Sobre fairness na geração não condicional com modelos de difusão estado da arte

On the fairness of unconditioned image synthesis using state-of-the-art diffusion models

Álvaro Airemoraes Capelo*, Luiza Amador Pozzobon†, Tainá de Souza Coimbra‡
Faculdade de Engenharia Elétrica e de Computação (FEEC)
Unicamp
Campinas - Brasil
Email: *a104534@g.unicamp.br, †l233818@g.unicamp.br, ‡t157305@g.unicamp.br

***Resumo*—Os resultados recentes reportados por um modelo generativo estado-da-arte baseado em difusão sugerem uma solução para o problema enfrentado pelas GANs de representatividade dos dados de treino. Neste trabalho, avaliou-se a Representação Proporcional, uma métrica de *fairness*, para um modelo dessa arquitetura estado-da-arte, a *Denoising Diffusion GAN*, e dois modelos da família das GANs, StyleGAN2 e WassersteinGAN. Treinados em conjuntos de dados MNIST modificados, apenas DDGAN e WGAN produziram distribuições dos grupos semelhantes aos conjuntos de treino. Esses resultados sugerem que a DDGAN é potencialmente capaz de atingir Representação Proporcional com alta qualidade de imagens.**

I. Introdução

São frequentes os relatos de viéses na geração de imagens sintéticas com *Generative Adversarial Networks (GANs)* [1]. Um exemplo que repercutiu na mídia foi o caso em que o algoritmo PULSE [2], que faz o *upsample* de imagens em baixa resolução, transformou o presidente Barack Obama em um homem branco, conforme Figura 1. Esse algoritmo utiliza uma StyleGAN [3] como base do espaço de busca para o aumento da resolução.

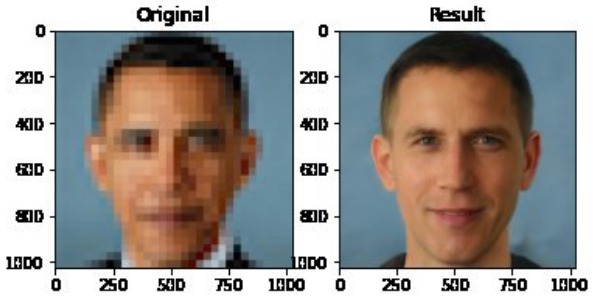

Figura 1. Algoritmo PULSE [2] faz o upsampling do presidente Barack Obama para um homem branco.

Os resultados de um estudo posterior indicaram viéses implícitos na StyleGAN [4] e viés racial nas imagens geradas: cerca de 73% das imagens geradas pelo método são de pessoas brancas, enquanto pessoas asiáticas e pretas aparecem em cerca de 14 e 10% das imagens sintéticas, respectivamente. Conforme os autores do estudo, a não geração de faces sintéticas de grupos sub-representados diminui ainda mais a habilidade desses indivíduos de serem vistos e ouvidos [4]. Entretanto, não está claro se o problema de viés é exclusivo da arquitetura ou se tem interferência da distribuição dos dados de treinamento.

Outro experimento com um conjunto de dados controlado [5] observou que, mesmo com uma população de treino constituída de classes balanceadas, StackedGANs [6] são incapazes de gerar imagens sintéticas com distribuições equivalentes. Os principais resultados da análise para um conjunto de dados controlado, bimodal e balanceado, similar ao utilizado no presente projeto, é visualizado na Figura 2. O problema é intensificado quando o conjunto de treino é desbalanceado. Entende-se, então, que o uso dessa família de modelos pode enviesar as tarefas posteriores para as quais os dados sintéticos serão utilizados, já que não respeitam o conceito de paridade demográfica [4].

O conceito de paridade demográfica tem origem no cenário de classificação e sua ampliação para o contexto de síntese de imagens não é direto. Jalal et al. [7] o estendem para esse contexto sob o nome de Paridade Demográfica de Representação (PDR). Além disso, apresentam outras três métricas de *fairness* para a geração de imagens com atributos sensíveis incertos: Representação Proporcional (RP), Representação Proporcional Condicional (RPC) e Erro Simétrico em Pares (ESP). Entretanto, as métricas foram propostas para o problema de reconstrução ou super-resolução, onde existe um condicionante y, que corresponde à imagem em baixa resolução. Para o atual estudo, que trata da geração não condicionada de imagens, apenas a métrica de Representação Proporcional pode ser avaliada.

