Twotechniques of Preventing Unauthorized Misappropriation of works of art–Mist And Glaze, do they truly work?

Minyuan Zhu
minyuanzhu@umass.edu
University of Massachusetts Amherst
Amherst, MA, USA

ABSTRACT

Recently, the pervasive integration of AI-based platforms into everyday life has made advanced technologies accessible even to preschool children. Despite their growing utility, the rapid expansion of AI tools raises profound concerns about copyright violations, the protection of intellectual property, and related legal matters, especially in the realms of visual arts and music. While many AI models assert the use of authorized datasets, none can guarantee complete compliance with copyright laws, leaving a significant risk of misuse. This paper discusses some open-source techniques that individuals can utilize to safeguard their creative works and private photographs from potential copyright infringements.

1 INTRODUCTION

Recently, there has been an explosion of AI-based platforms entering into our daily lives. Even a senior who just learned how to use computers or a preschool child can understand how to download and use AI-supported platforms.
According to the statistics from [17]. More than 15 billion images created using text-to-image algorithms since 2022, if we compare it with the number of photographs, it takes photographers more than 150 years to do that. Since the launch of DALLE-2, people have created an average of 34 million images per day. And Midjourney has over 15 million users. Approximately 80 % of the images were created using models, services, platforms, and applications based on an open-source model, Stable Diffusion.

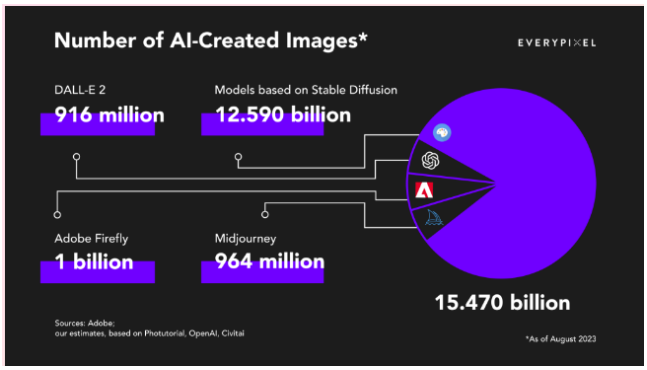

Figure 1: Data from [17], it shows the usage of different AI models

This increasing use rate of AI tools is significant, more and more people are getting used to AI techniques. According to the data from [9], by the sample of survey, over 56 % of businesses are using AI techniques to help them improve their business. However, this increasing use rate has caused strong concern about the security and privacy of AI techniques. Do AI companies or individuals who are training AI models use a dataset that was not authorized by the copyright owner? Is it possible that individual photos shared on the social platform might be used in the training of AI? This is worthy of consideration. OpenAi did talk about how their dataset came from [10], said: "As noted above, ChatGPT and our other services are developed using (1) information that is publicly available on the internet, (2) information that we license from third parties, and (3) information that our users or human trainers provide." Although they announced where their dataset comes from, this still does not solve the concerns about the security and privacy of AI. Yes, the dataset is from publicly available websites or platforms, but the companies did not ask the owner of the information or pictures if they were willing to do so.

Moreover, the "dark industry" is also using AI to do illegal things behind the internet.[14] A Florida man accused of using AI to create child porn and [15] a Stanford Internet Observatory (SIO) investigation identified hundreds of known images of child sexual abuse material (CSAM) in an open dataset used to train popular AI text-to-image generation models, such as Stable Diffusion, deepfake tools, etc. This is terrible if we do not have techniques to prevent it, it could happen just around you and you could even never know it.

Thus, to solve these kinds of concerns, in this paper, we are going to talk about various techniques that can be implemented in your original paintings, pictures, photos, and music.

The ideas of these techniques are similar, but the actual implementations are different. Although most of the techniques we are going to talk about in this paper are about adding "something" into your copyright things, this "something" here is different. Some of them are similar to the traditional ways of security implementation like digital signature, some of them change the pixel of the pictures so that the AI model can not correctly distinguish it and use it to train. All of them have advantages and disadvantages, we will also talk about in what situation it is better to use what techniques to protect your copyright and privacy.

