FPW: FREQUENCY-DOMAIN PIXEL-BY-PIXEL WATERMARKING AGAINST UNAUTHORIZED IMAGES USED ON TRAINING GENERATIVE MODEL

Jan Chi-Yuan, Hong-Han Shuai

Department of Electronics and Electrical Engineering, National Yang Ming Chiao Tung University

ABSTRACT

The proliferation of diffusion-based generative models has significantly advanced the field of synthetic image generation. However, the lack of transparency in training datasets, often not open-sourced, raises substantial concerns about copyright infringement, potentially involving the unauthorized use of proprietary images. To address this issue, we introduce Forensic Provenance Watermarking (FPW), a novel watermarking approach designed specifically for diffusion models. Specifically, FPW embeds a detectable pattern into the training data that, when used without permission, manifests in the generated images. This unique pattern serves as forensic evidence of data misuse, enabling data owners to legally challenge entities that misuse their copyrighted material. Our method simplifies watermark designs and utilizes spatial relationships within the images, employing a surrogate generative model to ensure robustness across various conditions. Experimental results on multiple datasets manifest that FPW maintains high detection rates even at low poisoning ratios, making it an effective tool for copyright protection in the era of advanced generative models. The code is available on the project page.

***Index Terms*—** Watermark, Diffusion-based Model, Generative Model, Deep-Learning, Machine-Learning

1. INTRODUCTION

In recent years, diffusion-based models like DALL·E [1] and Stable Diffusion [2] have revolutionized image generation, making these technologies widely accessible. While their applications in industries such as gaming and digital art are transformative, they also raise significant intellectual property (IP) concerns, particularly for online images. For instance, there is artistic style mimicry, where artworks shared online to promote creativity become vulnerable to unauthorized use. Recent there is a case that Los Angeles-based artist Hollie Mengert discovered her portfolio was used to train a model that convincingly replicated her unique style, blurring the line between her original works and AI-generated imitations.

This highlights the urgent need for mechanisms to prevent such exploitation. To address this, we propose a novel watermarking method that embeds detectable patterns into images before they are shared online. These patterns transfer to model outputs if the images are used in training, enabling the identification of unauthorized usage. By shifting from post-hoc verification to proactive safeguarding, our method offers a robust, preemptive defense against copyright infringement, empowering artists to protect their IP rights effectively.

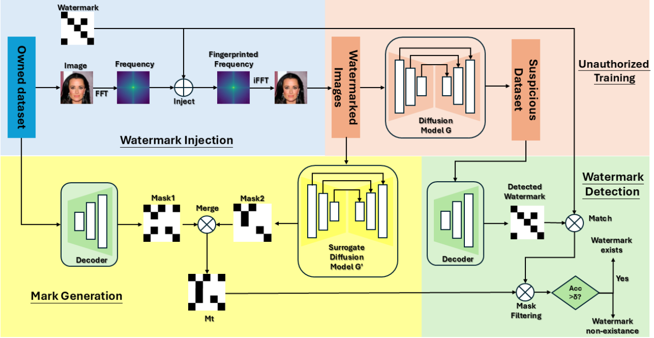


Fig. 1: We first inject the watermark to the image on FFT domain. After training the detector to distinguish the watermark, we generate a mask filtering out the pixels with poor performance. Afterward, the suspect images can be detected through the trained detector with masking.

Related research explores the potential of using encoded images to train or fine-tune generative models, enabling them to reproduce detectable watermarks as part of their output [3, 4]. The primary objective of these studies is to watermark the generative models themselves. That is, if a model generates images, the original model owner can verify their authorship by detecting the embedded watermark. However, this approach typically requires complete access to the entire training dataset. In contrast, our research addresses a more complex scenario where only a subset of the watermarked images is incorporated into the training process, presenting a significantly higher challenge. The limitations section of [5] further underscores this point, noting that a low poison ratio complicates the task of ensuring effective intellectual property protection. Our work aims to tackle this gap by developing a method capable of maintaining watermark detectability even when only a fraction of the training data is watermarked.

