Self-Consuming Generative Models Go MAD

Josue Casco-Rodriguez,* Sina Alemohammad,* Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, Richard G. Baraniuk
Department of Electrical and Computer Engineering, Rice University

Abstract

Seismic advances in generative AI algorithms have led to the temptation to use AI-synthesized data to train next-generation models. Repeating this process creates autophagous ("self-consuming") loops whose properties are poorly understood. We conduct a thorough analysis using state-of-the-art generative image models of three autophagous loop families that differ in how they incorporate fixed or fresh real training data and whether previous generations' samples have been biased to trade off data quality versus diversity. Our primary conclusion across all scenarios is that *without enough fresh real data in each generation of an autophagous loop, future generative models are doomed to have their quality (precision) or diversity (recall) progressively decrease.* We term this condition Model Autophagy Disorder (MAD) and show that appreciable MADness arises in just a few generations.

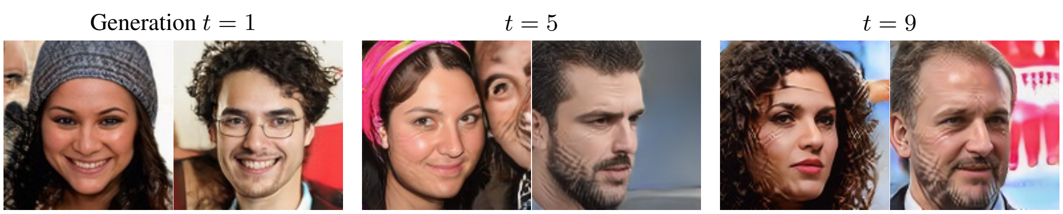

Figure 1: **Training generative artificial intelligence (AI) models on synthetic data progressively amplifies artifacts.** As AI-generated data proliferates, future models will train on both real and synthetic data in *autophagous ("self-consuming") loops*. To highlight a consequence of autophagy, we trained a sequence of StyleGAN2 [1] models wherein the model at generation $t \geq 2$ trains only on synthetic data from generation $t-1$: a fully synthetic loop (Figure 3) without sampling bias ($\lambda = 1$). The cross-hatched artifacts (possibly an architectural *fingerprint* [2]) are progressively amplified.

1 Introduction

Synthetic data from *generative artificial intelligence (AI)* models like Stable Diffusion and ChatGPT [3, 4] is rapidly proliferating on the Internet; soon, synthetic may outnumber real data. Today's AI models use Internet-scraped data, and thus unwittingly train on synthetic data (Figure 2). Moreover, AI-synthesized data is increasingly popular [5–10] because it is convenient [11, 12], anonymous [13–16], can augment real data [17, 18], and can match AI models' ever-increasing sizes [19–21].

Generative models can train on synthetic data repeatedly, forming **autophagous ("self-consuming") loops** (Figure 3), which vary not only on how they use real and synthetic data, but also on whether they incorporate *sampling biases* to trade off perceptual *quality* versus *diversity*. If synthetic data is in our training datasets today, then future autophagous loops are inevitable—and yet, their effects are poorly understood. In one direction, autophagy may amplify synthetic biases and artifacts (*fingerprints*), as in Figure 1. In another direction, autophagy with sampling biases could dilute data diversity, as in Figure 5. We describe these and other symptoms of autophagy as *Model Autophagy Disorder (MAD)*.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

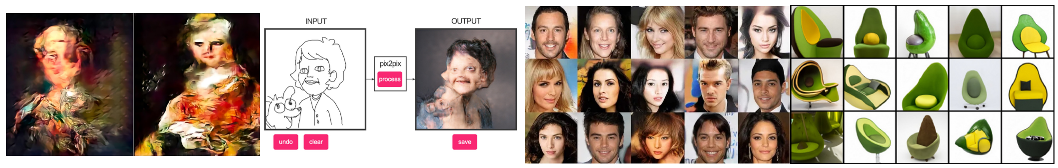

Figure 2: **Today's large-scale image training datasets contain AI syntheses,** including LAION-5B, which trains Stable Diffusion and has samples from AICAN, Pix2Pix, StyleGAN, and DALL-E [1, 3, 22–25]. Generative models using LAION-5B thus close an autophagous loop (Figure 3).

