WIP: Auditing Artist Style Pirate in Text-to-image Generation Models

Linkang Du[†]*, Zheng Zhu[†]*, Min Chen[‡], Shouling Ji[†], Peng Cheng[†], Jiming Chen[†], Zhikun Zhang[‡†§¶]
[†] Zhejiang University, Hangzhou 310017, China
Email: linkangd@gmail.com, zjuzhuzheng@zju.edu.cn, sji@zju.edu.cn, saodiseng@gmail.com, cjm@zju.edu.cn
[‡]CISPA Helmholtz Center for Information Security, Saarbrücken 66123, Germany
Email: min.chen@cispa.de
[§]Stanford University, Stanford, California 94305, USA
Email: zhikun@stanford.edu

***Abstract*—The text-to-image models based on diffusion processes, capable of transforming text descriptions into detailed images, have widespread applications in art, design, and beyond, such as DALL-E, Stable Diffusion, and Midjourney. However, they enable users without artistic training to create artwork comparable to professional quality, leading to concerns about copyright infringement. To tackle these issues, previous works have proposed strategies such as adversarial perturbation-based and watermarking-based methods. The former involves introducing subtle changes to disrupt the image generation process, while the latter involves embedding detectable marks in the artwork. The existing methods face limitations such as requiring modifications of the original image, being vulnerable to image pre-processing, and facing difficulties in applying them to the published artwork.**

To this end, we propose a new paradigm, called StyleAuditor, for artistic style auditing. StyleAuditor identifies if a suspect model has been fine-tuned using a specific artist's artwork by analyzing style-related features. Specifically, StyleAuditor employs a style extractor to obtain the multi-granularity style representations and treats artwork as samples of an artist's style. Then, StyleAuditor queries a trained discriminator to gain the auditing decisions. The results of the experiment on the artwork of thirty artists demonstrate the high accuracy of StyleAuditor, with an auditing accuracy of over 90% and a false positive rate of less than 1.3%.

I. Introduction

Text-to-image models based on diffusion processes represent a groundbreaking advancement in the field of generative artificial intelligence (AI), such as DALL-E [31], Stable Diffusion [32], and Midjourney [18], which can generate detailed and realistic images from textual descriptions. These models typically function by gradually refining a random pattern of pixels into a coherent image that matches the text description, making them suitable for a variety of creative and practical applications [24], [30], [36], [22], [29], [42], [5], [25]. Due to the stunning effect, these models are swiftly springing up among users and platforms. For instance, Midjourney receives around 32 million pageviews per day at around 7.5 pageviews per visit [16]. The downloads of the latest Stable Diffusion (v2.1) exceed 0.4 million per month.

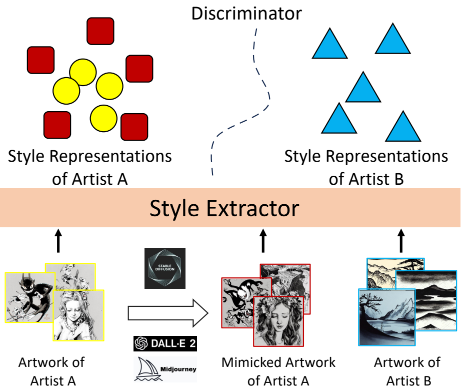
Discriminator
Style Representations of Artist A
Style Representations of Artist B
Style Extractor
DALL·E 2
Midjourney
Artwork of Artist A
Mimicked Artwork of Artist A
Artwork of Artist B

Fig. 1. Intuitive explanation of StyleAuditor. The figures with yellow borders represent the original artwork of artist A, the figures with red borders represent mimicked artwork after the fine-tuning of the diffusion model, and the blue borders represent the artworks of artist B. After style extraction, the images with red and yellow borders are clustered in the feature space far away from the representation of images with blue borders. The discriminator can judge the red picture as an imitation of the yellow picture based on the distributions of the features.

With the rapid development of text-to-image models, a user with no painting foundation can use simple prompts to generate artistic works at the level of a professional painter. As one of the sensational events, Jason M. Allen created his digital artistic work with Midjourney and took first place in the digital category at the Colorado State Fair [33]. Recently, many platforms have allowed users to upload artistic works and train the models that can generate works of similar style [8], [36], [28], which caused panic among the artist community regarding the infringements of their works [39]. To protect the intellectual property (IP) of the artists, a series of strategies have been proposed by the researchers [6], [41], [39], [7], [46], [10], [27], [26], [43], [11].