Therefore, to address this challenging problem, we propose a novel approach, namely, Forensic Provenance Watermarking (FPW). We embed the watermark without using an image encoder, as doing so increases the diversity of the watermark features. According to previous work [6], pattern uniformity and watermark detection rate shows a positive correlation. Therefore, we embed the watermark pixel by pixel and train a detector to identify whether a pixel is watermarked. Since the watermark can be easily removed by denoising method if the added watermark behaves similarly to random noise, we apply the watermarking method on FFT domain, which highly improves the detection rate of the watermarked images among generated images. Additionally, we employ a surrogate generative model to help filter out the pixels that the generative model struggles to detect. Moreover, it is much harder to detect the watermark in the high resolution case. To this end, we propose to change the pixel watermark to patch watermark, which makes the watermark flexible and remain detectable while the regenerated watermark is defected. The entire detection process is illustrated in Fig. 1. Our contributions are multifaceted and advance the field of digital watermarking in generative models:

- Introduction of Forensic Provenance Watermarking (FPW): We develop a novel watermarking technique, FPW, that operates in the frequency domain. This method is specifically tailored for scenarios where only a fraction of the training dataset is watermarked, thus addressing a more complex challenge of unauthorized data usage in generative model training.
- Enhanced Watermark Detectability: Unlike traditional methods that require complete access to the dataset, our approach maintains high watermark detectability even with low poisoning ratios. We achieve this by embedding watermarks pixel by pixel and employing a surrogate generative model to enhance the identification of watermarked pixels, particularly in high-resolution images where detection is typically more challenging.
- Innovative Benchmark for Watermark Reproducibility: We introduce a new benchmark, termed DR, which is designed to rigorously test the reproducibility and detectability of watermarks in images generated by diffusion-based models. This benchmark helps in evaluating the effectiveness of our watermarking approach under varied and challenging conditions.

2. RELATED WORK

HiDDeN [7] utilizes a DNN-based encoder and decoder to encode watermarks with improved efficacy. Subsequently, FNN [8] incorporates adversarial attacks within this encoder-decoder framework, improving accuracy but extending encoding time. LISO [9] speeds up this process by employing an encoder-decoder and critic network that mimics FNN's optimized results. Further, DiffusionSheild [6] promotes uniform watermarking to enhance detectability, employing patch watermarking optimized by trained detectors. Additionally, a recent method by [10] successfully employs poisoned image-text watermarking to ensure effectiveness in T2I image generation under low poison ratios. Besides image watermarking, other studies have explored watermark protection for various types of content. For instance, [11, 12, 13] focus on 3D object watermarking, while [14, 15] addresses watermarking for text documents. Different from previous work, which uses encoder-decoder networks or adversarial optimization, the proposed FPW method operates in the frequency domain and employs a surrogate model to filter out the uncertain ones to ensure the accuracy, enhancing the robustness and detectability even under low poisoning ratios and high-resolution conditions.

3. METHODOLOGY

3.1. Overview

Assume a data owner possesses a dataset D, which contains two subsets: D^{shadow} and D^{clean}. D^{shadow} is used for training the detector, while D^{clean} mimics the dataset that a malicious user might collect from external sources, ensuring that the detector's training is independent of D^{clean}. The first step involves using D^{shadow} to train the detector, with the specific training method discussed in a later subsection. Next, the owner selects a watermark W, shaped identically to the images, and embeds this watermark into a portion of the images. By embedding the watermarked images into the clean dataset D^{clean}, we create a new dataset D', simulating a scenario where a malicious user downloads images online to build their own training dataset. We then train a generative model $\mathcal{G}$ using the dataset D'. Since D' contains watermarked images, a portion of the synthetic images generated by the model also carry the watermark, allowing the detector to identify watermarked images. The detector outputs a detected watermark W_d, which is then compared with the original watermark W to determine whether the generative model was trained on watermarked images.

3.2. Detector Training

We first introduce the null hypothesis H_0. For images without embedded fingerprints, the detected output is expected to be unmatched with the selected watermark W. Inspired by HiDDeN [7], we assume under the null hypothesis H_0 that the detector produces a random binary sequence. By analyzing the null hypothesis and the number of matching bits M between W and the detected watermark W_d, the probability that the match is due to random guessing using the following formula is calculated by:

$$P = \Sigma_{i=M}^{N}(\frac{N!}{(i!)(N-i)!}) \cdot 2^{-N}, \tag{1}$$

where N denotes the fingerprint length, equivalent to the pixel number in the image. Given its large size, even with low detection accuracy, the likelihood of randomly matching a significant portion of the watermark is statistically negligible.