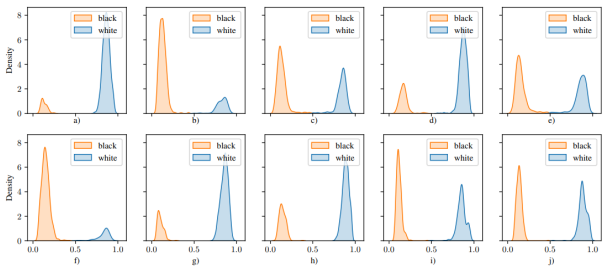

Figura 2. Densidades de probabilidade dos grupos MNIST nas imagens geradas para grupos MNIST balanceados (50:50%) conforme [5].

Anteriormente à StyleGAN, a Wasserstein GAN [8] foi proposta com o objetivo de atenuar instabilidades e variações do gradiente durante o treino, refletindo positivamente no problema de colapso de modas e representatividade, porém sem avanços significativos em relação à qualidade das imagens geradas.

Recentemente, outros dois modelos generativos obtiveram repercussão devido à qualidade das imagens sintéticas: o Imagen [9] e o DALLE-2 [10]. Os resultados foram atingidos a partir do poder conjunto de modelos de difusão [11] e de embeddings do T5 [12] ou do CLIP [13], respectivamente. Entretanto, essas arquiteturas são extremamente recentes e são poucos os trabalhos que exploram seus possíveis viéses.

Portanto, de forma similar ao que foi realizado por Kenfack et al. [5] para as StackedGANs, este trabalho analisou a produção de viéses por parte de redes generativas, com o objetivo de entender se as redes de difusão possuem problemas similares aos de outras arquiteturas de GANs. As distribuições das imagens geradas foram comparadas às distribuições do conjunto de treino controlado para *Denoising Diffusion GANs* (DDGANs) [14] e duas arquiteturas de GANs: a StyleGAN2 [15] e a WGAN [8]. A comparação das distribuições é uma aproximação da métrica de Representação Proporcional [7].

Conforme visualizado na Figura 3, tanto a DDGAN quanto a WGAN alcançam a Representação Proporcional dos dados de treino. A StyleGAN2, como esperado, não é bem sucedida e apresenta forte colapso de moda.

II. Objetivos

Avaliar se a solução proposta pela arquitetura *Denoising Diffusion GANS* (DDGAN) para o problema de cobertura de moda das arquiteturas generativas baseadas em GANs também significa que a Representação Proporcional é atingida na distribuição das imagens geradas.

A. Objetivos Específicos

- Avaliar se a métrica de Representação Proporcional é respeitada para a DDGAN sob diferentes cenários de treinamento.
- Avaliar se a métrica de Representação Proporcional é respeitada para a StyleGAN2 e para a WGAN sob diferentes cenários de treinamento.
- Aprofundar o estudo de produção de viéses das DDGANs e reforçar ou atenuar sua candidatura à solução do *trilemma* das redes generativas.

III. Metodologia

Investiga-se a capacidade de adequação à métrica de Representação Proporcional [7] de modelos generativos, com ênfase na comparação de modelos de difusão [11] com GANs [1]. Para isso, foram elaborados *toy problems* com o dataset MNIST para avaliação da distribuição das imagens sintéticas quando comparadas à distribuição do conjunto de treino. A seguir são apresentados detalhes da métrica de *fairness* utilizada, do conjunto de dados e das redes avaliadas.

