NEPOTISTICALLY TRAINED GENERATIVE IMAGE MODELS COLLAPSE

Matyas Bohacek
Stanford University
maty@stanford.edu

Hany Farid
University of California, Berkeley
hfarid@berkeley.edu

ABSTRACT

Trained on massive amounts of human-generated content, AI-generated image synthesis is capable of reproducing semantically coherent images that match the visual appearance of its training data. We show that when retrained on even small amounts of their own creation, these generative-AI models produce highly distorted images. We also show that this distortion extends beyond the text prompts used in retraining, and that once affected, the models struggle to fully heal even after retraining on only real images.

1 INTRODUCTION

From text to image, audio, and video, today's generative-AI systems are trained on large quantities of human-generated content. Most of this content is obtained by scraping a variety of online sources (Zhang & Tang, 2024; Srinivasan et al., 2021; Schuhmann et al., 2022). As generative AI becomes more common, it is reasonable to expect that future data scraping will invariably catch generative AI's own creations (Martínez et al., 2023b;a). We ask what happens when these generative systems are trained on varying combinations of human-generated and AI-generated content.

Despite the rapidly accelerating capabilities of generative AI, evidence suggests that retraining a model on its own creation — what we call model self-poisoning — leads to artifacts in the output of the newly trained model. It has been shown, for example, that when retrained on their own output, large language models (LLMs) contain irreversible defects that cause the model to produce gibberish (Shumailov et al., 2023; 2024). Similarly, on the image generation side, it has been shown (Alemohammad et al., 2023) that when retrained on its own creations, StyleGAN2 (Karras et al., 2020) generates images (of faces or digits) with visual and structural defects. Interestingly, the authors found that there was a deleterious effect as the ratio of AI-generated content used to retrain the model ranged from 0.3% to 100%.

It has been shown that in addition to GAN-based image generation, diffusion-based text-to-image models are also vulnerable (Xing et al., 2024). The authors in (Martínez et al., 2023b;a) showed that in a simplified setting, retraining on one's own creation can lead to image degradation and a loss of diversity. Examining the impact of self-poisoning of ID-DPM (Nichol & Dhariwal, 2021), Hataya et al. (2023) report somewhat conflicting results depending on the task at hand (recognition, captioning, or generation). With respect to generation, the authors report a relatively small impact on image quality but do note a lack of diversity in the retrained models.

Building on these earlier studies, we show that the popular open-source model Stable Diffusion[1] (SD) is highly vulnerable to data self-poisoning. In particular, we show that when iteratively retrained on faces of its own creation, the model — after an initial small improvement — quickly collapses, yielding highly distorted and less diverse faces. Somewhat surprisingly, even when the retraining data contains only 3% of self-generated images, this model collapse persists. We also investigate the extent of this model self-poisoning beyond the prompts used for retraining and examine the ability of the poisoned model to heal with further retraining on only real images.

[1] https://github.com/Stability-AI/StableDiffusion

Figure 1: Examples of real images (top) used to seed image-to-image generation (bottom).

2 METHODS

2.1 IMAGES

Starting with the FFHQ image dataset containing 70,000 faces of size 1024×1024 pixels (Karras et al., 2019), we automatically classified (Serengil & Ozpinar, 2021; Rezgui, 2019) each face based on gender (man/woman), race (asian, black, hispanic, indian, and white), and age (young, middle-aged, old). A total of 900 images were randomly selected constituting 30 images from each of 30 demographics (2 [gender] $\times$ 5 [race] $\times$ 3 [age]).

These real images were used as input to the image-to-image synthesis pipeline of Stable Diffusion (v.2.1) (Rombach et al., 2022) to generate 900 images consistent with the demographic prompt "a photo of a [age] [race] [gender]." The strength parameter was set to 0.8 and the number of inference steps to 250, with all other parameters left at their defaults (a strength of 0.0 reproduces the input image and a strength of 1.0 effectively the input image). Shown in Figure 1 are examples of real (top) and generated faces (bottom). As described next, these 900 generated faces were used to seed the iterative model retraining. We used this image-to-image synthesis instead of the unconstrained text-to-image generation to balance the comparison of model retraining on healthy data and self-generated data.

2.2 EVALUATION

A standard Fréchet inception distance (FID) (Heusel et al., 2017) and Contrastive Language-Image Pre-training score (CLIP) (Radford et al., 2021; Hessel et al., 2021) are used to assess the quality of generated images. These metrics are commonly used in text-to-image model benchmarks.