2 BACKGROUND

In this section, we will talk about some background information and knowledge. It is useful and worthy to read this section before entering the actual techniques and examples. We will briefly talk about how AI is being trained, how AI generates the output pictures, and basic ideas about some famous AI models

Figure 2: A highschool life image| created by Dall-E[1]

Figure 3: A highschool life image| created by Stable-Diffusion

2.1 How AI being trained?

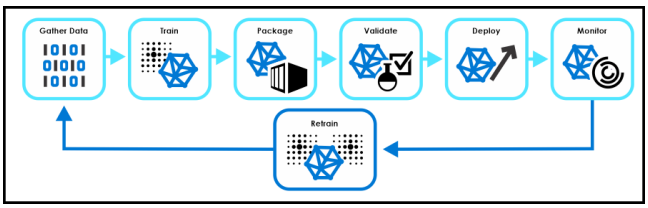

Figure 4: AI training process from Sumit Singh[13]

The first step of training an AI model is collecting the dataset, which is the part we are concerned about. Many companies and labs use the dataset from some dataset companies, and although they claim all data are legal and authorized, it's not fully true.

The second step is processing the training data, for our topics, pictures, and music, usually, trainers need to cut the pictures and music into the same length and label the styles of each picture and music. Of course, there is more processing needed to be implemented into the data, we are not going to specifically talk about them here.
The third step is choosing the model and designing the architecture. Usually, in AI painting, people will choose a stable diffusion model to train their new model, since stable diffusion is open-source and easy to be used.
The fourth step is training. We need to set some hyperparameters such as learning rate, batch size, number of training epochs, and others. These need to be tuned based on your specific dataset and training goals.
The fifth step is monitoring the training to see if the training satisfies your expectations
Then after training, you will need to decide using the model or retrain based on the performance.
The key part is the first step, the quality of the dataset decides lots of things in training AI models, if the dataset has some polluted or toxic data, it will destroy the entire model(which is one technique we will discuss in the following section). To make sure the dataset has a high quality, lots of people chose to use the data from social media since most of the pictures and sentences in social media are sent by real humans, if the quality of the pictures is low, people tend to not sending them to the social media. Here is our concern about copyrighted pictures and paintings. Lots of artists are nice such that they post their original paintings on some art social platform, and adversarials can just select the highest popular paintings and download them, Compared to the hard work of artists, the cost of downloading the pictures and violating the copyright is just one sand in the desert.

2.2 How AI generate the output pictures?

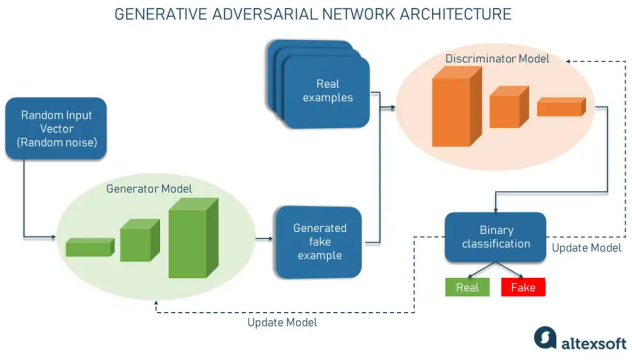

Figure 5: How AI generate a picture from altexsoft[2]

AI models, particularly those designed for image generation like GANs (Generative Adversarial Networks) or VAEs (Variational Autoencoders), learn from vast amounts of data to create new images. During training, these models don't gain an understanding of objects as humans do but instead learn to detect and replicate patterns found in the training dataset. For instance, if the input

dataset contains different kinds of cars, AI will learn what's the features of cars, for example, cars usually has four wheels, then it comprehends a complex array of features that typically define cars in the dataset it was trained on.

These features are mathematical and involve high-dimensional data representations that the model uses to generate images. When tasked with creating a new picture, the AI does not literally stitch together parts from existing images. Instead, it synthesizes new content by adjusting and combining these learned features in novel ways, guided by the statistical patterns it has internalized.

If you ask the model to generate a specific image, say of a cat, it doesn't just piece together parts of previous cat images. Instead, it employs the underlying "concept" of a cat as defined by the training data to produce a new image that fits the description or matches an input picture. This process is akin to what is humorously referred to in some creative communities as "suturing body parts"—the AI composes a new image from the 'corpse' of its dataset, though not through direct assembly but through a sophisticated recombination of learned features.

While it might seem like AI can originate completely new creations, its ability to generate images is heavily dependent on the style, quality, and diversity of its training data. The output often mirrors the characteristics of the dataset, making the AI's "creativity" constrained by what it has been exposed to during training. This highlights a critical point: while AI can produce work that feels novel and unique, it is always reflecting some aspect of its training environment. This relationship underscores the importance of the dataset quality and diversity, which fundamentally shapes what the AI can generate.

2.3 Laws about AI art

According to U.S. District Judge Beryl Howell [5] only works with human authors can receive copyrights, which means the artworks generated by AI can not be considered as "creative" since it has more "machine" than human work.