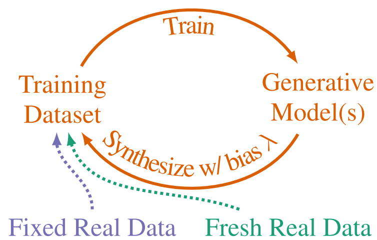

Figure 3: **Recursively training generative models on synthetic data from other models produces an autophagous ("self-consuming") loop.** In this paper, we study three autophagous loop variants: the *fully synthetic loop* (only synthetic data), the *synthetic augmentation loop* (synthetic + fixed real data), and the *fresh data loop* (synthetic + fresh real data). Each generation samples with a bias λ that trades off sample quality versus diversity.

Contributions. We thoroughly study AI autophagy via generative image models; our findings apply to any data type (e.g., text) and unify contemporary results. Our three key contributions establish that, *without enough fresh real data each generation, future generative models are doomed to go MAD.* Moreover, we demonstrate that MADness occurs in only a handful of generations.

1. Realistic models for autophagous loops. We propose 3 types of self-consuming loops (Figure 3):

The fully synthetic loop (Section 2), where each generation's training data is entirely synthesized by previous generations; e.g., training a model on its own outputs [26]. We show that, in this case, *the synthetic quality (precision) or diversity (recall) decreases over generations.*

The synthetic augmentation loop (Section 3), where each generation's training data includes previous generations' syntheses and a fixed set of real data; e.g., training on real and self-generated data [27]. We show that *fixed real training data only slows the degradation of synthetic quality or diversity.*

The fresh data loop (Section 4), where each generation's training data includes previous generations' syntheses and a fresh set of real data; e.g., training on both real and synthetic Internet data (Figure 2). We show that, *with enough fresh real data, the synthetic quality and diversity do not degrade.*

2. Sampling bias plays a key role in autophagous loops. Practitioners often favor synthetic quality, whether through curation or model-intrinsic mechanisms that boost quality (*precision*) and sacrifice diversity (*recall*) [28], like truncation and guidance [29–33]. We unify these different *sampling biases* under a universal parameter $\lambda \in [0, 1]$. Decreasing λ generally increases quality and decreases diversity. Specific definitions for λ include: sampling from $\mathcal{N}(\boldsymbol{\mu}, \lambda\boldsymbol{\Sigma})$ for any Gaussian $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, defining $\lambda = \Psi$ for StyleGAN2 with truncation $\Psi \in [0, 1]$, and defining $\lambda = \frac{1}{1+w}$ for diffusion with guidance [32] $w \in [0, \infty]$. We show that, *without these biases ($\lambda = 1$), MADness degrades quality and diversity, while with them ($\lambda < 1$), quality can persevere but diversity degrades even faster*.

3. Autophagous loop behaviors hold across various generative models and datasets. We use multivariate Gaussian, Gaussian mixture, diffusion (DDPM), StyleGAN2, Wasserstein GAN (WGAN), and Normalizing Flow [30, 34–36] models on datasets like FFHQ and MNIST [37, 38].

Related work. Contemporary works on AI autophagy support our conclusions. [39] show that variational autoencoders and Gaussian mixture models in fully synthetic loops, and language models in synthetic augmentation and fresh data loops, can go MAD. However, they do not incorporate sampling biases, and they fine-tune some of their models, while we train ours from scratch. Meanwhile, [40, 41] conduct fully synthetic and synthetic augmentation loops with diffusion models and report the same conclusions on sampling bias as us. Finally, [42] find that even one synthetic augmentation loop generation can induce MADness, hurting downstream tasks like classification.