The FID is used to quantify the realism of generated images by comparing the set of generated images to matched real images. This metric compares two sets of N images – one real and one AI-generated – with a $1:1$ prompt correspondence. The FID is calculated as the Fréchet distance between the mean and covariance matrices of the corresponding images' Inception-v3 embeddings (Szegedy et al., 2015). A smaller FID corresponds to a higher similarity between the AI-generated and corresponding real images. In our case, the reference image set consists of 820 real faces from the FFHQ dataset. This is less than the full set of 900 because the FFHQ dataset is not sufficiently diverse across all demographic categories. We compare 820 generated images, consisting of the maximum number of images per 30 demographic groups, to this set of 820 demographically similar real images.

The CLIP score calculates the cosine similarity between the image and caption embeddings. This metric relies on the latent space of a pre-trained CLIP model to gauge whether the generated image is semantically consistent with the caption. In our case, the CLIP score is evaluated across all 900 generated images using the CLIP ViT-H/14 model trained with the LAION-2B English subset of LAION-5B[2]. A larger CLIP score corresponds to higher semantic coherence.

[2] https://huggingface.co/laion/CLIP-ViT-H-14-laion2B-s32B-b79K

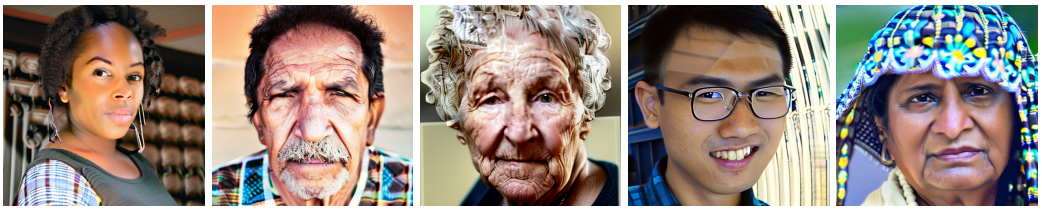

Figure 2: Examples of low-quality generated images that are replaced in the retraining control experiment.

2.3 POISIONING

We define model self-poisoning as the process of retraining a generative model on data with varying amounts of images generated by a base model. Unlike adversarial training (Carlini et al., 2023), the training image captions are consistent with the image contents.

The architecture of Stable Diffusion comprises three main modules: CLIP text encoder, U-Net, and variational autoencoder (VAE). The CLIP text encoder converts the user text prompt into a CLIP embedding, a high-dimensional representation in a multimodal vector space. The base SD model does not train a new CLIP text encoder but instead uses a pre-trained model with frozen weights. The CLIP embedding is provided as input into the U-Net module. The U-Net, guided by a scheduler, gradually denoises what was originally a random noise image towards a representation that is semantically consistent with the text embedding. At the end of this diffusion process, the VAE converts the latent U-Net image representation into pixel space to yield a final image.

The model retraining proceeds as follows. Leaving the CLIP text encoder and VAE modules intact, we retrained the denoising U-Net module of the base Stable Diffusion model (SD v.2.1) using the recommended parameters (a constant learning rate of 2×10^{-6}, 50 epochs, 512×512 image resolution, no mixed precision, and random horizontal flip augmentation). The rationale for this is that these modules handle supportive tasks – the conversion from text to the latent space and from the latent space back to the pixel space – outside of the core diffusion process. Freezing of these modules is standard practice for retraining. The model was initially retrained on the 900 SD-generated images and demographic captions. Another 900 images with the same demographic prompts were generated using this retrained model. These images were then used to retrain the model. This process was repeated for a total of five iterations.

This entire process was repeated with different compositions of faces ranging from the above 100% SD-generated and 0% real faces to a $50\%/50\%$, $25\%/75\%$, $10\%/90\%$, $3.3\%/96.7\%$, or $0\%/100\%$ (the odd-ball 3.3% composition corresponds to one of the 30 images per demographic group being SD-generated with the other 29 real).

2.4 HEALING

We define model healing as the process of retraining a generative model on only real images. The goal is to determine whether a self-poisoned model can be healed to produce images similar to the base model. This healing process proceeds in the same way as the self-poisoning described in Section 2.3 except that the images are real, not AI generated.