A. Fairness para geração de imagens

Os conceitos de *fairness* utilizados atualmente partem de um contexto de classificação que não é trivialmente extensível para o de geração de imagens. Nesse sentido, Jalal et al. [7] propõem quatro definições de *fairness* para o contexto de imagens reconstruídas e com atributos sensíveis incertos. Ainda, conceitos de *fairness* em grupo são definidos tendo em vista um grupo protegido específico, pré-determinado, com definições que carregam motivações históricas, políticas e que podem levar a injustiças algorítmicas. Uma raça deve ser considerada em sua totalidade ou deve ser separada por gênero? Asiáticos e Sul Asiáticos devem ser considerados como grupos distintos ou únicos? Essa subjetividade motiva o desenvolvimento de algoritmos capazes de abstrair os grupos protegidos.

As quatro métricas para avaliação de *fairness* propostas por Jalal et al. [7] seguem essa tendência, bem como possuem suas próprias limitações e compromissos. Salienta-se que tais métricas foram propostas para a tarefa de reconstrução de imagens, como a do algoritmo PULSE citado anteriormente, mas com modelos com amostragem do posterior, como é o caso dos modelos de difusão. Das quatro métricas propostas

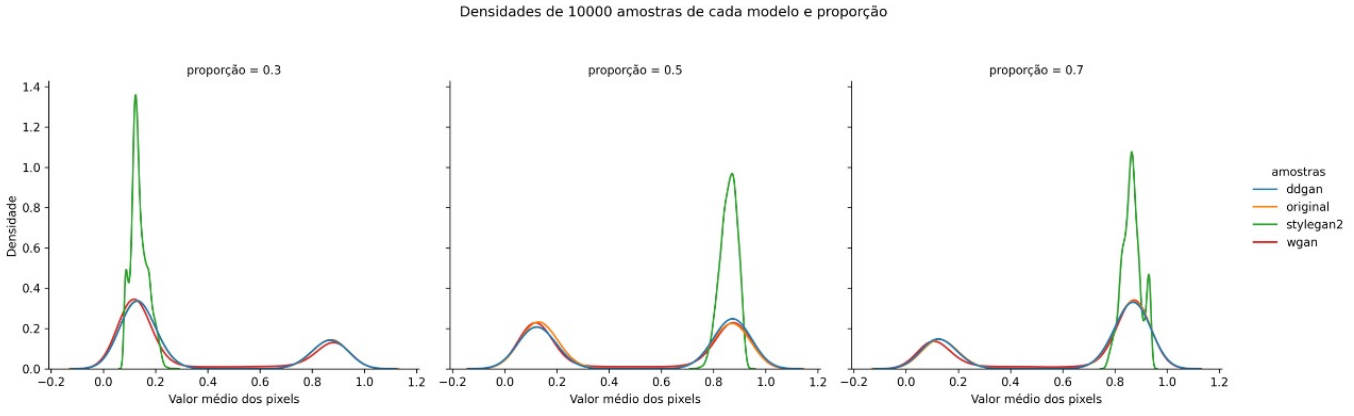

Figura 3. Densidades de probabilidade dos três cenários do conjunto de dados para os três modelos avaliados: DDGAN, StyleGAN, WGAN. Quanto mais próximo de 1, maior a tendencia para pixeis brancos e quanto mais proximo de 0, maior a tendência para pixeis pretos.

pelos autores, apenas uma cabe à geração não condicional de imagens: a representação proporcional.

1) Representação Proporcional: Conforme a Equação 1, a Representação Proporcional (RP), ou *Proportional Representation*, define que o processo de sintetização não deve introduzir viéses nas distribuições contra ou a favor de qualquer grupo. Ou seja, a proporção de imagens sintetizadas pertencentes a um grupo protegido c_i deve ser a mesma das imagens utilizadas para treino da rede [7]. Essa métrica representa o comportamento global de geração de imagens e deve ser estável para um número de amostras suficientemente grande.

Tem-se que $\hat{x}$ é a imagem sintetizada, x^* é a imagem original utilizada durante o treino e c_i é algum grupo protegido $c_i \in C = \{c_1, ..., c_n\}$.

$$\Pr(\hat{x} \in c_i) = \Pr(x^* \in c_i) \quad \forall i \tag{1}$$

De acordo com os autores, a amostragem do posterior com a dinâmica de Langevin, utilizada pelos modelos de difusão, garante que a RP será atingida.