Next, we train the detector to produce a random sequence of binary bits when provided with clean images. The detector is specifically trained to extract watermarks from images with randomly implanted watermarks. Through optimization using binary cross-entropy loss, it learns to distinguish between watermarked and non-watermarked pixels. As a result, when processing an image without a watermark, the detector generates random predictions. The training process is guided by the binary cross-entropy loss function, defined as:

$$L(W, W_d) = W * log(W_d) + (1 - W) * log(1 - W_d). \quad (2)$$

Since spatial-domain watermarks are easily detected and removed, we shift to frequency-domain watermarking. During training, images are transformed via FFT, where a fixed perturbation is applied to the frequency components. The perturbed data is then converted back using inverse FFT (iFFT), mimicking standard image encoding. Finally, the watermarked image undergoes FFT again before being processed by the detector, which extracts the detected fingerprint W_d.

3.3. Masking Detection

The amplitude of frequency components is unevenly distributed, with higher concentrations in the center. Adding a uniform watermark value in the frequency domain can make the watermark difficult to be detected in the central region due to the high amplitude values. As a result, the detector cannot accurately identifying the watermark and tends to make random guesses. This results in approximately 50% accuracy in the low-frequency domain as amplitude values are generally higher.

Observing that pixels with large frequency amplitudes consistently yield similar detection outcomes, we identify this repetitive behavior as a detriment to accurate detection. Directly intensifying the watermark in the low-frequency domain could overly distort images, making them impractical for use. Instead of modifying the watermark strength, we opt to exclude defective pixels from the detection process, thereby preserving image quality and maintaining unaffected watermark detection. To implement this, we use a mask M_1 to block these pixels. M_1 is calculated by feeding clean images, randomly watermarked, into the detector. If the detection accuracy for any pixel strays notably from the expected 50%, that pixel is deemed unreliable and is consequently blocked by M_1. The formula for generating M_1 is as follows:

$$M_1(x, y) = \begin{cases} 1 & \text{if} \quad 0.5 - \delta_1 < acc(x, y) < 0.5 + \delta_1 \\ 0 & \text{otherwise} \end{cases}, \quad (3)$$

where δ_1 is a preset accuracy threshold. Moreover, $acc(x, y)$ denotes the accuracy at position (x, y), which is calculated by

$$acc(x, y) = \frac{1}{N} \Sigma_{i=1}^{N} (\mathbb{1}(\mathcal{D}(I_i)(x, y) \geq 0) \cdot W_i^r(x, y)), \quad (4)$$

where N is the total number of images in D^{shadow}, I_i is the image with random watermark W_i^r, δ_1 is a preset accuracy threshold, and $\mathcal{D}$ is trained decoder.

To effectively utilize the surrogate generative model, we design an additional mask, M_2, which filters out the pixels with low detection accuracy in the generated images from surrogate model during detection. First, we watermark a subset of images from D^{shadow} to construct a new training dataset, D^{sur}. By training the surrogate generative model $\mathcal{G}_{sur}$ on D^{sur}, we obtain a refined generator. By sampling N_{sur} images from $\mathcal{G}_{sur}$, we can generate M_2 by blocking the pixels with accuracy lower than δ_2. After combining M_1 and M_2, we obtain the final mask M:

$$M = M_1 \wedge M_2. \quad (5)$$

It is worth noting that setting a large δ_2 may lead to the final mask M causing the detector to output the selected watermark M, thereby compromising detection accuracy. To mitigate this issue, we start with a relatively high initial value for δ_2 and apply mask M to clean images. We then gradually decrease δ_2, carefully monitoring whether the detection results for clean images become excessively high. This optimization process ensures that M enhances detection performance while maintaining robustness.

