Adversarial Perturbations Cannot Reliably Protect Artists From Generative AI

Anonymous Authors[1]

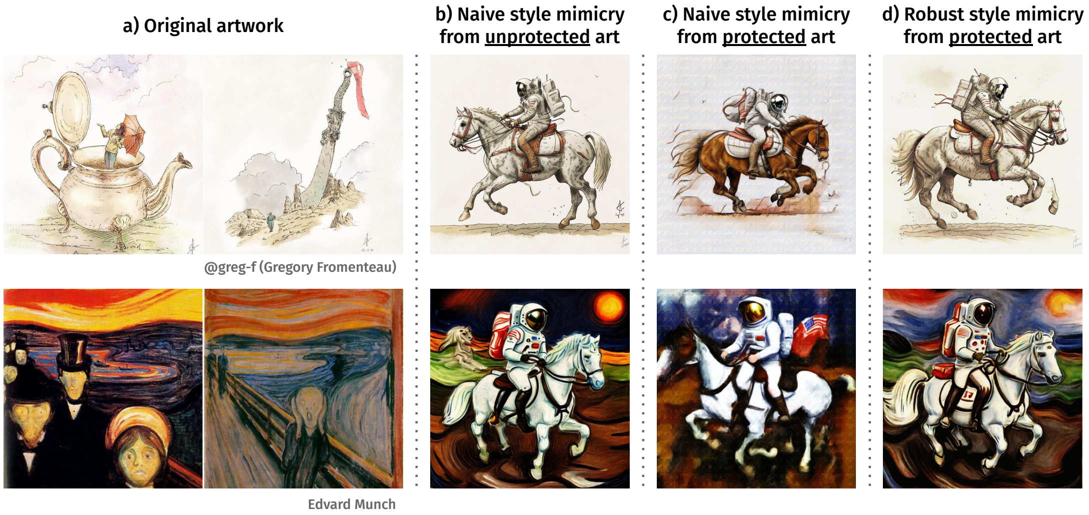
a) Original artwork
b) Naive style mimicry from unprotected art
c) Naive style mimicry from protected art
d) Robust style mimicry from protected art
@greg-f (Gregory Fromenteau)
Edvard Munch

Figure 1. Style mimicry protections provide a false sense of security. While they can prevent naive mimicry (c), we show that simple methods can achieve style mimicry from protected art (d) that is indistinguishable from mimicry without protections (b).

1. Introduction

Style mimicry is a popular application of modern text-to-image generative models. Given a few images with a specific style, a model can be finetuned to generate new images in that style (e.g., a spaceship in the style of Van Gogh). But style mimicry has the potential to cause significant harm if misused. Many contemporary artists worry that others could now produce images that copy their unique art style, and potentially steal away customers (Heikkilä, 2022). As a response, several protections have been developed to protect artists from style mimicry (Shan et al., 2023a; Van Le et al., 2023; Liang et al., 2023). These protections add adversarial perturbations to images that artists publish online, in order to inhibit the finetuning process. These protections have received significant attention from the media—with features in the New York Times (Hill, 2023), CNN (Thorbecke, 2023) and Scientific American (Leffer, 2023)—and have been downloaded over 1.5M times (Shan et al., 2023a).

In this work, we show that state-of-the-art style protection tools—*Glaze* (Shan et al., 2023a), *Mist* (Liang et al., 2023) and *Anti-DreamBooth* (Van Le et al., 2023)—are ineffective when faced with simple *robust mimicry methods*. Our robust mimicry methods only require combining standard image processing techniques *which already existed at the time that these protection tools were first introduced!* We validate our results with a user study, which reveals that robust mimicry methods can produce results indistinguishable in quality from those obtained from unprotected artworks (see Figure 1 for an illustrative example).

We argue that existing protection tools provide a false sense of security and their limitations are inherent. Artists are necessarily at a disadvantage since they have to act first (i.e., once someone downloads protected art, the protection can no longer be changed). To be effective, protective tools face the challenging task of creating perturbations that transfer to *any* finetuning technique, even ones chosen adaptively in the future. We thus caution that *adversarial machine learning techniques will not be able to reliably protect artists from generative style mimicry*, and urge the development of non-technological protective measures, such as stronger legislation or community norms, to protect artists.

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

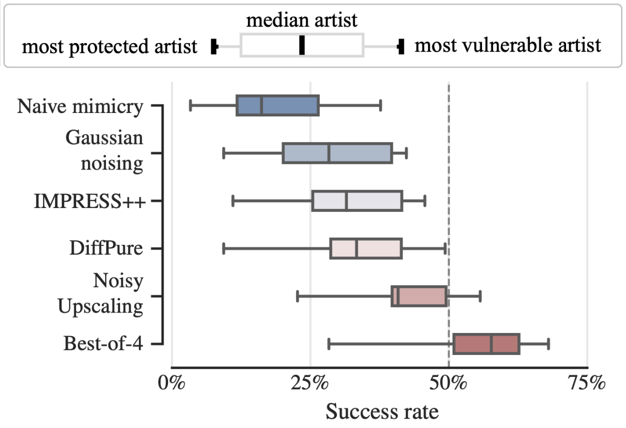

Figure 2. **Success rate** of our robust mimicry methods for 10 artists averaged across protections. 50% indicates perfect robust mimicry.