Existing Solutions. The main-stream solutions can be classified into two categories by the underlying technologies, *i.e.*, the adversarial perturbation-based methods [39], [41], [7], [46]

*The first two authors made equal contribution.
¶Zhikun Zhang is the corresponding author.

Workshop on AI Systems with Confidential Computing (AISCC) 2024
26 February 2024, San Diego, CA, USA
ISBN 979-8-9894372-4-5
https://dx.doi.org/10.14722/aiscc.2024.23009
www.ndss-symposium.org

and the watermarking-based methods [10], [27], [26], [47]. The adversarial perturbation-based methods introduce subtle perturbations that alter the latent representation in the diffusion process, causing models to be unable to generate images as expected. The watermarking-based methods inject well-designed watermarks into the artistic works before sharing them. The diffusion model collects and learns the original images with the injected watermarks in the training process. Then, the artists can validate the potential infringements by checking if the features of the watermarks exist in the generated images. However, the previous strategies face three main limitations for real-world application. First, both the adversarial perturbation-based and the watermarking-based strategies need to manipulate the original images, *i.e.*, perturbation or injecting watermark, inevitably affecting the normal image generation quality of the model. Second, the effectiveness of these solutions relies on the integrity of the injected features and the adversary can utilize the basic image processing methods (*e.g.*, spatial smoothing (SS) [44] and JPEG compression (JC) [12]), to reduce the protection effectiveness. Recently, several perturbation purification methods have been proposed to purify the image (*e.g.*, IMPRESS [4]), which weaken the protection of the existing work. Third, since many artistic works are published in the real world, adversarial perturbation-based and watermarking-based strategies require manipulation of the original images before sharing. Thus, they may not suit the artwork posted online.

Our proposal. In this paper, we propose the first practical artistic style auditing paradigm, called StyleAuditor, for the text-to-image models. We are inspired by the fact that artistic works within an artist's style share certain representation similarities. The adversary fine-tunes the text-to-image models to mimic the style of the artistic works and generate more images in this style. Thus, the auditor can mine the style-related features in an artist's works to form the auditing basis. Figure 1 provides a schematic diagram of StyleAuditor, where the core components are the style extractor and discriminator. Since the entire feature space retains a variety of information about the original image (*e.g.*, objects, locations, color, style), the auditor needs to utilize the style extractor to filtrate the irrelevant information. Then, the auditor adopts the discriminator to conduct the auditing. The discriminator will output a positive result if the feature representations of the generated images closely match those of the original images. Otherwise, the discriminator produces a negative prediction.

Evaluations. The experimental results show that the auditing accuracy of StyleAuditor exceeds 90% with false positive rates less than 1.3%. By comparing original artworks and mimicked artworks, we find that StyleAuditor can accurately identify imitations that are difficult to detect by the naked eye.

Contributions. Our contributions are two-fold:

- To our knowledge, StyleAuditor is the first dataset auditing method for the text-to-image models, using the extracted style representations as an intrinsic fingerprint of the artist.
- StyleAuditor is an efficient solution that allows the artist to easily perform the auditing on consumer-grade GPU.
- We demonstrate the effectiveness of StyleAuditor on Stable Diffusion with the artwork from thirty artists.

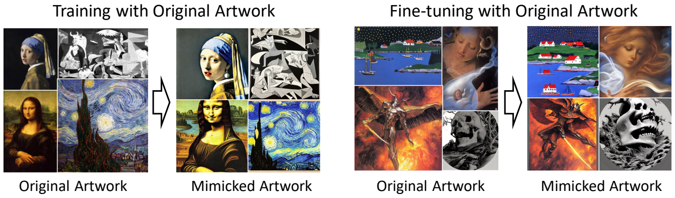

Fig. 2. An example of stylistic imitation based on Stable Diffusion. Left: original artwork. Right: generated artwork by Stable Diffusion trained or fine-tuned on the original artwork.

II. Background

A. Text-to-Image Generation

Generative adversarial network (GAN) [15], [9], [19] and diffusion model (DM) [31], [32], [18] have been used in text-to-image tasks. GAN in this space might struggle with the fidelity and diversity of the images. Diffusion models, inspired by the physical process of diffusion where particles spread over time, represent a significant development in generative models. These models function through a two-phase process: a forward process that gradually adds noise to an image over a series of steps until it becomes random noise and a reverse process where the model learns to reverse this, reconstructing the image from noise. The forward process gradually adds noise to an image x_0 over a series of steps T. This process can be represented as a Markov chain where each step adds Gaussian noise.

$$x_t = \sqrt{\alpha_t}x_{t-1} + \sqrt{1-\alpha_t}\epsilon_t, \tag{1}$$

where x_t is the noisy image at step t, x_{t-1} is the image from the previous step, ϵ_t is the noise added at step t sampled from a normal distribution, *i.e.*, $\epsilon_t \sim \mathcal{N}(0, I)$. α_t is a variance schedule determining how much noise to add at each step. It's a predefined sequence of numbers between 0 and 1.