3.4. Patch Watermark

Embedding a watermark into high-resolution images is challenging, particularly in maintaining its detectability in generated images. Several factors contribute to this difficulty. First, the wide dynamic range of frequency components in high-resolution images makes it harder for the watermark to remain both imperceptible and robust. Additionally, high-resolution image generation is inherently more complex, making it unsurprising that diffusion models struggle to accurately reproduce embedded watermarks. To address these challenges, we propose a patch-based watermarking approach, replacing the traditional pixel-by-pixel embedding method.

Specifically, by structuring the watermark at the patch level, we simplify its design while enhancing its detection robustness. A larger patch size not only reduces complexity but also increases the flexibility of watermark detection. Even if a portion of the watermark is distorted within a patch, the diffusion model's attempt to reproduce it introduces detectable distortions in the overall image distribution. This adaptation enhances the watermark's resilience against degradation and improves detection in high-resolution generated images.

4. EXPERIMENT

4.1. Experiment Setup

Dataset. We evaluate our method against several baseline approaches using three datasets: CIFAR-10 [16], CelebA [17], and FFHQ [18], with image resolutions of 32x32, 64x64, and

Table 1: Comparisons of different approaches on detection rate and image loss.

			HiDDeN (ECCV2018)	FNN (ICLR2022)	LISO (ICLR2023)	FPW	FPW$_M$
CelebA	Budget	LPIPS ↓	0.0211	0.0631	0.0712	**0.0155**	**0.0155**
		l_2 ↓	18.3551	8.2709	12.8207	**2.4135**	**2.4135**
	DR, p=0.3	δ_a^{low} ↑	0.1693	0.0823	0.0823	0.1497	**0.2693**
		δ_a^{high} ↑	0.1693	0.0820	0.0820	0.1477	**0.2680**
	DR, p=0.1	δ_a^{low} ↑	0.0267	0	0	**0.0680**	**0.0680**
		δ_a^{high} ↑	0.0260	0	0	**0.0680**	0.0670
CIFAR-10	Budget	LPIPS ↓	0.0261	0.0389	0.0953	**0.0230**	**0.0230**
		l_2 ↓	27.7659	10.1356	20.3372	**6.4018**	**6.4018**
	DR, p=0.3	δ_a^{low} ↑	0.1890	0.0440	0.1063	0.1527	**0.3163**
		δ_a^{high} ↑	0.1850	0	0.0233	0.1527	**0.3153**
	DR, p=0.1	δ_a^{low} ↑	0.0397	0	0.0357	0.0340	**0.0650**
		δ_a^{high} ↑	0.0383	0	0.0040	0.0340	**0.0650**

Table 2: Comparison on FFHQ with HiDDeN

		HiDDeN	FPW	FPW(BL-16)
FFHQ	LPIPS ↓	**0.0176**	0.0433	0.0447
	DR p=0.3 ↑	0.1457	0	**0.2777**
	DR p=0.1 ↑	0	0	**0.1517**

256x256 pixels, respectively. Each dataset is divided equally into two subsets: one subset serves as clean images to simulate malicious data collection, while the other is used for training the detector $\mathcal{D}$ and the surrogate generative model $\mathcal{G}_{sur}$. We employ poison ratios of 10% and 30% for training the generative model $\mathcal{G}$. For $\mathcal{G}_{sur}$, we use the full poison ratio rate, with a detailed analysis of varying poison ratios for $\mathcal{G}_{sur}$ to be discussed in the ablation study.

Baseline. We choose up-to-date steganography methods as baselines which include the encoder-decoder based method HiDDeN [7], the optimization-based method FNN [8], and the mimicking optimization-based method LISO [9]. The baseline implement detail can be found in the supplementary material.

Implementation Details. For the generative diffusion models, we adopt DDPM-IP [19] as our test model, which has demonstrated state-of-the-art (SOTA) performance on the 64×64 CelebA generation task. All generation tasks are trained for 50,000 steps while keeping other hyperparameters at their default values. For watermark detection, we employ a convolutional-based deep neural network (DNN) model, which is resolution-agnostic and supports efficient training. Notably, since the encoding method is independent of the detector, images can be released first, with the detector trained later if necessary. Furthermore, our approach allows multiple users to share a single detector, as the intellectual property protection key is derived from a common seed.