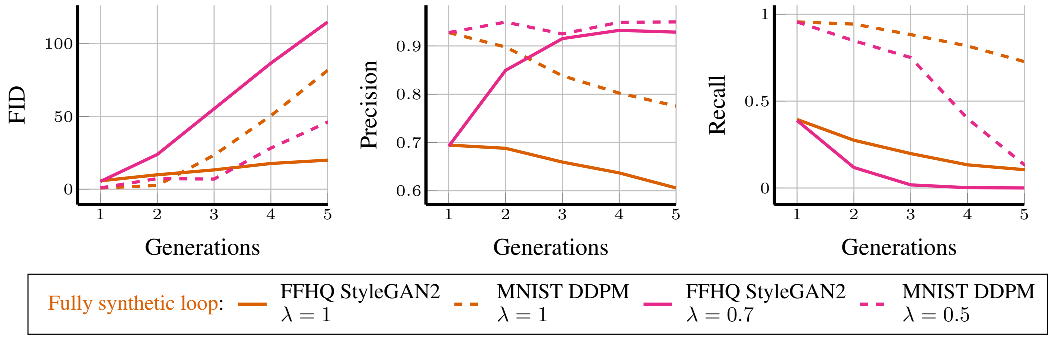

Figure 4: **Training generative models in a fully synthetic loop reduces synthetic quality and/or diversity, depending on sampling bias.** We plot the FID, precision (quality), and recall (diversity) of synthetic FFHQ and MNIST images from fully synthetic loops with unbiased ($\lambda = 1$) and biased ($\lambda < 1$) StyleGAN2 and DDPM models (for MNIST FIDs, we use LeNet [43]). FID increases and diversity decreases. However, sampling bias can salvage quality at the expense of diversity.

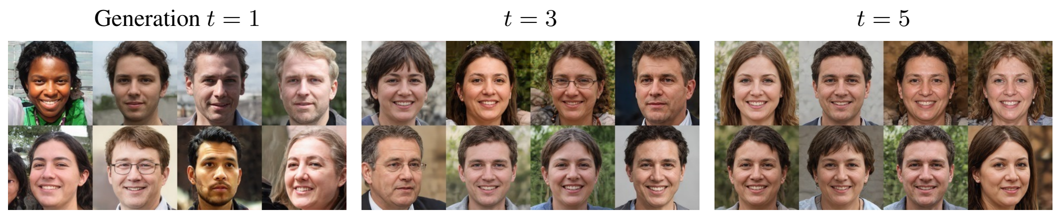

Figure 5: **Training generative models on biased synthetic data in a fully synthetic loop progressively loses diversity.** We repeat the Figure 1 experiment but with sampling bias $\lambda = 0.7$ (Figure 4).

2 The fully synthetic loop: Training only on synthetic data leads to MADness

Unbiased sampling degrades synthetic data quality and diversity. In Figure 4 we empirically study the fully synthetic loop using FFHQ StyleGAN2 and MNIST DDPM models with ($\lambda < 1$) and without ($\lambda = 1$) sampling bias. In the latter case, synthetic data distributions undergo random walks that deviate from the reference distribution because each generation's training data is finite. Consequently, the models go MAD: FID [44] increases, while quality and diversity steadily decrease.

Biased sampling can boost synthetic data quality, but at the expense of diversity. As for the biased FFHQ StyleGAN2 and MNIST DDPM models ($\lambda = 0.7$ and 0.5) in fully synthetic loops (Figure 4), sampling bias increases precision, but also accelerates losses in diversity (shown clearly in Figure 5) compared to unbiased models. Moreover, the FID still increases, indicating MADness.

3 The synthetic augmentation loop: Fixed real data only slows MADness

Fully synthetic loop analysis is tractable, but practitioners will use real data when available. Figure 6 shows how keeping the full FFHQ dataset in a StyleGAN2 synthetic augmentation loop still produces the same symptoms (albeit more slowly) as the fully synthetic loop: the distance from the real dataset (FID) increases, while the quality (precision) and diversity (recall) of synthetic samples still decrease without sampling bias. In fact, we see the same artifacts appear as in Figure 1. Additional MNIST DDPM experiments confirm these trends for synthetic augmentation loops with sampling bias.