The model learns to generate images by reversing the noise addition in the reverse process. At step t, the model predicts the noise ϵ_t added in the forward process and then uses this to compute the previous step's image x_{t-1}.

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right), \tag{2}$$

where $\epsilon_\theta(x_t, t)$ is the noise predicted by the model (parameterized by θ), given x_t and the time step t. $\bar{\alpha}_t$ is the cumulative product of α_i up to step t, *i.e.*, $\bar{\alpha}_t = \prod_{i=1}^{t}\alpha_i$. The model starts with a sample of pure noise $x_T \sim \mathcal{N}(0, I)$ and applies this denoising step iteratively to arrive at a generated data point x_0. The model's training involves learning the parameters θ to predict the noise ϵ_t at each step accurately. Diffusion models excel in generating highly detailed and coherent images, showing great flexibility and stability in training, making them less prone to issues like mode collapse.

B. Style Mimicry

Style Mimicry Technique. The concept of style mimicry in the text-to-image field refers to using diffusion models to create images that closely resemble a specific artistic style. The first way is to train the diffusion models from scratch on a large dataset of images that includes the target artists' artwork. This training allows the model to understand and replicate these styles. A naive mimicry attack directly queries

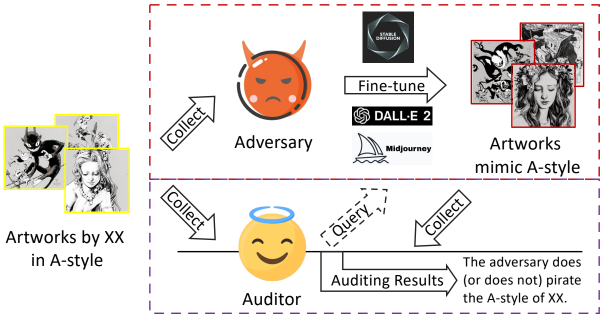

Fig. 3. An example of the application scenario. The artist acquires the auditing results by comparing the style representations between the original and generated artwork.

a generic text-to-image model using the name of the victim artist. For instance, in the left of Figure 2, we utilize Stable Diffusion to imitate artwork.

However, since the huge overhead for training the mainstream diffusion models, the adversary tends to fine-tune diffusion models for style mimicry, *i.e.*, adjusting the diffusion models by a small set of the target artist's artwork [13], [17], [21], [34]. This dataset encompasses unique elements like specific brushwork, color schemes, and compositional techniques characteristic of the style. The fine-tuning process involves continuous learning and adjustment to enhance the model's ability to apply these style characteristics accurately to various images. In the right of Figure 2, we demonstrate the model's imitation performance after fine-tuning.

Ethical Concerns. The ease of generating art using AI might devalue the skill, creativity, and expression involved in human-made art, diminishing the appreciation of human creativity. For instance, the artists feel that their unique styles are being appropriated when the market is flooded with AI-generated mimicked artwork. This raises questions about IP rights and copyright infringement. It is important to accurately determine whether infringement has occurred based on the images generated by the model, especially when reproducing styles that are distinct from specific creators.

III. Problem Statement

A. System and Threat Model

Application Scenarios. Comparing training the diffusion models from scratch, the adversary can easily implement the style mimicry on a low-end consumer GPU by fine-tuning the models. Thus, we mainly consider the fine-tuning scenarios in this work, where the adversary collects a small set of artwork from an artist and adjusts the models' parameters to mimic the artist's style. Figure 3 illustrates a typical application cases. Since many artists post their works online, adversaries can easily collect them from the internet by searching for the artist's name. Then, they fine-tune the diffusion model to generate artwork miming the artist's style. The artist stumbles upon the model's ability to generate artwork similar to his/her style and thus suspects the model's unauthorized use of his/her work for fine-tuning. The artist adopts StyleAuditor to audit the suspected model and obtains decisions.

Auditor's Background Knowledge and Capability. The artist has full access to his/her artwork and a low-end consumer GPU to extract the style representations. We consider the auditor to have black-box access to the suspect text-to-image model. Note that this is the most general and challenging scenario for the auditor. An adversary leverages several artists' works to fine-tune the text-to-image model and open model services to generate artwork with specific artistic styles. The artist collects the imitated images by querying the suspect model with pre-defined prompts.