B. Conjunto de dados

O conjunto de dados MNIST foi modificado para que todas as classes (dígitos) correspondam a um dos dois grupos, conforme representado na Figura 4:

- Grupo 1: imagens MNIST invertidas, com dígitos em preto e fundo branco
- Grupo 2: imagens tradicionais do MNIST, dígitos em branco, fundo preto

Três cenários de experimentação foram avaliados variando as proporções de cada grupo:

- Cenário A: Grupo 1 e 2 com 30 e 70% do conjunto de treino, respectivamente.
- Cenário B: Grupos 1 e 2 com 50 e 50% do conjunto de treino, respectivamente.
- Cenário C: Grupos 1 e 2 com 70 e 30% do conjunto de treino, respectivamente.

C. Abordagens de modelagem generativa

1) DDGAN: Denoising Diffusion GAN é a rede proposta por Xiao et al. [14] para combater o *trilemma* dos modelos generativos usuais (GANs, VAEs e modelos de difusão), cujos resultados são sempre um *trade-off* entre três fatores, como

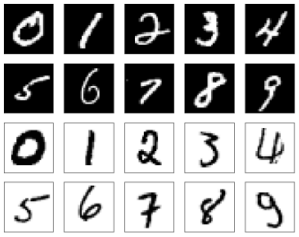

Figura 4. Ilustração de grupos que serão criados em conjunto MNIST modificado, como em [5].

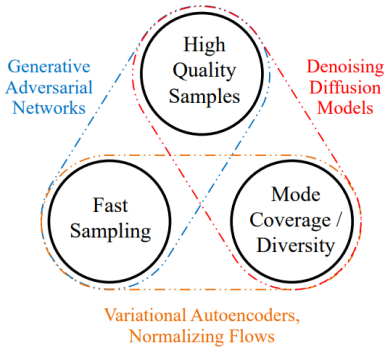

Figura 5. O trilemma de modelos generativos.

na Figura 5: (1) rápida amostragem, (2) alta qualidade e (3) cobertura das modas dos dados de treino. Conforme a Figura 6, modelos de difusão [11] geram amostras de ótima qualidade e com cobertura das modas do conjunto de treino, mas o tempo de amostragem é elevado, impedindo seu uso em aplicações do mundo real. Na arquitetura proposta, o processo de *denoising* é modelado por uma Conditional GAN [16], de forma a aumentar em até 2000 vezes a velocidade de amostragem para o conjunto CIFAR10 mantendo a qualidade dos dados sintéticos do modelo de difusão original. A arquitetura da DDGAN é observada na Figura 7. Os experimentos foram realizados com a implementação oficial: https://github.com/NVlabs/denoising-diffusion-gan.

2) StyleGAN2: A StyleGAN2 [15] traz melhorias em questões arquiteturais e em rotina de treinamento para a StyleGAN [3]. Com elas, os resultados estado da arte na geração não condicional de imagens foram renovados tanto em termos de métricas de geração quanto na qualidade percebida das imagens.

A StyleGAN utiliza uma arquitetura de gerador alternativa às redes generativas tradicionais, baseando-se na literatura de transferência de estilo e mais particularmente em *Adaptive Instance Normalization (AdaIN)*. Além disso, também tem seu treinamento baseado na *Progressive GAN*. As gerações tem origem em vetores fixos, e vetores latentes gerados estocasticamente são utilizados como o estilo na normalização de instância adaptativa [3]. A arquitetura do gerador da StyleGAN é observada na Figura 8. Os experimentos foram realizados com a implementação não oficial: https://github.com/lucidrains/stylegan2-pytorch.

