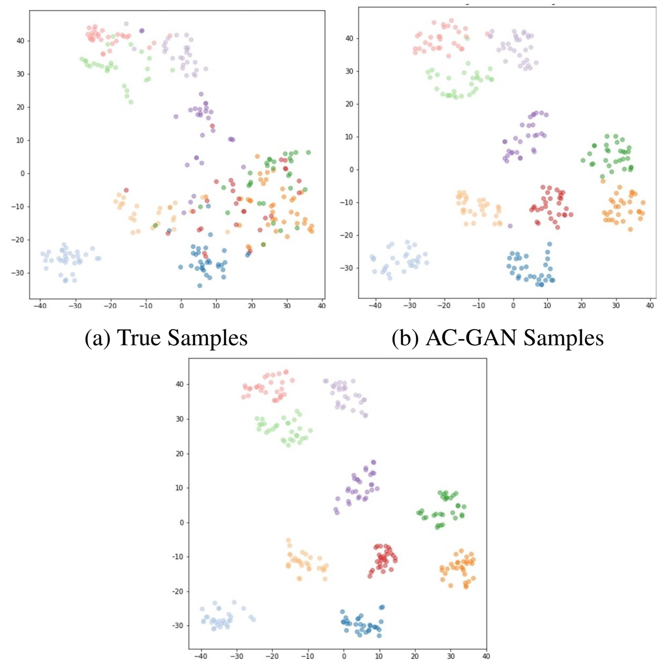


Fig. 2: Graphical comparison of average accuracy on the test set for the five models as training set size increases. All methods show comparable performance. The WAC-GAN-GPT consistently improves over the baselines, especially in the lower data regime.

feed them through the CNN to obtain feature embeddings, subsequently applying t-SNE to transform the embeddings to a two-dimensional space and visualizing the data based on either the true label for real images or the label given to the generator for generated images. An example is shown in Figure 3. The average distance for each data point to the center of its class cluster is 7.83, 5.16, and 3.94 for the CNN, AC-GAN, and WAC-GAN-GPT respectively. The standard deviations for these distances is 4.71, 2.17, and 1.76, respectively. Multiple runs of t-SNE confirm these trends. Although greater diversity of training data is vital for improving generalizability, based on the higher performance of the WAC-GAN-GPT in the low data regime, we find that the samples closer to the mean are informative in these early stages of learning. The classifier needs to first learn the most common features following a simple distribution before learning more complex, diverse ones. The standard AC-GAN without truncation may produce confusing images to train a classifier, some of which may cross the decision boundaries, which may harm training.

On the CIFAR10 dataset, the WAC-GAN-GPT method outperforms the standard CNN with traditional augmentation and the standard AC-GAN on both the CIFAR10 and the CIFAR10.1v6 dataset, as seen in Table 2. The drop in absolute accuracy from CIFAR10 to CIFAR10.1v6 is explained by a natural distribution shift in this new test set [3]. The discrepancy in the different test set accuracies is reduced from $\sim 17\%$ with the CNN to $\sim 14\%$ with the WAC-GAN-GPT. This suggests a greater generalization ability from one dataset set to another for the WAC-GAN-GPT. We consider that this is due to the nature of *implicit density estimation*: the generator's sampled images from the estimated p_g distribution does not perfectly match p_{data}. As the discriminator trains its classification task on both p_g and p_{data}, it does not overfit to p_{data} and may be able to better adapt to slightly varied distributions.

The potential for the AC-GAN for image classification is further corroborated with experiments in a higher resolution space: 128x128x3 images from the COVID-19 Radiography Database. Results are seen in Table 2 above.



(a) True Samples

(b) AC-GAN Samples

(c) WAC-GAN-GPT Samples

Fig. 3: T-SNE of Fashion MNIST true and generated samples.

	AlexNet	AC-GAN	WAC-GAN-GPT
CIFAR 10	$70.5\% \pm 0.5$	$70.1\% \pm 0.8$	$\mathbf{72.9}\% \pm 0.7$
CIFAR 10.1v6	$53.5\% \pm 1.0$	$56.4\% \pm 1.1$	$\mathbf{59.3}\% \pm 0.6$

	CNN	AC-GAN	WAC-GAN-GPT
COVID-19	$94.0\% \pm 1.5$	$95.5\% \pm 0.5$	$97.6\% \pm 0.9$

Table 2: 95% confidence intervals for accuracy on CIFAR10/CIFAR10.1v6 test sets and the COVID-19 Radiography Database validation set.