Evaluation Metrics. We consider the detection rate (DR) among generated images as the metric, as it reflects the reproducibility of the watermark. To precisely define the detection rate in our scenario, we introduce the following procedure. A dataset is created by embedding a specific watermark W into a subset of images while keeping the remaining images clean. A generative model $\mathcal{G}$ is then trained on this dataset. After generating K_d images from $\mathcal{G}$, each generated image is analyzed by extracting its detected watermark W_d and comparing it with W. If the extracted watermark meets or exceeds a predefined accuracy threshold δ_a, the image is considered successfully detected. The detection rate is then computed as the proportion of successfully detected images among the K_d generated samples.

The value of the threshold δ_a directly affects the probability P of successful detection. Additionally, the number of bits in the watermark fingerprint influences P. To ensure a fair comparison across different methods, we adjust δ_a so that all methods are evaluated under the same probability P. To measure the impact of watermark embedding on image quality, we use *LPIPS loss* [20] and the l_2 norm to assess the difference between watermarked and original images. For the setting of δ_a, we use accuracy thresholds of 70% and 80% with 100-bit watermarking as baselines, adjusting all methods to match the same P. The sample number K_d is set to 3,000.

4.2. Experimental Result

As illustrated in Table 1, unlike baseline methods, our watermarking technique supports adjustable perturbation values. Since the added perturbation correlates with detection rate,

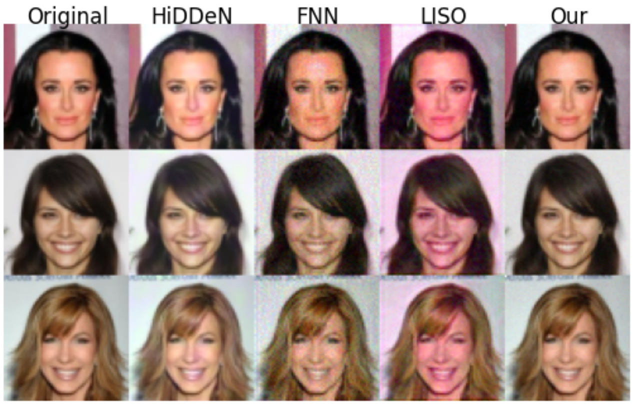


Fig. 2: Qualitative result comparison between ours and compared previous method.

we select a value that minimizes LPIPS loss compared to other methods, ensuring a fair comparison. Notably, while FNN and LISO achieve high detection accuracy, they incur significant image loss, making the perturbation harder to conceal as shown in Fig. 2.

Before applying masking, our method achieves a strong detection rate but slightly underperforms compared to HiDDeN. However, with masking—where low-detection regions (primarily low-frequency components) are excluded—the detection rate improves significantly, closely matching the original poison ratio. As shown in Fig. 2, perceptual comparisons highlight that our watermark introduces minimal disruption to the image, demonstrating its stealthiness.

Among the evaluated methods, HiDDeN achieves the second-best detection rate. This is likely because methods like FNN involve more complex watermarking operations, making it harder for the diffusion model to effectively learn the watermark. For LISO, the model is trained for 100 epochs with only 1000 images in the original setting, leading to overfitting and significant image distortion in low-resolution cases. We reduced the training to 20 epochs, which mitigated distortion but still caused noticeable artifacts. While LISO maintains high accuracy on watermarked images, it struggles to regenerate the watermark in synthetic images.

For the FFHQ dataset, we select HiDDeN as the baseline due to its strong performance on CelebA and CIFAR-10, as shown in Table 2. The table demonstrates that FPW initially struggles to be detected on generated images, but patch-based embedding significantly improves detection performance, outperforming HiDDeN. We opt for a slightly higher image distortion to ensure effective watermarking in 10% of cases. With lower distortion, FPW could further surpass HiDDeN in detection accuracy.