However, using other people's copyrighted art to train AI is hard to judge since it's hard to find if the company or individual truly used the copyrighted art to train their model. According to the Chip Law Group[4], some groups of artists suspected some companies of using their artwork to train AI and sued some companies, but the result hasn't come out and lots of experts in the law industry thought the result might be bad for the artists.

2.4 People's opinion about AI art and stealing other people's images to train ai

Within the artistic community, many find the use of AI to generate paintings quite controversial. On a popular social media platform, numerous accounts are dedicated to identifying AI-generated art, particularly criticizing the use of copyrighted materials without permission for training these models.[6]

Conversely, the sentiment within the computer science community tends to be more neutral regarding the use of others' images to train AI. Responses to the question "Why do so many people criticize AI-generated paintings?" on Zhihu[16] reveal that most disagree with the criticism. Many respondents, who identify as enthusiasts of computer science, express support for AI in art even using the unauthorized data sets, "Preventing people using your paintings to train AI is like destroying the railway tracks after the invention of trains" said by a zhihu user.[**?**]

3 MIST2.0

In this section, we will introduce the first technique you can use to protect your paintings from stealing by others. This work is done by these excellent researchers from different institutions.[8].

3.1 Background information about Generative Modeling and Mist

A generative model learns from data $x \sim p(x)$ and holds a distribution $p_\theta(x)$ such that the generated data can be sampled. The generative models based on the latent variables have proven effective in generative tasks. In other words, basically generative model has a dataset that it can learn from and it will generate the output based on the input dataset which has a distribution under$p_\theta(x)$. The goal of Mist is to attack the generative model. Mist used adversarially attacking generative diffusion models to attack the model. This can make sure the model can not correctly distinguish the input pictures without much harm to the quality of the original pictures.

Thus, Mist used a specific algorithm to add imperceptible noise to the original pictures, it is like a very light glaze on the surface of the picture, and human eyes are hard to find it. Of course, this algorithm is based on the stable diffusion model, when doing the process we still need to use the characteristics of the stable diffusion model. Actually, we use the stable diffusion model to process the picture by changing some sets.

Mist mainly used two tricks to reach the goal, the first trick is combining two terms of existing adversarial loss in an effective approach and the second trick is picking a compatible target image for generating targeted adversarial examples.

In the first trick, the researchers merged the semantic loss and textual loss together into a joint loss function.

For the second trick, the researchers found that the input images should have high contrast ratio, by doing this, it can improve the effectiveness of adversarial examples and its robustness against noise purification.

The researchers also compared Mist with Glaze, they found Mist can remain highly robust even after using techniques like screenshots to try to decrease the efficiency of Mist, but Glaze can not. The verification and proof is included in the paper done by[7].

3.2 Experiment of Mist2.0

These pictures are from some famous artist, and their paintings have already into the Public Domain. We use Mist to process the pictures and compare the new images with the original ones.

The set ups are Epochs:5,LoRA Steps:10,Attacking Steps:30,The learning rate of LoRA:0.0001,The learning rate of PGD:0.005,LoRA Ranks:4,The weight of prior loss:0.1,The weight of vae loss:0.00001. It takes about 3 minutes to process a single picture under the GPU: RTX 3070ti. Which is a fair running time.

As we can see, by using Mist to process the pictures, it's hard to tell the difference between the original pictures and the processed pictures, although we do have the feeling that something changed in the pictures, it's like a glaze. By using the processed pictures,

Figure 6: The original version of Staryy Night by Van Gogh

Figure 7: The original version of Sunflower by Van Gogh

Figure 8: The original version of Het meisje met de parel by Johannes Vermeer

Figure 9: Processed version of Sunflower by Van Gogh(mist)

Figure 10: Processed version of Starry Night by Van Gogh(mist)

Figure 11: Processed version of Het meisje met de parel by Johannes Vermeer(mist)

the AI model can not fully correctly label the images, although the generated images do have some same features like colors, the main features of the original images can not be labeled and imitated. However, Mist has two significant disadvantages. First, the size of

Figure 12: Using processed version(mist) of Sunflower to generated the picture under Stable Diffusion official model1.5

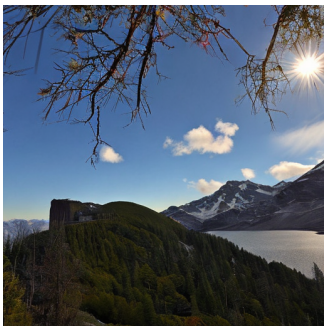

Figure 13: Using processed version(mist) of Starry Night to generated the picture under Stable Diffusion official model1.5

the original picture can only be 512×512 pixels, if the input size does not satisfy the condition, the algorithm will automatically change it into 512×512 pixels. Second, although it is hard for human eyes to distinguish the difference between the original images and the processed ones, it's still possible to find the difference, and due to the difference implemented by Mist, the processed images may not be accepted by some artists.