4 The fresh data loop: Fresh real data can prevent MADness

We imagine that a data pool (e.g., the Internet) contains real and synthetic data. Independently drawing n^t samples from this pool yields n_r^t real and n_s^t synthetic samples ($n_r^t + n_s^t = n^t$) to train the t-th generation model. This fresh data loop reveals two intriguing phenomena:

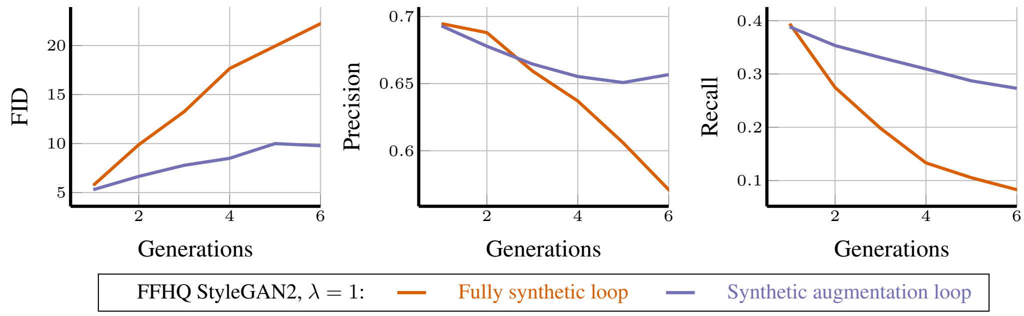

Figure 6: **Training generative models in a synthetic augmentation loop reduces synthetic quality and/or diversity, albeit more slowly than in the fully synthetic loop.** We show the FID, precision (quality), and recall (diversity) of FFHQ syntheses from unbiased ($\lambda = 1$) synthetic augmentation (where the original dataset is kept) and fully synthetic (for reference, from Figure 4) loops.

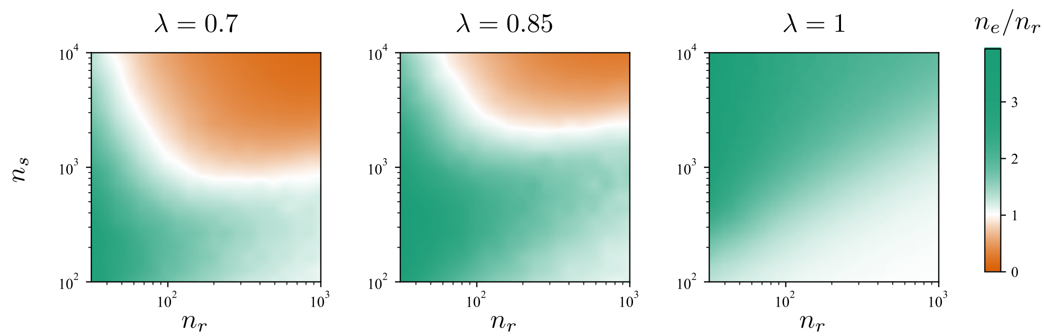

Figure 7: **In a fresh data loop, the benign amount of synthetic data does not increase with the amount of real data.** As the real data count n_r increases, the synthetic data count n_s for which $n_e \geq n_r$ (**green** area) converges. Synthetic data is only likely to be helpful for small n_r.

Initial models will eventually be forgotten in the fresh data loop. For both MNIST DDPM and Gaussian models with constant $n_r^t = n_r$ and $n_s^t = n_s$ for all t, the FID and Wasserstein distance [45] converged depending on n_r, n_s, and λ, not on the initial models or the initial dataset size n_s^1. These distances converging instead of always increasing means that *fresh real data can prevent MADness*.

Modest (but not excessive) amounts of synthetic data can help a fresh data loop. We Monte-Carlo simulate fresh data loop asymptotic Wasserstein distances in autophagous Gaussian models, and calculate the *effective sample size* n_e that an alternative model would need to perform the same as the asymptote from scratch. If n_e/n_r is greater (or less) than 1, synthetic data effectively increases (or decreases) the real sample size. In Figure 7, the non-MAD region $n_e/n_r \geq 1$ grows with λ and shrinks with n_s. Practitioners generally sample with bias, so $\lambda < 1$ conclusions are more useful.