B. Design Challenges

There arise two challenges in the design of StyleAuditor. The task of discerning and isolating style-specific characteristics from the artwork of a particular artist presents a significant challenge. The primary obstacle lies in the absence of a robust mathematical framework to precisely define and quantify "artistic styles." Generally, the style of an artwork is defined by a multifaceted combination of elements, each contributing to its unique aesthetic and thematic identity, *e.g.*, the nuances in brushwork, choice of color palette, subject matter, composition, perspective, the interplay of light and shadow, texture, as well as the historical and cultural context in which the artwork is created. For instance, Claude Monet is regarded as the quintessential impressionist. Monet's work is characterized by his fascination with light and its effects on the natural world. His series of paintings, like those of water lilies and the Rouen Cathedral, showcases his exploration of light and color at different times of day and in varying weather conditions. Monet's brushstrokes are fluid and seemingly spontaneous, capturing fleeting moments of natural beauty. Edgar Degas is also considered an impressionist, his style differs significantly from that of Monet. Degas's work is noted for its dynamic compositions and his skill in depicting movement and human anatomy, often using unusual perspectives.

The adversary might fine-tune the diffusion model on artwork from different artists to disturb the auditing. If an adversary wants to imitate a certain artist's artistic style, fine-tuning with exclusive images of this style tends to achieve a better artistic style imitation effect. However, this often results in the image generated by the fine-tuned model being too similar to the original image. The artist can easily sense the style plagiarism with the naked eye. Thus, the adversary can fine-tune the model by collecting artwork with different styles to mitigate the impact of a particular artist's work.

IV. Methodology

A. Intuition

Inspired by work in style migration [14], [45], we leverage latent space representations at different layers from the convolutional neural networks (CNNs) as the artist's style representation. In a CNN, the initial layers typically capture low-level features such as edges, colors, and textures. The latent space representations here are more closely related to the raw images. The deeper layers capture higher-level features. These may represent more abstract aspects of the images, like object parts or complex shapes. Then, we train a discriminator to compress the high-dimensional style representations into a confidence score for the final auditing.

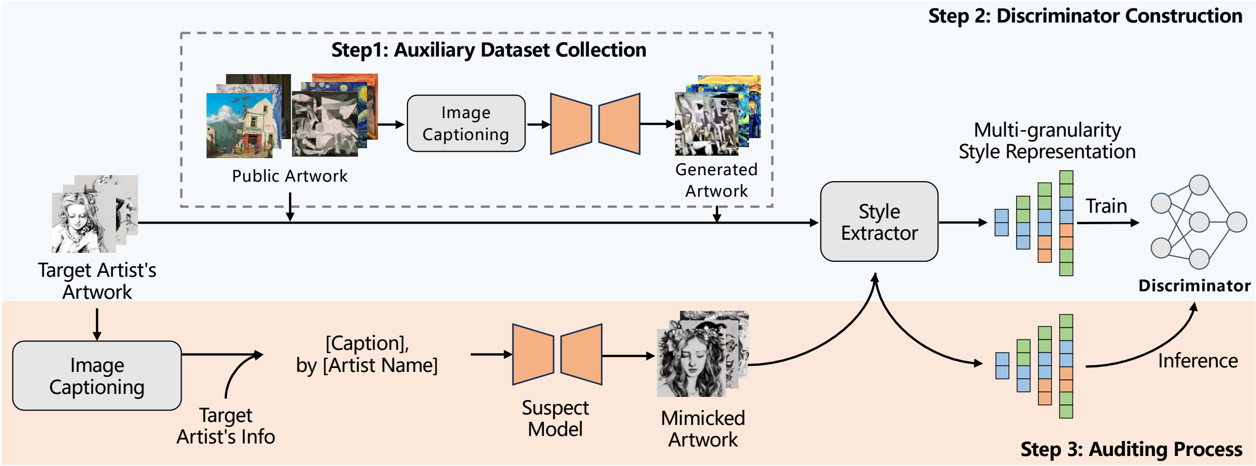

Fig. 4. The workflow of StyleAuditor contains three steps, *i.e.*, auxiliary dataset collection, discriminator construction, and auditing process. StyleAuditor first collects the public artwork and generated artwork by the target text-to-image model, then extracts the multi-granularity style representation to train the discriminator. Finally, StyleAuditor extracts the style features of mimicked artwork and inputs them into the discriminator to obtain the auditing results.