6. CONCLUSION

We have demonstrated that the AC-GAN, in fact, can achieve competitive performance with standard CNNs across datasets of varying complexity and resolution, with particular performance gains in lower data regimes. We have presented some methods which can be employed to improve accuracy and save efforts on training a separate GAN. We hope this work inspires further efforts to interface GANs with image classification as a 'one-stop-shop.' Future work can look into using AC-GANs with more diverse datasets, higher resolution images, and interfacing them with more advanced techniques, such as adaptive discriminator augmentation or progressive growing [27, 28]. This could further facilitate training with limited data and generating higher fidelity images to potentially improve upon the current benchmarks in image classification.

7. REFERENCES

[1] Zhaoqing Pan, Weijie Yu, Xiaokai Yi, Asifullah Khan, Feng Yuan, and Yuhui Zheng, "Recent progress on generative adversarial networks (gans): A survey," *IEEE Access*, vol. 7, pp. 36322–36333, 2019.

[2] Yaxing Wang, Chenshen Wu, Luis Herranz, Joost van de Weijer, Abel Gonzalez-Garcia, and Bogdan Raducanu, "Transferring gans: generating images from limited data," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 218–234.

[3] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar, "Do cifar-10 classifiers generalize to cifar-10?," *arXiv preprint arXiv:1806.00451*, 2018.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, pp. 2672–2680, 2014.

[5] Hoo-Chang Shin, Neil A Tenenholtz, Jameson K Rogers, Christopher G Schwarz, Matthew L Senjem, Jeffrey L Gunter, Katherine P Andriole, and Mark Michalski, "Medical image synthesis for data augmentation and anonymization using generative adversarial networks," in *International workshop on simulation and synthesis in medical imaging*. Springer, 2018, pp. 1–11.

[6] Augustus Odena, Christopher Olah, and Jonathon Shlens, "Conditional image synthesis with auxiliary classifier gans," in *International conference on machine learning*. PMLR, 2017, pp. 2642–2651.

[7] Mehdi Mirza and Simon Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[8] Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan, "Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, 2018.

[9] Simon Vandenhende, Bert De Brabandere, Davy Neven, and Luc Van Gool, "A three-player gan: generating hard samples to improve classification networks," in *2019 16th International Conference on Machine Vision Applications (MVA)*. IEEE, 2019, pp. 1–6.

[10] J. Wang, F. Gao, J. Dong, and Q. Du, "Adaptive dropblock-enhanced generative adversarial networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–14, 2020.

[11] Dongyi Wang, Robert Vinson, Maxwell Holmes, Gary Seibel, Avital Bechar, Shimon Nof, and Yang Tao, "Early detection of tomato spotted wilt virus by hyperspectral imaging and outlier removal auxiliary classifier generative adversarial nets (or-acgan)," *Scientific reports*, vol. 9, no. 1, pp. 1–14, 2019.

[12] Caidan Zhao, Caiyun Chen, Zeping He, and Zhiqiang Wu, "Application of auxiliary classifier wasserstein generative adversarial networks in wireless signal classification of illegal unmanned aerial vehicles," *Applied Sciences*, vol. 8, no. 12, pp. 2664, 2018.

[13] Tong Che, Yanran Li, Athul Paul Jacob, Yoshua Bengio, and Wenjie Li, "Mode regularized generative adversarial networks," *arXiv preprint arXiv:1612.02136*, 2016.

[14] Ian Goodfellow, "Nips 2016 tutorial: Generative adversarial networks," *arXiv preprint arXiv:1701.00160*, 2016.

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

[16] Xin Yi, Ekta Walia, and Paul Babyn, "Generative adversarial network in medical imaging: A review," *Medical image analysis*, vol. 58, pp. 101552, 2019.

[17] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C Courville, "Improved training of wasserstein gans," in *NIPS*, 2017.

[18] Fabio Henrique Kiyoiti dos Santos Tanaka and Claus Aranha, "Data augmentation using gans," *arXiv preprint arXiv:1904.09135*, 2019.

[19] Andrew Brock, Jeff Donahue, and Karen Simonyan, "scale gan training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.

[20] Han Xiao, Kashif Rasul, and Roland Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[21] Alex Krizhevsky, Geoffrey Hinton, et al., "Learning multiple layers of features from tiny images," 2009.

[22] Tawsifur Rahman, "Covid-19 radiography database," Mar 2021.

[23] Antonio Torralba, Rob Fergus, and William T Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 11, pp. 1958–1970, 2008.

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[25] Atsuhiro Noguchi and Tatsuya Harada, "Image generation from small datasets via batch statistics adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2750–2758.

[26] Laurens Van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[27] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila, "Training generative adversarial networks with limited data," *arXiv preprint arXiv:2006.06676*, 2020.

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