5 Discussion

We extrapolate what may happen as generative models become ubiquitous and train future models in autophagous (self-consuming) loops: without enough fresh real data, future models are doomed to Model Autophagy Disorder (MAD), progressively losing quality (precision) or diversity (recall), and amplifying artifacts. Uncontrolled MAD, even after just five generations, could poison the Internet's data quality and diversity (Figures 1 and 5). Practitioners who deliberately use synthetic training data should heed our warning, while those who unknowingly train on synthetic data could try identifying [46–48] and rejecting synthetic data, perhaps through *watermarking* [49–55]. However, watermarking inserts hidden artifacts that autophagy could uncontrollably amplify.

Future works could combine or alternate our autophagous loop families, examine how MADness affects downstream tasks (e.g., classification), and use other data types. We have focused on imagery, but other data types, like text, cannot avoid autophagy [27, 56, 57] and MADness [39].

References

[1] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.

[2] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *NeurIPS*, 2021.

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[4] Roberto Gozalo-Brizuela and Eduardo C. Garrido-Merchan. ChatGPT is not all you need. a state of the art review of large generative ai models. *arXiv preprint arXiv:2301.04655*, 2023.

[5] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? In *ICLR*, 2023.

[6] Hritik Bansal and Aditya Grover. Leaving reality to imagination: Robust classification via generated datasets. *arXiv preprint arXiv:2302.02503*, 2023.

[7] Shaobo Lin, Kun Wang, Xingyu Zeng, and Rui Zhao. Explore the power of synthetic data on few-shot object detection. *arXiv preprint arXiv:2303.13221*, 2023.

[8] Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Lichao Sun, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. AugGPT: Leveraging ChatGPT for text data augmentation. *arXiv preprint arXiv:2302.13007*, 2023.

[9] Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*, 2023.

[10] Mert Bulent Sariyildiz, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transferable representations from synthetic ImageNet clones. In *CVPR*, 2023.

[11] Walter H. L. Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F. Da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M. Jorge Cardoso. Brain imaging generation with latent diffusion models. In *Deep Generative Models*. Springer Nature, 2022.

[12] Chengyuan Deng, Shihang Feng, Hanchen Wang, Xitong Zhang, Peng Jin, Yinan Feng, Qili Zeng, Yinpeng Chen, and Youzuo Lin. OpenFWI: Large-scale multi-structural benchmark datasets for full waveform inversion. In *NeurIPS*, 2022.

[13] Lorenzo Luzi, Ali Siahkoohi, Paul M Mayer, Josue Casco-Rodriguez, and Richard Baraniuk. Boomerang: Local sampling on image manifolds using diffusion models. *arXiv preprint arXiv:2210.12100*, 2022.

[14] Marvin Klemp, Kevin Rösch, Royden Wagner, Jannik Quehl, and Martin Lauer. LDFA: Latent diffusion face anonymization for self-driving applications. *arXiv preprint arXiv:2302.08931*, 2023.

[15] Kai Packhäuser, Lukas Folle, Florian Thamm, and Andreas Maier. Generation of anonymous chest radiographs using latent diffusion models for training thoracic abnormality classification systems. *arXiv preprint arXiv:2211.01323*, 2022.

[16] August DuMont Schütte, Jürgen Hetzel, Sergios Gatidis, Tobias Hepp, Benedikt Dietz, Stefan Bauer, and Patrick Schwab. Overcoming barriers to data sharing with medical image generation: a comprehensive evaluation. *NPJ Digital Medicine*, 2021.

[17] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*, 2023.

[18] Max F Burg, Florian Wenzel, Dominik Zietlow, Max Horn, Osama Makansi, Francesco Locatello, and Chris Russell. A data augmentation perspective on diffusion models and retrieval. *arXiv preprint arXiv:2304.10253*, 2023.