B. *Workflow of* StyleAuditor

We refer to the artist whose artwork is being audited as the *target artist*, and the *target text-to-image model* is the suspect model. If the suspect model is fine-tuned on the target artist's artwork, the discriminator should output a positive auditing result for the suspect model; otherwise, a negative auditing result. Figure 4 illustrates the workflow of StyleAuditor.

Step 1: Auxiliary Dataset Collection (ADC). The auxiliary dataset consists of artwork from two sources, *i.e.*, the public artwork and the generated artwork. The former represents the images published online, *e.g.*, WikiArt [40] and Artbench [23]. For public artwork, there are many world-famous images commonly included in the pre-training of the diffusion model [32], [37], such as the paintings of Picasso and Da Vinci. We collect the generated artwork by imitating the artwork of world-famous artists based on the suspect model. Since the diffusion model has distortion when imitating the artistic style, *i.e.*, there is a deviation between the original image and the generated image even under the same prompts. This bias will cause the discriminator to mistakenly judge negative samples as positive. Thus, we integrate the distortion in the discriminator training by measuring the difference between the generated artwork and the corresponding public artwork in Step 2.

Step 2: Discriminator Construction (DC). In the second step, the auditor trains a discriminator based on target artists' artwork, public artwork, and generated artwork. For ease of reading, we denote the above three kinds of artwork as X_t, X_p, and X_g respectively. Recalling the design challenges in Section III-B, we leverage a VGG model as the style extractor ϕ and select the output features of the four layers as the style representations. Then, for each image, we merge the style representations to form the training sample $\phi(x)$. We use 1 and -1 as the label value y, where $y = 1$ represents the artwork that originates from the target artist ($y = -1$ if it does not), to build the training set for the discriminator. During training, we optimize the parameters of the discriminator f_θ using the following loss function.

$$
\begin{aligned}
&\mathcal{L} = \mathcal{L}_{\text{regression}} + \mathcal{L}_{\text{distortion}}, \qquad (3)\\
&\mathcal{L}_{\text{regression}} = \left(y - f_\theta(\phi(x))\right)^2,\\
&\mathcal{L}_{\text{distortion}} = \left(f_\theta(\phi(x_g)) - f_\theta(\phi(x_p))\right)^2,
\end{aligned}
$$

where $\mathcal{L}_{\text{regression}}$ guides the discriminator to distinguish between the target artist's and other artists' artwork (*i.e.*, $x \in \{X_t, X_p\}$), and the distortion loss $\mathcal{L}_{\text{distortion}}$ to characterize the difference between the generated artwork and the corresponding public artwork (*i.e.*, $x_g \in X_G, x_p \in X_p$).

Step 3: Auditing Process (AP). After the above two steps, the auditor obtains the trained discriminator and then conducts the auditing process. To gather the suspect model's output for auditing, the auditor needs to create a set of prompts to encourage the suspect to produce the mimicked artwork. We adopt the clip-interrogator [2] as the image captioning to generate the prompts for each artwork of the target artist. To encourage the model to incorporate more features of the target artists in the generated artwork, we include the target artists' information in the generated caption. The auditor employs the style extractor to process the mimicked artwork and obtain their style representations. Then, the discriminator predicts the confidence scores based on the style representations. The auditor draws a conclusion by comparing the average confidence score with the preset threshold.

V. Evaluation

In this section, we evaluate StyleAuditor's overall auditing performance. We first describe the experimental setup in Section V-A. Then, we present the target model performance and auditing performance in Section V-B.

A. *Experimental Setup*

Target Models. We adopt Stable Diffusion (v2.1) in our experiments, which is a popular and high-performing text-to-image model trained on 11.5 million images from the LAION dataset [38]. It utilizes diffusion methods to generate images and achieves state-of-the-art performance on several benchmarks [32].

Metrics. We adopt four popularly used metrics to evaluate the performance of StyleAuditor, *i.e.*, accuracy, area under the curve (AUC), F1 score, and false positive rate (FPR), and please refer to Section A for more details.