2.5 CONTROLS

We carried out two control experiments to determine the impact of our iterative retraining of the SD model (Section 2.3). Having noticed that, compared to real images, SD-generated images tend to be of higher contrast and saturation, we wondered if these color differences would impact the iterative retraining. In this first control, the color histogram of each generated image is matched to a real image. Each generated image is histogram matched – in the three-channel luminance/chrominance (YCbCr) space – to a real image with the most similar color distribution (measured as the image with the minimal Kullback-Leibler divergence averaged across all three YCbCr channels). This histogram matching is performed on each retraining iteration.

Figure 3: Examples of images generated by the baseline version of Stable Diffusion (prompt: "older hispanic man").

We also noticed that among the 900 generated images there are occasional images with obvious artifacts, including misshapen facial features, as shown in Figure 2. In this second control, we removed from the retraining dataset any generated image with a single-image FID score (Shaham et al., 2019) greater than the mean single-image FID between the first batch of 900 generated images and their corresponding 900 real images. This culling is performed on each retraining iteration.

3 RESULTS

Shown in Figure 3 are five representative images generated from the baseline Stable Diffusion model for a single demographic group ("older hispanic man"). Generally speaking, images from the baseline model are consistent with the text prompt and of high visual quality.

Shown in the first row of Figure 4 are representative images generated from iterative retraining of the baseline SD model on images of real faces taken from the FFHQ facial dataset. The generated images are semantically consistent with the text prompts, exhibit the prototypical alignment property of the faces in the FFHQ dataset, and show no signs of distortion. Shown in Figure 5 are the FID and CLIP scores for the full set of generated images (labeled 0%), from which we see that iterative retraining on real images causes no degradation in the resulting model.

Also shown in the top portion of Figure 4 are representative images generated from iterative retraining of the baseline SD model with a different mixture of self-generated and real images. Regardless of the mixture ratio, the iterative retraining eventually leads to collapse by the fifth iteration, at which point the generated images are highly distorted. This model collapse can be seen qualitatively by the appearance of the generated images and quantitatively by the significant deviation of the FID and CLIP score from the baseline model. This collapse is apparent in Figure 5 where we see that after a small improvement in quality on the first iteration, both the FID and CLIP scores reveal a significant degradation in image quality (a high FID and a low CLIP correspond to lower quality images).

The filled plot symbols (diamond and square) in Figure 4 correspond to the two control conditions in which the retraining dataset is color matched to real images (diamond) and any low-quality generated images are replaced with high-quality generated images prior to retraining (square). Even these curated retraining datasets lead to model collapse at the same rate as the other datasets.

In addition to the degradation in image quality, and consistent with previous reports (Hataya et al., 2023), we also note that model self-poisoning leads to a lack of diversity in terms of the appearance of the generated faces. This can be seen in Figure 4 where, particularly when the self-poisoning is greater than 10%, the generated faces are highly similar across the latter iterations.

Shown in the lower two rows of Figure 4 are representative examples of images generated with text prompts distinct from the demographic prompts used in the model retraining. The text prompts used to generate the images in the lower panel of Figure 4 are: "A dog walker training a dog amidst a field of sunflowers"; "A football team grilling hamburgers at a snowy ski resort"; "A group of skateboarders practicing martial arts in a mysterious foggy landscape"; "A marine biologist scuba diving amidst a field of sunflowers"; and "An elderly woman writing in a journal in a charming village square". The images generated by the model retrained on entirely real images (0%) produce semantically coherent images with no obvious visual artifacts. The images generated by the model retrained on 25% SD-generated faces often exhibit the same textured artifacts as seen in the faces in

Figure 4: Examples generated after iterative retraining for different compositions of the retraining dataset: 0% SD-generated and 100% real to 100% SD-generated faces and 0% real. Shown in the lower panel are examples generated with text prompts distinct from those used in the retraining.