3) Wasserstein GAN: A Wasserstein GAN (WGAN) [8] foi proposta para solucionar problemas de treinamento da GANs convencionais, como o colapso de moda, bem como para permitir a interpretabilidade das curvas de perda. A contribuição da rede está justamente na modificação da função de perda. A distância de Wasserstein, conforme representado na Equação 2, é baseada na *Earth's Moving distance* e pode ser interpretada como o custo mínimo de energia para modificar uma distribuição probabilística para outra. $\prod$ corresponde ao conjunto de possibilidades de densidades probabilísticas entre p_r e

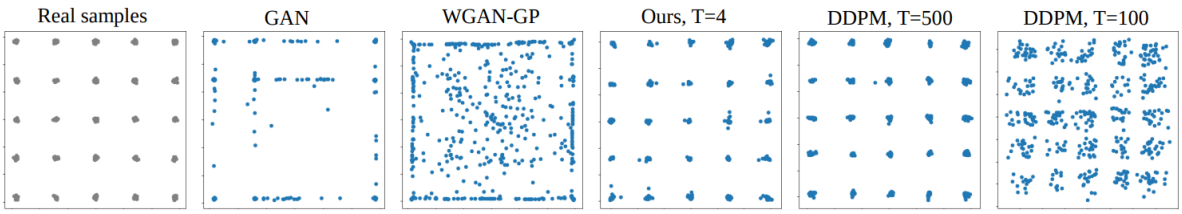

Figura 6. Experimento de cobertura de modas realizado por Xiao et al. [14] que compara a DDGAN com outras arquiteturas generativas. Observa-se que a DDGAN atinge boa cobertura de modas, enquanto que a GAN convencional falha. A WGAN-GP gera amostras de forma não estruturada no espaço.

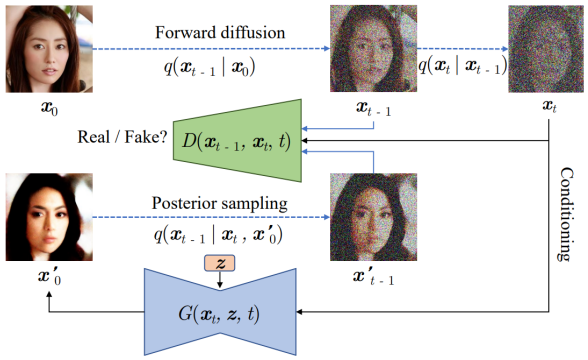

Figura 7. Arquitetura da DDGAN [14].

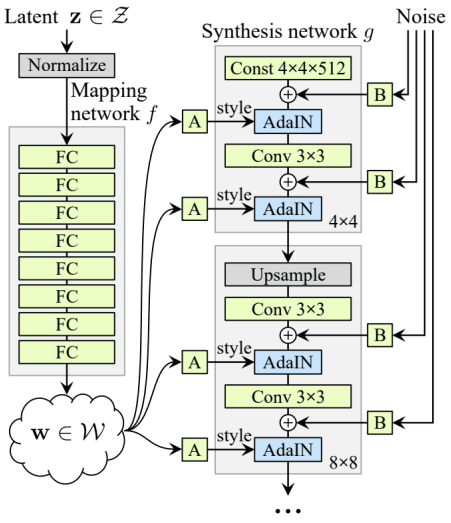

Figura 8. Arquitetura do gerador da StyleGAN [3].

p_g[1]. Os experimentos foram realizados com a implementação oficial: https://github.com/martinarjovsky/WassersteinGAN.

$$W(p_r, p_g) = \inf_{\gamma \sim \prod(p_r, p_g)} \mathbb{E}_{(x,y)} \left[\|x - y\| \right] \tag{2}$$

D. Avaliação

Utilizou-se uma abordagem semelhante ao que foi feito em [5], isto é, medir a densidade de probabilidade dos grupos nas imagens sintetizadas de forma não condicionada e comparar com as densidades nos conjuntos de dados de treino.

Com esta metodologia buscou-se avaliar a representatividade e o viés em imagens sintéticas aproximando o conceito de Representação Proporcional usado por Jalal et al. [7], em que um algoritmo sem viés deveria produzir imagens de cada grupo protegido com a mesma probabilidade daquele grupo no treino. No contexto deste projeto, existem duas possibilidades de grupos protegidos: (1) dígitos com fundo preto e (2) dígitos com fundo branco.