4 NIGHTSHADE AND GLAZE

Nightshade[12] and glaze[11] are two different techniques invented by the team of the University of Chicago. The difference between nightshade and glaze is nightshade is used to attack the model while glaze is used to defend, like Mist.

Figure 14: Using processed version(mist) of Het meisje met de parel to generated the picture under Stable Diffusion official model1.5

4.1 Background information about nightshade and glaze

The nightshade is computed as a multi-objective optimization that minimizes visible changes to the original image, human eyes are hard to see the changes. However, by doing this, we can offend the AI model to make the model not correctly label and distinguish the pictures. This is like poisoning, the goal isn't to let the model not correctly distinguish your pictures' elements and features but to let the model not work normally.

The design of Glaze is complicated, here is a simpler explanation, for full proof and design please check their paper.

Glaze has some unique functions to compute the style cloaks of the image, and there are some precomputed style cloaks in the model. When a user gives an input image, it will first compute the style cloaks and randomly choose another style's cloak to replace the original input picture's cloaks. By changing the cloaks, other AI models will misunderstand the style(feature) of the images and can not correctly mimic the style of the images.

4.2 Experiments

Still, we are using previous pictures from famous artists that the copyright has in the public domain.

Unfortunately, the download link of Nightshade is closed, we have to only do glaze for now.

As we can see, the glaze has one more advantage than mist, the glaze does not limit the input picture's size, and the output picture's size is the same as the original one. However, we are not going to do the generated experiments here, because Glaze has already been cracked, see section 5.1 to learn more.

Figure 15: Processed version of Starry Night by Van Gogh(glaze)

Figure 16: Processed version of Sunflower by Van Gogh(glaze)

Figure 17: Processed version of Het meisje met de parel by Johannes Vermeer(glaze)

5 DISCUSSION

From previous sections, we can see the result of these techniques to prevent unauthorized stealing pictures to train AI.
We can see that the results of these experiments are significant, however, this is under the ideal situations. Although most of the time, these techniques do have positive results, we can not forget the real world is not the experiment. One common attack on these techniques is just using screenshot tools to take a screenshot of the original pictures, in this case, the techniques can not work as fine as the experiment situation.
Other techniques like digital signature and also be successfully attacked by screenshots or blur the original pictures.
Moreover, the reason techniques like mist and glaze can work is they have some specific algorithm to process the input pictures by changing the position of pixels or changing a small amount of pixel's color to let AI models be confused such that they can not correctly label the input pictures. However, this is not always the ideal situation, sometimes these small changes can be observed by human eyes which will affect the quality of the original pictures.
Furthermore, although these techniques seem very useful, just a few days after the release of Glaze, one famous AI field person, the inventor of ControlNet wrote 15 lines code to crash the Glaze, by using these 15 lines of code, the process of Glaze became useless. Even without using these codes, the influence of Glaze and Mist is far smaller than the researchers expected.
Not only other models, but also different variants of Stable Diffusion can still correctly label and distinguish the images after being processed by these two techniques. AI technology improves much faster than we expected, we need more researchers to focus on the copyright questions and invent more powerful, reliable techniques to prevent unauthorized stealing.

6 RELATED WORK

There are also some other related works about AI-generated images. We can not only protect the images from being stolen but also track the images and check whether the images are generated by AI.
In this paper[18], the author talked about a reverse engineering method to check if the image is generated by AI models. Unlike previous methods, this reverse engineering method does not hurt the image's quality, it uses the input to a model that would have generated a given image is reconstructed to infer the image's origin. This method effectively differentiates between images produced by different models and actual photographs, without altering the model or the images.

An alternate method is checking the noise of the images, the real images, for example, drawn by humans, photoed by humans, etc, have different noise features than AI-generated ones. In this paper[3], the authors developed some algorithms to distinguish the AI images and real-world images by checking the noise of the pictures.

7 CONCLUSION

This paper has examined two key technologies, Mist and Glaze, designed to protect copyrighted artwork from unauthorized AI training. Our findings indicate a complex interplay between the effectiveness of these methods and the ease with which they can be circumvented. While both techniques initially demonstrated promising results in obscuring artwork details sufficiently to confuse AI training algorithms, they also revealed significant vulnerabilities. For instance, Glaze's protections were quickly neutralized by straightforward coding interventions, highlighting the ongoing arms race between copyright protection technologies and methods to breach them.