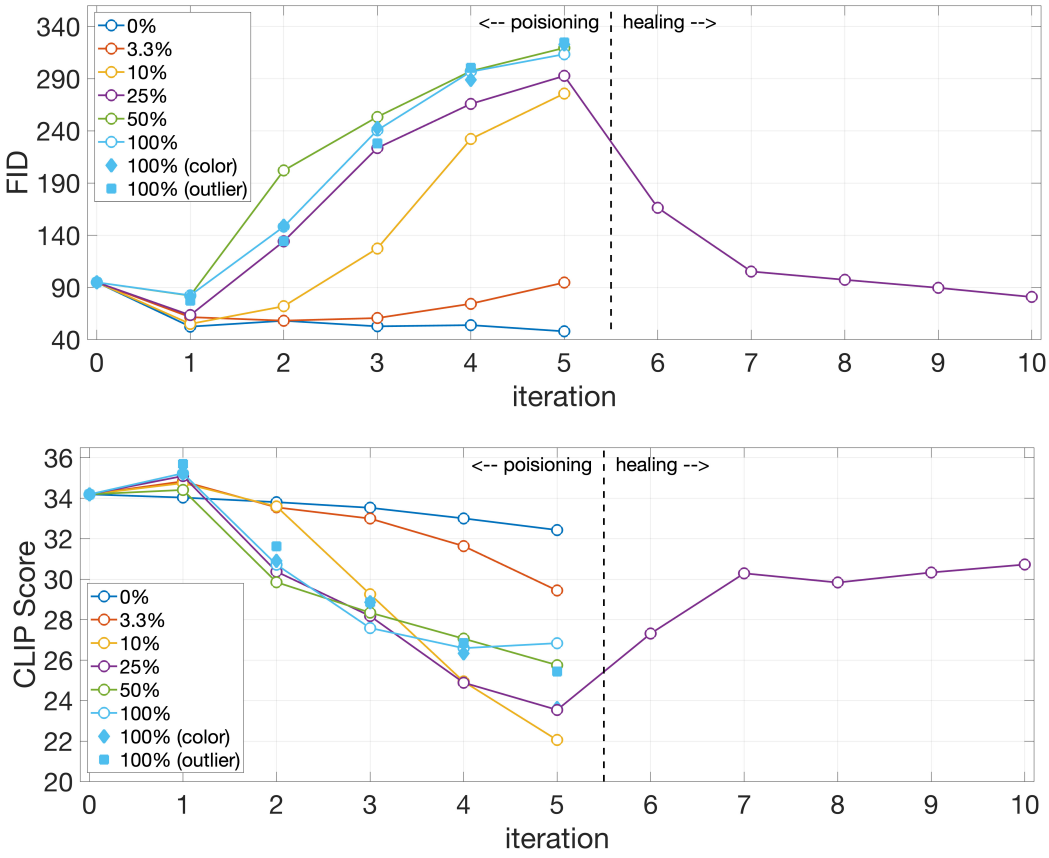

Figure 5: Shown are the FID and CLIP as a function of the number of retraining iterations and the composition of the retraining dataset ranging from 100% SD-generated faces and 0% real faces to 0% SD-generated and 100% real ("poisoning"). The diamond plot symbol corresponds to the 100%/0% condition in which the retraining dataset is color matched to the real faces. The square plot symbol corresponds to the the same condition in which the retraining dataset was curated on each iteration to remove low quality faces. The trend is the same for both metrics: the presence of generated faces leads to a degradation in quality across iterations (a higher FID and a lower CLIP correspond to lower image quality). Also shown is the FID and CLIP score for the 25% model retrained on an additional five iterations (6-10) on only real images ("healing").

the upper portion of this figure. This means that the model self-poisoning is not limited to a specific category of images used in the retraining but seems to impact the model more broadly.

Lastly, we wondered if, once self-poisoned, the model could be "healed" by retraining on only real images.. The model self-poisoned for five iterations with 25% SD-generated images was retrained for another five iterations on only real images. Shown in Figure 6 are representative examples of faces generated from five different demographic groups across these five additional iterations. Although in some cases, by the last iteration, the generated faces have fewer artifacts, in other cases, the artifacts persist. Shown in the right portion of Figure 5 are the FID and CLIP scores for these healing iterations in which we see that the FID recovers to the original base model and the CLIP score almost recovers to base model levels.

Although the mean FID and CLIP score recovers, we clearly see remnants of the self-poisoning in some of the faces in Figure 6. This larger variation is evident in the standard deviation of the CLIP score, which is 2.8 for the base model (with a mean of 35.1) but is 4.2 for the healed model after five iterations (with a mean of 27.3). It appears that the model can partially – but not entirely – heal.

Figure 6: Examples generated after retraining of the 25% self-poisoned model with only real images. In some cases, the self-poisoning persists, while in others the model appears to partially heal itself.

4 Discussion

We find that at least one popular diffusion-based, text-to-image generative-AI system is surprisingly vulnerable to data poisoning with its own creations. This data poisoning can occur unintentionally by, for example, indiscriminately scraping and ingesting images from online sources. Or, it can occur from an adversarial attack where websites are intentionally populated with poisoned data, as described in (Carlini et al., 2023). Even more aggressive adversarial attacks can be launched by manipulating both the image data and text prompt on as little as 0.01% to 0.0001% of the dataset (Carlini & Terzis, 2021).