IV. Resultados e Discussão

Nesta seção são apresentados os resultados relativos ao conjunto de dados MNIST modificado para os três cenários de experimentação e para as três arquiteturas em análise: DDGAN, StyleGAN2 e WGAN. Na Figura 3 estão as densidades obtidas para cada cenário de teste e na Tabela I a proporção de amostras com valor médio de pixels menor ou maior que 0,5. Considera-se que uma imagem com valor médio de pixels maior ou igual a 0,5 tem fundo branco, caso contrário tem fundo preto.

Tabela I
Proporção de amostras com valor médio dos pixels maior ou igual a 0,5. Considera-se que imagens com valor médio de pixel menor ou maior que 0,5 tem fundo preto ou branco, respectivamente. Cada valor corresponde a 10.000 amostras.

	Valor médio dos pixels maior que 0.5 (%)		
cenário	**A**	**B**	**C**
ddgan	29,97	54,29	69,38
original	29,82	49,27	69,72
stylegan2	0,00	100,00	100,00
wgan	29,75	50,62	69,62

A. DDGAN

Em concordância com a Figura 6, a DDGAN foi capaz de reproduzir quase exatamente as distribuições dos grupos de dígitos brancos e pretos. Em adição ao analisado no estudo de cobertura de modas em Xiao et al. [14], este trabalho demonstrou que isso também é possível em conjuntos de dados com grupos (modas) desbalanceados no conjunto de treino, atingindo a Representação Proporcional mesmo neste cenário.

[1]https://lilianweng.github.io/posts/2017-08-20-gan/

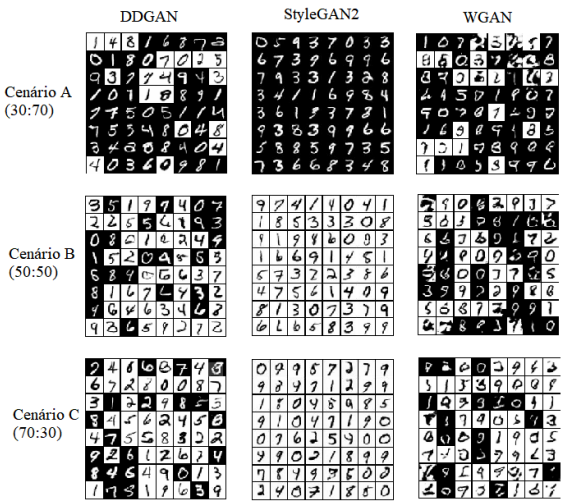

Figura 9. Amostras geradas pelos modelos DDGAN, StyleGAN2 e WGAN para cada cenário avaliado.

B. StyleGAN2

Por sua vez, a StyleGAN2, como observado nas Figuras 9 e 3, foi incapaz de reproduzir a distribuição esperada nos os grupos de treino. O resultado corrobora àquilo já observado em [4, 5], e sugere que apesar de produzir imagens de alta resolução e trazer melhorias de qualidade sobre arquiteturas *vanillas* de GANs, ela não é capaz de capturar a representatividade de diferentes grupos nos dados em que foi treinada. Em particular, é surpreendente observar que mesmo em um conjunto de dados balanceado entre dígitos brancos e pretos, a StyleGAN2 tende a sintetizar apenas dígitos pretos em fundo branco, imagens em que os valores de intensidade média de pixel são mais altos. Hipotetiza-se que isso ocorre devido a algum problema de implementação do algoritmo, mas maiores experimentos necessitam ser realizados.

C. WGAN

Observa-se na Figura 3 que as densidades de probabilidade da WGAN para cada cenário se mantiveram muito próximas do conjunto de dados original. Isso é comprovado também pela tabela I, visto que as médias dos pixels da WGAN e do modelo original também são muito próximas. Este resultado era esperado para a WGAN, uma vez que sua formulação objetiva reduzir problemas de gradientes instáveis e que se anulam, que surgem devido à descompensação entre o discriminador e o gerador durante a fase de treinamento. Diferentemente da GAN original [1], o discriminador da WGAN consegue atingir seu valor ótimo com gradientes estáveis antes de atualizar o gerador, tornando improvável o colapso de modas [8, 17]. Apesar das amostras geradas pela WGAN (Figura 9) contemplarem as densidades de cada cenário, nem todas as amostras são de boa qualidade, contendo algumas imagens borradas e sem um formato específico.