TABLE I. A SUMMARY OF THE DATASET SETTINGS

Dataset		**Fine-tune Text-to-image Model**		**Train Discriminator**	
#.Users	#.Images per User	D^+	D^-	#.Training Images	#.Validation Images
(U)	*(I)*	*(50%*U*I)*	*(50%*U*I)*	*(80%*I*2)*	*(20%*I*2)*
30	20	300	300	32	8

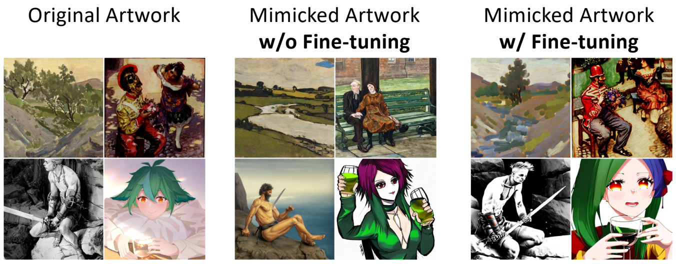

Fig. 5. Target model performance. The first column displays the original artwork created by the artists. The second column displays imitations generated by the text-to-image model before its fine-tuning on the original artwork. The final column showcases the imitations created after fine-tuning.

Datasets. From [37], [32], the training dataset of Stable Diffusion includes almost all popular-used public datasets, such as WikiArt [35], Artistic-Faces Dataset [1], and Painting-91 [20]. In order to evaluate the performance of StyleAuditor on data that did not participate in the model training and fine-tuning process (*i.e.*, negative samples), we built a new dataset containing the artwork of thirty artists based on fresh-published datasets [23] and publicly licensed artwork on the internet. In order to construct the dataset, we used a filtering approach that relied on clip-retrieval [3] to calculate the similarity score for each image with respect to a designated dataset. Subsequently, we picked out images whose similarity score is less than 0.8, *i.e.*, obviously different from the images training dataset of Stable Diffusion. To ensure an equal representation of all artists, we randomly selected twenty artwork from each artist.

Experimental Settings. We randomly split the thirty artists into two groups and utilized the artwork created by the first group to fine-tune the model. For ease of reading, we note the first group's artwork as D^+ and the second group's artwork as D^-. For the prompts, we use CLIP [2] to generate a description for each artwork and include the artist's name in the caption, following the previous work [39]. We fine-tune Stable Diffusion using dataset D^+. During the training of each artist's discriminator, we used the original artwork of each user as positive samples and further divided them into training samples and validation samples at a ratio of 8:2. For negative samples, we randomly sampled from the other twenty-nine artists' artwork while maintaining a positive-to-negative ratio of 1:1. We provide a summary of the dataset settings in Table I. In the auditing process, the threshold is set to zero.

B. Overall Auditing Performance

Target Model Performance. We first investigate the stylistic imitation ability of the target model, as shown in Figure 5. On the left-hand side, you will find the original artwork created by artists. The second part displays mimicked artwork generated without fine-tuning the target models with the original artwork. On the other hand, the third part showcases mimicked artwork produced when the target models are fine-tuned on the original artwork. By comparing these three parts in Figure 5, it becomes apparent that the target model, after being fine-tuned on the original artwork, exhibits a discernible ability to imitate artistic styles. However, detecting the imitation of certain artwork is not immediately evident, making it challenging to ascertain through direct visual inspection, as exemplified by the image in the lower left corner of Figure 5. This underscores the necessity of utilizing StyleAuditor to identify potential infringements.

TABLE II. OVERALL AUDITING PERFORMANCE FOR FOUR EVALUATION METRICS. WE REPORT THE MEAN AND STANDARD VARIANCE OF FIVE REPEATED EXPERIMENTS WITH DIFFERENT RANDOM SEEDS.

	Mean	Std
Accuracy	0.900	0.021
AUC	0.993	0.005
F1 Score	0.890	0.022
False Positive Rate	0.013	0.027

Auditing Performance. We evaluate the overall auditing performance of StyleAuditor. The experiments are repeatedly conducted on five random seeds $\{1, 2, 3, 4, 5\}$. The auditing performance is shown in Table II. We observe that StyleAuditor achieves good auditing performance with an auditing accuracy of up to 90% and an average false positive rate of 1.3%.

VI. CONCLUSION

In this work, we propose a novel artwork auditing method for the text-to-image models relying on the insight that the multi-granularity latent representations of the CNN model can serve as the artist's style fingerprint. Through multiple experiments, we show that StyleAuditor is an effective and efficient solution to protect the IP of the artist. The auditing accuracy of StyleAuditor can exceed 90% with less than 1.3% false positive rate, and the artist can easily perform auditing on consumer-grade GPU.

ACKNOWLEDGMENT

We would like to thank the anonymous reviewers for their constructive comments. This work was partly supported by the National Key Research and Development Program of China under No. 2022YFB3102100, NSFC under Grants 61833015, 62088101, and the Fundamental Research Funds for the Central Universities 226-2022-00107, 226-2023-00111. Min Chen was partly sponsored by the Helmholtz Association within the project "Trustworthy Federated Data Analytics" (TFDA) (No. ZT-I-OO1 4). Zhikun Zhang was supported by the CISPA-Stanford Center for Cybersecurity (FKZ:13N1S0762).