[19] The Economist. The bigger-is-better approach to AI is running out of road. *The Economist*, June 2023.

[20] The Economist. Large, creative AI models will transform lives and labour markets. *The Economist*, April 2023.

[21] Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho. Will we run out of data? an analysis of the limits of scaling datasets in machine learning. *arXiv preprint arXiv:2211.04325*, 2022.

[22] Christoph Schuhmann et al. LAION-5B: An open large-scale dataset for training next generation image-text models. In *NeurIPS Datasets and Benchmarks Track*, 2022.

[23] Ahmed Elgammal, Bingchen Liu, Mohamed Elhoseiny, and Marian Mazzone. CAN: Creative adversarial networks, generating "art" by learning about styles and deviating from style norms. *arXiv preprint arXiv:1706.07068*, 2017.

[24] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[25] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021.

[26] followfox.ai. The power of synthetic data: Infinite loop to improve fine-tuning results with stable diffusion models, February 2023. URL https://followfoxai.substack.com/p/the-power-of-synthetic-data-infinite.

[27] Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*, 2022.

[28] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *NeurIPS*, 2019.

[29] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *ICLR*, 2019.

[30] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020.

[31] Ahmed Imtiaz Humayun, Randall Balestriero, and Richard Baraniuk. Polarity sampling: Quality and diversity control of pre-trained generative networks via singular values. In *CVPR*, 2022.

[32] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[33] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[34] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

[35] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of Wasserstein GANs. *NeurIPS*, 2017.

[36] Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11), 2020.

[37] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.

[38] Li Deng. The MNIST database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6), 2012.

[39] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*, 2023.

[40] Gonzalo Martínez, Lauren Watson, Pedro Reviriego, José Alberto Hernández, Marc Juarez, and Rik Sarkar. Towards understanding the interplay of generative artificial intelligence and the Internet. *arXiv preprint arXiv:2306.06130*, 2023.

[41] Gonzalo Martínez, Lauren Watson, Pedro Reviriego, José Alberto Hernández, Marc Juarez, and Rik Sarkar. Combining generative artificial intelligence (AI) and the Internet: Heading towards evolution or degradation? *arXiv preprint arXiv:2303.01255*, 2023.

[42] Ryuichiro Hataya, Han Bao, and Hiromi Arai. Will large-scale generative models corrupt future datasets? *arXiv preprint arXiv:2211.08095*, 2022.

[43] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

[44] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *NeurIPS*, 2017.

[45] Leonid V Kantorovich. Mathematical Methods of Organizing and Planning Production. *Management Science*, 6(4), 1960.

[46] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. Deepfake detection by analyzing convolutional traces. In *CVPR workshops*, 2020.

[47] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint arXiv:2301.11305*, 2023.

[48] Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. The science of detecting llm-generated texts. *arXiv preprint arXiv:2303.07205*, 2023.

[49] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Manli Shu, Khalid Saifullah, Kezhi Kong, Kasun Fernando, Aniruddha Saha, Micah Goldblum, and Tom Goldstein. On the reliability of watermarks for large language models. *arXiv preprint arXiv:2306.04634*, 2023.

[50] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. *arXiv preprint arXiv:2301.10226*, 2023.

[51] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Ngai-Man Cheung, and Min Lin. A recipe for watermarking diffusion models. *arXiv preprint arXiv:2303.10137*, 2023.

[52] Sen Peng, Yufei Chen, Cong Wang, and Xiaohua Jia. Protecting the intellectual property of diffusion models by the watermark diffusion process. *arXiv preprint arXiv:2306.03436*, 2023.

[53] Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. *arXiv preprint arXiv:2305.20030*, 2023.

[54] Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. *arXiv preprint arXiv:2303.15435*, 2023.

[55] Jianwei Fei, Zhihua Xia, Benedetta Tondi, and Mauro Barni. Supervised GAN watermarking for intellectual property protection. In *Workshop on Information Forensics and Security (WIFS)*, 2022.

[56] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.

[57] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpaca: A strong, replicable instruction-following model, 2023.