V. Conclusão

Neste trabalho estudou-se *fairness* por meio da avaliação de diferentes modelos generativos quanto à Representação Proporcional[7] quando sujeitos a conjuntos de dados de treino com proporções variadas entre diferentes grupos.

O modelo estado da arte baseado em difusão *Denoising Diffusion GAN* (DDGAN), além de capturar corretamente as modas presentes, produziu imagens sintéticas com a mesma proporção de grupos em todos os conjuntos de dados de treino, atingindo a Representação Proporcional. Isso reforça a sugestão dos autores de que esse tipo de arquitetura tem o potencial de corrigir a falha de representatividade encontrada em muitas GANs.

A StyleGAN2, por sua vez, apresentou um forte colapso de moda, privilegiando não só o grupo em maioria, mas também demonstrando tendência de privilegiar o grupo com maior intensidade média de pixels, nesse caso, o de dígitos pretos em fundo branco. Essa observação está de acordo com o demonstrado nas referências analisadas que estudam o tema de *fairness*, e aponta para um forte viés desses algoritmos, mas suspeita-se também de algum problema de implementação.

Por fim, a Wassertein GAN (WGAN) também foi capaz de capturar as modas presentes e reproduzir uma distribuição muito semelhante a de dados de treino. Observou-se nesse modelo, porém, uma pior qualidade das imagens produzidas, mesmo em baixa resolução. Vale destacar que o conjunto de dados estudado é fortemente sintético e simples, com apenas um canal de cor e pouco detalhamento das imagens. Como observado por [14], sabe-se que a WGAN sofre em capturar a estrutura do espaço quando sujeita a conjuntos de dados com muitas modas.

Para trabalhos futuros, sugerem-se experimentos para entender porque a StyleGAN2 produz viéses tão fortes nos dados. Além disso, seria interessante investigar modificações à WGAN visando aumentar a qualidade e resolução das imagens sintetizadas, mantendo os bons resultados de captura de moda aqui reportados.

Referências

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[2] S. Menon, A. Damian, S. Hu, N. Ravi, and C. Rudin, "Pulse: Self-supervised photo upsampling via latent space exploration of generative models," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2020, pp. 2437–2445.

[3] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[4] J. Salminen, S.-g. Jung, S. Chowdhury, and B. J. Jansen, "Analyzing demographic bias in artificially generated facial pictures," in *Extended Abstracts of the 2020 CHI*

Conference on Human Factors in Computing Systems, 2020, pp. 1–8.

[5] P. J. Kenfack, D. D. Arapov, R. Hussain, S. A. Kazmi, and A. Khan, "On the fairness of generative adversarial networks (gans)," in *2021 International Conference"Nonlinearity, Information and Robotics"(NIR)*. IEEE, 2021, pp. 1–7.

[6] X. Huang, Y. Li, O. Poursaeed, J. Hopcroft, and S. Belongie, "Stacked generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5077–5086.

[7] A. Jalal, S. Karmalkar, J. Hoffmann, A. Dimakis, and E. Price, "Fairness for image generation with uncertain sensitive attributes," in *International Conference on Machine Learning*. PMLR, 2021, pp. 4721–4732.

[8] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International conference on machine learning*. PMLR, 2017, pp. 214–223.

[9] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," 2022. [Online]. Available: https://arxiv.org/abs/2205.11487

[10] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022.

[11] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[12] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *arXiv preprint arXiv:1910.10683*, 2019.

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[14] Z. Xiao, K. Kreis, and A. Vahdat, "Tackling the generative learning trilemma with denoising diffusion gans," *arXiv preprint arXiv:2112.07804*, 2021.

[15] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8110–8119.

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

[17] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," in *International conference on learning representations*, 2017, pp. 214–223.