References

[1] Artistic-Faces Dataset. https://faculty.runi.ac.il/arik/site/foa/artistic-faces-dataset.asp.

[2] clip-interrogator. https://github.com/pharmapsychotic/clip-interrogator?tab=readme-ov-file.

[3] R. Beaumont. Clip Retrieval: Easily Compute Clip Embeddings and Build a Clip Retrieval System with Them. https://github.com/rom1504/clip-retrieval, 2022.

[4] B. Cao, C. Li, T. Wang, J. Jia, B. Li, and J. Chen. IMPRESS: Evaluating the Resilience of Imperceptible Perturbations Against Unauthorized Data Usage in Diffusion-Based Generative AI. *ArXiv*, abs/2310.19248, 2023.

[5] J. Chen, J. Wang, X. Ma, Y. Sun, J. Sun, P. Zhang, and P. Cheng. QuoTe: Quality-oriented Testing for Deep Learning Systems. *ACM Transactions on Software Engineering and Methodology*, 2023.

[6] M. Chen, Z. Zhang, T. Wang, M. Backes, and Y. Zhang. FACE-AUDITOR: Data Auditing in Facial Recognition Systems. In *USENIX Security*, 2023.

[7] R. Chen, H. Jin, J. Chen, and L. Sun. EditShield: Protecting Unauthorized Image Editing by Instruction-guided Diffusion Models. *ArXiv*, abs/2311.12066, 2023.

[8] CIVITAI. What the heck is Civitai? https://civitai.com/content/guides/what-is-civitai, 2022.

[9] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 2017.

[10] Y. Cui, J. Ren, H. Xu, P. He, H. Liu, L. Sun, and J. Tang. DiffusionShield: A Watermark for Copyright Protection against Generative Diffusion Models. *ArXiv*, abs/2306.04642, 2023.

[11] L. Du, M. Chen, M. Sun, S. Ji, P. Cheng, J. Chen, and Z. Zhang. ORL-Auditor: Dataset Auditing in Offline Deep Reinforcement Learning. In *NDSS*, 2024.

[12] G. K. Dziugaite, Z. Ghahramani, and D. M. Roy. A Study of the Effect of JPG Compression on Adversarial Images. *ArXiv*, abs/1608.00853, 2016.

[13] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-or. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *ICLR*, 2022.

[14] L. A. Gatys, A. S. Ecker, and M. Bethge. A Neural Algorithm of Artistic Style. *arXiv preprint arXiv:1508.06576*, 2015.

[15] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative Adversarial Networks. *Communications of the ACM*, 2014.

[16] C. Heidorn. Mind-Boggling Midjourney Statistics in 2023. Tokenized, 2023.

[17] E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR*, 2021.

[18] N. Ivanenko. Midjourney v4: An incredible new version of the ai image generator, 2022.

[19] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila. Alias-free generative adversarial networks. In *NeurIPS*, 2021.

[20] F. S. Khan, S. Beigpour, J. Van de Weijer, and M. Felsberg. Painting-91: A Large Scale Database for Computational Painting Categorization. *Machine Vision and Applications*, 2014.

[21] N. Kumari, B. Zhang, R. Zhang, E. Shechtman, and J.-Y. Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023.

[22] H. Li, Y. Yang, M. Chang, H. Feng, Z. hai Xu, Q. Li, and Y. ting Chen. SRDiff: Single Image Super-Resolution with Diffusion Probabilistic Models. *Neurocomputing*, 2021.

[23] P. Liao, X. Li, X. Liu, and K. Keutzer. The ArtBench Dataset: Benchmarking Generative Models with Artworks. *arXiv preprint arXiv:2206.11404*, 2022.

[24] G. Liu. The world's smartest artificial intelligence just made its first magazine cover. Cosmopolitan, 2022.

[25] P. Liu, J. Liu, et al. How ChatGPT is Solving Vulnerability Management Problem. *arXiv preprint arXiv:2311.06530*, 2023.

[26] G. Luo, J. Huang, M. Zhang, Z. Qian, S. Li, and X. Zhang. Steal My Artworks for Fine-tuning? A Watermarking Framework for Detecting Art Theft Mimicry in Text-to-Image Models. *ArXiv*, abs/2311.13619, 2023.

[27] Y. Ma, Z. Zhao, X. He, Z. Li, M. Backes, and Y. Zhang. Generative Watermarking Against Unauthorized Subject-Driven Image Synthesis. *ArXiv*, abs/2306.07754, 2023.