In the face of these vulnerabilities, some reasonable measures could be taken to mitigate these risks. First, there is a large body of literature for classifying images as real or AI-generated (e.g., Zhang et al. (2019); Chai et al. (2020); Gragnaniello et al. (2021); Liu et al. (2022); Corvi et al. (2023)). An ensemble of these types of detectors could be deployed to exclude AI-generated images from being ingested into a model's retraining. A complementary approach can automatically and robustly watermark all content produced by a model. This can be done after an image is generated using standard techniques (Cox et al., 1999) or can be baked into the synthesis by watermarking all the

training data (Yu et al., 2021). Lastly, more care can be taken to ensure the provenance of training images by, for example, licensing images from trusted sources.

These interventions are, of course, not perfect. Passive detection of AI-generated images is not full proof: a sophisticated adversary can remove a watermark, and provenance is not always available or completely reliable. Combined, however, these strategies will most likely mitigate some of the risk of data poisoning by significantly reducing the number of undesired images.

Open questions remain. What about the model or training data causes the data poisoning? Will data poisoning generalize across synthesis engines? Will, for example, Stable Diffusion retrained on DALL-E or Midjourney images exhibit the same type of model collapse? Can generative-AI systems be trained or modified to be resilient to this type of data poisoning? If it turns out to be difficult to prevent data poisoning, are there specific techniques or data sets that can accelerate healing?

REFERENCES

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G Baraniuk. Self-consuming generative models go mad. arXiv:2307.01850, 2023.

Nicholas Carlini and Andreas Terzis. Poisoning and backdooring contrastive learning. arXiv:2106.09667, 2021.

Nicholas Carlini, Matthew Jagielski, Christopher A Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. Poisoning web-scale training datasets is practical. arXiv:2302.10149, 2023.

Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? Understanding properties that generalize. In *European Conference on Computer Vision*, pp. 103–120, 2020.

Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5. IEEE, 2023.

Ingemar J Cox, Matt L Miller, JMG Linnartz, and Ton Kalker. A review of watermarking principles and practices. *Digital Signal Processing for Multimedia Systems*, 2:461–482, 1999.

Diego Gragnaniello, Davide Cozzolino, Francesco Marra, Giovanni Poggi, and Luisa Verdoliva. Are GAN generated images easy to detect? A critical analysis of the state-of-the-art. In *IEEE International Conference on Multimedia and Expo*, pp. 1–6, 2021.

Ryuichiro Hataya, Han Bao, and Hiromi Arai. Will large-scale generative models corrupt future datasets? In *IEEE/CVF International Conference on Computer Vision*, pp. 20555–20565, 2023.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. arXiv:2104.08718, 2021.

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. *Advances in Neural Information Processing Systems*, 30, 2017.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8110–8119, 2020.

Bo Liu, Fan Yang, Xiuli Bi, Bin Xiao, Weisheng Li, and Xinbo Gao. Detecting generated images by real images. In *European Conference on Computer Vision*, pp. 95–110. Springer, 2022.

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:2303.01255, 2023a.

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:2306.06130, 2023b.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171. PMLR, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Zohra Rezgui. Détection et classification de visages pour la description de l'égalité femme-homme dans les archives télévisuelles, 2019.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

Sefik Ilkin Serengil and Alper Ozpinar. HyperExtended lightface: A facial attribute analysis framework. In *International Conference on Engineering and Emerging Technologies*, pp. 1–4. IEEE, 2021.

Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Singan: Learning a generative model from a single natural image. In *IEEE/CVF International Conference on Computer Vision*, pp. 4570–4580, 2019.

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:2305.17493, 2023.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. WIT: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pp. 2443–2449, 2021.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, 2015.

Xiaodan Xing, Fadong Shi, Jiahao Huang, Yinzhe Wu, Yang Nan, Sheng Zhang, Yingying Fang, Mike Roberts, Carola-Bibiane Schönlieb, Javier Del Ser, et al. When AI eats itself: On the caveats of data pollution in the era of generative ai. 2024.

Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In *IEEE/CVF International Conference on Computer Vision*, pp. 14448–14457, 2021.

Nonghai Zhang and Hao Tang. Text-to-image synthesis: A decade survey. 2024.

Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in GAN fake images. In *IEEE International Workshop on Information Forensics and Security*, pp. 1–6, 2019.