Furthermore, Mist's limitations in terms of image resolution and detectable alterations underscore the practical challenges in deploying these protections without degrading the aesthetic or functional quality of the original artworks. These issues are indicative of a broader dilemma within AI development related to ethical uses of data and the protection of intellectual property without stifling innovation.

As AI capabilities continue to evolve at a rapid pace, the need for robust, adaptable, and legally compliant protection mechanisms becomes increasingly urgent. This research underscores the necessity for a collaborative approach among artists, technologists, and legislators to forge solutions that uphold creators' rights while fostering the beneficial uses of AI technologies. Moving forward, it

will be crucial to develop more resilient and less intrusive methods of protecting copyrighted content from unauthorized use, ensuring that innovation does not come at the expense of ethical considerations.

REFERENCES

[1] [n. d.]. Command given by Minyuan Zhu to ChatGPT, 21/04/24: Create a high-scool life image.

[2] altexsoft. 2023. *AI Image Generation Explained: Techniques, Applications, and Limitations.* Retrieved April 21, 2024 from https://www.altexsoft.com/blog/ai-image-generation/

[3] Xiuli Bi Bin Xiao Weisheng Li Xinbo Gao Bo Liu, Fan Yang. 2022. Detecting Generated Images by Real Images. *European Conference on Computer Vision* (2022).

[4] Pramod Chintalapoodi. 2023. Does using art to train AI violate copyright law? https://www.lexology.com/library/detail.aspx?g=6849db8b-9549-4f04-a7fb-5e1d3c992d93. Accessed: 2023-04-15.

[5] Beryl A. Howell. 2023. AI COPYRIGHT LAWSUIT thalerdecision. https://fingfx.thomsonreuters.com/gfx/legaldocs/lbvgooeoqvq/AI%20COPYRIGHT%20LAWSUIT%20thalerdecision.pdf. Accessed: 2024-04-15.

[6] Cyber Cadaver Institute. 2024. Cyber Cadaver Institute. https://weibo.com/u/7152334518. Accessed: 2024-04-16.

[7] Chumeng Liang and Xiaoyu Wu. 2023. Mist: Towards Improved Adversarial Examples for Diffusion Models. *arXiv preprint arXiv:2305.12683* (2023).

[8] Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. 2023. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. In *International Conference on Machine Learning*. PMLR, 20763–20786.

[9] Emily Matzelle. 2024. Top Artificial Intelligence Statistics and Facts for 2024. https://connect.comptia.org/blog/artificial-intelligence-statistics-facts. Accessed: 2024-04-16.

[10] openai. 2024. How ChatGPT and our language models are developed. https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed. Accessed: 2024-04-16.

[11] Emily Wenger Haitao Zheng Rana Hanocka Ben Y. Zhao Shawn Shan, Jenna Cryan. 2023. Glaze: Protecting Artists from Style Mimicry by Text-to-Image Models. *USENIX Security Symposium* (2023).

[12] Josephine Passananti Stanley Wu Haitao Zheng Ben Y. Zhao Shawn Shan, Wenxin Ding. 2024. Nightshade: Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models. *IEEE Symposium on Security and Privacy* (2024).

[13] Sumit Singh. 2023. *Everything you need to know about AI Model Training.* Retrieved April 21, 2024 from https://www.labellerr.com/blog/everything-you-need-to-know-about-ai-model-training/

[14] CBS Miami Team. 2024. Florida man accused of using AI to create child porn. https://www.cbsnews.com/miami/news/florida-man-accused-of-using-ai-to-create-child-porn/. Accessed: 2024-04-16.

[15] David Thiel. 2023. Investigation Finds AI Image Generation Models Trained on Child Abuse. https://cyber.fsi.stanford.edu/news/investigation-finds-ai-image-generation-models-trained-child-abuse. Accessed: 2024-04-16.

[16] users. 2023. Why do so many people bash AI painting? https://www.zhihu.com/question/569042768/answer/2932538066, note = Accessed: 2024-04-16.

[17] Alina Valyaeva. 2023. Everypixel Journal. https://journal.everypixel.com/ai-image-statistics. Accessed: 2024-04-16.

[18] Yi Zeng Lingjuan Lyu Shiqing Ma Zhenting Wang, Chen Chen. 2023. Where Did I Come From? Origin Attribution of AI-Generated Images. *Conference on Neural Information Processing Systems (NeurIPS 2023)* (2023).