2. Robust Style Mimicry

We assume two parties, the *artist* who places art online, and a *forger* who performs style mimicry using these images. The challenge for the forger is that the artist *protects* their art, using a state-of-the-art protection tool. If a mimicry method can emulate an artist's style using only *protected* artwork, we call it *robust*. We consider style forgers who finetune a model on an artist's images—the most successful style mimicry method to date (Shan et al., 2023a).

Although mimicry protections were originally evaluated against methods that forgers could use for rubst mimicry, we note a number of limitations in these methods and their evaluation in Section 2.1. We then address these limitations by proposing our own methods (Section 2.2) and evaluating them across all protections (Section 3).

2.1. Limitations of Prior Robust Mimicry

We highlight 3 main limitations of prior attempts to evaluate protections against robust mimicry methods. (1) We found that some protections' effectiveness depend on a specific experimental setup and do not generalize to different finetuning hyperparameters that an attacker may choose. (2) Existing robust mimicry attempts designed to evaluate protections are sub-optimal. Our methods show that slight variations of methods that were dismissed by previous work can result in very successful robust mimicry. (3) Comparing the robustness of protections and mimicry from previous work is challenging since they use different experimental setup and metrics. We perform a *unified evaluation on a common set of prompts and rely only on human evaluations*.

2.2. Our Robust Mimicry Methods

We design *four* robust mimicry methods. We prioritize *three* simple methods that only require *preprocessing* protected images (Gaussian Noising, DiffPure, and Noisy Upscaling). These methods present a higher risk because they are more accessible, and can be used in black-box scenarios. For completeness, we further propose *one* white-box method (IMPRESS++), inspired by IMPRESS (Cao et al., 2024).

3. Results

Evaluation setup. To measure the success of each style mimicry scenario, we rely only on human evaluations since previous mimicry work found automated metrics (e.g., using CLIP (Radford et al., 2021)) to be unreliable (Shan et al., 2023a;b). Our user study asks annotators to compare image pairs, where one image is generated by a robust mimicry method, and the other by a model finetuned directly on *unprotected* art. We define *success rate* of a robust mimicry method as the percentage of annotators who prefer outputs from the robust mimicry method over those from *unprotected* art. A perfectly robust mimicry method would generate images indistinguishable from those generated from unprotected art and thus obtain a success rate of 50%. Figure 2 includes a summary of the success rate for each robust mimicry method averaged over all protections.

All protections are easily circumvented. The main finding is that *all protective tools create a false sense of security and leave artists vulnerable to style mimicry*. Noisy upscaling is the most effective method for robust mimicry, with a median success rate above 40%. A style forger does not have to choose a single robust mimicry method beforehand, but can test all of them and select the most successful. We define "best-of-4" as a forger who selects the best method for each prompt and artist. "Best-of-4" almost always beats the generations from unprotected art.

Protections are highly non-uniform across artists. Since artists cannot *check* how vulnerable they are, these tools still provide a false sense of security for all artists. Protections should hold for *all* artists alike, while a mimicry method might successfully target only specific artists.

Adversarial perturbations will not protect artists from mimicry. The goal of our research is to help artists better decide how to protect their artwork and business. We do not focus on creating the *best* mimicry method, but rather on highlighting limitations in popular perturbation tools. We argue that having *no* protection tools is preferable to having *insecure* ones. Insecure protections may mislead artists to believe it is safe to release their work, putting them in a worse situation than if they had been more cautious in the absence of protection. We thus caution that *adversarial machine learning techniques will not be able to reliably protect artists from generative style mimicry*, and urge the development of non-technological protective measures, such as stronger legislation or community norms, to protect artists.

References

Cao, B., Li, C., Wang, T., Jia, J., Li, B., and Chen, J. Impress: Evaluating the resilience of imperceptible perturbations against unauthorized data usage in diffusion-based generative ai. *Advances in Neural Information Processing Systems*, 36, 2024.

Heikkilä, M. This artist is dominating ai-generated art. and he's not happy about it. *Technology Review*, 2022.

Hill, K. This tool could protect artists from ai-generated art that steals their style. *The New York Times*, 2023.

Leffer, L. Your personal information is probably being used to train generative ai models. 2023.

Liang, C., Wu, X., Hua, Y., Zhang, J., Xue, Y., Song, T., Xue, Z., Ma, R., and Guan, H. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. In *International Conference on Machine Learning*, pp. 20763–20786. PMLR, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Shan, S., Cryan, J., Wenger, E., Zheng, H., Hanocka, R., and Zhao, B. Y. Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 2187–2204, 2023a.

Shan, S., Wu, S., Zheng, H., and Zhao, B. Y. A response to glaze purification via impress. *arXiv preprint arXiv:2312.07731*, 2023b.

Thorbecke, C. It gave us some way to fight back: New tools aim to protect art and images from ai's grasp. 2023.

Van Le, T., Phung, H., Nguyen, T. H., Dao, Q., Tran, N. N., and Tran, A. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2116–2127, 2023.