[28] V. Madan, H. Hotz, and X. Ma. Fine-tune Text-to-image Stable Diffusion Models with Amazon SageMaker JumpStart. https://aws.amazon.com/blogs/machine-learning/fine-tune-text-to-image-stable-diffusion-models-with-amazon-sagemaker-jumpstart/i, 2023.

[29] J. Meng, Z. Yang, Z. Zhang, Y. Geng, R. Deng, P. Cheng, J. Chen, and J. Zhou. SePanner: Analyzing Semantics of Controller Variables in Industrial Control Systems based on Network Traffic. In *ACSAC*, 2023.

[30] N. Popli. He Used AI to Publish a Children's Book in a Weekend. Artists Are Not Happy About It. https://time.com/6240569/ai-childrens-book-alice-and-sparkle-artists-unhappy/, 2022.

[31] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generation, 2021.

[32] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[33] K. Roose. An A.I.-Generated Picture Won an Art Prize. Artists Aren't Happy. https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html, 2022.

[34] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023.

[35] B. Saleh and A. Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature, 2015.

[36] Scenario.gg. AI-generated Aame Assets. https://www.scenario.gg/, 2022.

[37] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al. Laion-5b: An Open Large-scale Dataset for Training Next Generation Image-text Models. *NeurIPS*, 2022.

[38] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022.

[39] S. Shan, J. Cryan, E. Wenger, H. Zheng, R. Hanocka, and B. Y. Zhao. Glaze: Protecting Artists from Style Mimicry by Text-to-Image Models . *arXiv preprint arXiv:2302.04222*, 2023.

[40] W. R. Tan, C. S. Chan, H. Aguirre, and K. Tanaka. Improved ArtGAN for Conditional Synthesis of Natural Image and Artwork. *IEEE Transactions on Image Processing*, 2019.

[41] T. Van Le, H. Phung, T. H. Nguyen, Q. Dao, N. N. Tran, and A. Tran. Anti-DreamBooth: Protecting Users from Personalized Text-to-image Synthesis. In *ICCV*, 2023.

[42] K. Wang, J. Wang, C. M. Poskitt, X. Chen, J. Sun, and P. Cheng. K-ST: A Formal Executable Semantics of the Structured Text Language for PLCs. *IEEE Transactions on Software Engineering*, 2023.

[43] C. Wei, W. Meng, Z. Zhang, M. Chen, M. Zhao, W. Fang, L. Wang, Z. Zhang, and W. Chen. LMSanitator: Defending Task-agnostic Backdoors Against Prompt-tuning. In *NDSS*, 2024.

[44] W. Xu, D. Evans, and Y. Qi. Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. *ArXiv*, abs/1704.01155, 2017.

[45] Y. Zhang, F. Tang, W. Dong, H. Huang, C. Ma, T.-Y. Lee, and C. Xu. Domain Enhanced Arbitrary Image Style Transfer via Contrastive Learning. In *ACM SIGGRAPH*, 2022.

[46] Z. Zhao, J. Duan, K. Xu, C. Wang, R. Guo, and X. Hu. Can Protective Perturbation Safeguard Personal Data from Being Exploited by Stable Diffusion? *ArXiv*, abs/2312.00084, 2023.

[47] H. Zhu, M. Liu, C. Fang, R. Deng, and P. Cheng. Detection-Performance Tradeoff for Watermarking in Industrial Control Systems. *IEEE Transactions on Information Forensics and Security*, 2023.

Appendix

A. *The Details of Evaluation Metrics*

We use the following four metrics to evaluate the performance of StyleAuditor.

- **Accuracy.** We use accuracy to measure the auditing success rate. Concretely, accuracy measures the correct prediction of the total test.
- **Area Under the Curve (AUC).** For binary classification, AUC is the measure of the ability of a classifier to distinguish between classes when the decision threshold varies. The AUC is a measure of the model's ability to distinguish between positive and negative classes. An AUC score of 1 signifies that a model has achieved perfect prediction, while a score of 0.5 indicates random guessing.
- **F1 Score.** F1 Score is a harmonic mean of precision (the proportion of true positive cases to the member classes) and recall (the proportion of true positive cases to all correctly predicted classes), which can provide a better measure of the incorrectly classified cases than the accuracy metric. A higher F1 Score indicates better auditing performance.
- **False Positive Rate (FPR).** FPR evaluates the proportion of incorrect ownership claims to the total cases. In practice, a higher FPR degrades the credibility of StyleAuditor.