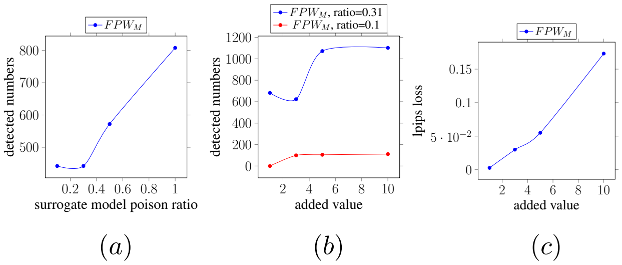
4.3. Ablation Study

We conduct experiments on watermarking in the spatial domain. The results show that the watermark often fails to reproduce in the diffusion model. Specifically, we evaluated fine-tuning and training generative models using spatial-domain and frequency-domain watermarked image datasets. As shown in Table 3, spatial-domain watermarking resulted in lower detection rates despite higher image distortion. Additionally, we test the masking method across various poison ratios, as presented in Fig. 3a. The results indicate that a surrogate generative model trained on a fully watermarked dataset provides a more effective filtering mask for detection. Furthermore, we examine the impact of perturbation values on watermark effectiveness. Fig. 3b and 3c shows that the detection rate improves as image distortion increases.

Table 3: Comparison watermarking on spatial-domain and frequency-domain

		Spatial-Domain	Frequency-Domain
CelebA	LPIPS ↓	0.0210	**0.0155**
	DR p=0.1 ↑	0.0210	**0.0680**

We test the two different-domain watermark on CelebA with threshold δ_a^{high}, and show that the frequency-domain show higher reproducibility.



(a) (b) (c)

Fig. 3: (a) Performance on surrogate models trained with different poison ratio. (b) Effect of added perturbation values on DR. (c) The relation between added perturbation values and LPIPS loss.

5. CONCLUSION

We propose a novel watermarking method, FPW, designed to protect the intellectual property of images from unauthorized use in training generative models. Our approach embeds a long-length watermark into the image's frequency components, making it not only easier for generative models to reproduce but also more discreet and harder to detect. By leveraging a surrogate model, we introduce a masking technique to filter out pixels that are challenging for generative models to replicate, achieving a higher detection rate compared to previous methods while maintaining low image distortion. Furthermore, transitioning from pixel-wise to patch-wise watermarking simplifies the watermark's complexity and demonstrates high reproducibility in diffusion-based generative models, as evidenced by our experimental results.

6. REFERENCES

[1] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever, “Zero-shot text-to-image generation,” in *Proceedings of the 38th International Conference on Machine Learning*. 18–24 Jul 2021, vol. 139 of *Proceedings of Machine Learning Research*, pp. 8821–8831, PMLR.

[2] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, “High-resolution image synthesis with latent diffusion models,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10674–10685.

[3] Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz, “Artificial fingerprinting for generative models: Rooting deepfake attribution in training data,” in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 14428–14437.

[4] Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon, “The stable signature: Rooting watermarks in latent diffusion models,” in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 22409–22420.

[5] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao, “Glaze: Protecting artists from style mimicry by {Text-to-Image} models,” in *32nd USENIX Security Symposium (USENIX Security 23)*, 2023, pp. 2187–2204.

[6] Yingqian Cui, Jie Ren, Han Xu, Pengfei He, Hui Liu, Lichao Sun, Yue Xing, and Jiliang Tang, “Diffusionshield: A watermark for data copyright protection against generative diffusion models,” *ACM SIGKDD Explorations Newsletter*, vol. 26, no. 2, pp. 60–75, 2025.

[7] Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei, “Hidden: Hiding data with deep networks,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 657–672.

[8] Varsha Kishore, Xiangyu Chen, Yan Wang, Boyi Li, and Kilian Q. Weinberger, “Fixed neural network steganography: Train the images, not the network,” in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. 2022, OpenReview.net.

[9] Xiangyu Chen, Varsha Kishore, and Kilian Q. Weinberger, “Learning iterative neural optimizers for image steganography,” in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. 2023, OpenReview.net.

[10] Haonan Wang, Qianli Shen, Yao Tong, Yang Zhang, and Kenji Kawaguchi, “The stronger the diffusion model, the easier the backdoor: Data poisoning to induce copyright BreachesWithout adjusting finetuning pipeline,” in *Proceedings of the 41st International Conference on Machine Learning*. 21–27 Jul 2024, vol. 235 of *Proceedings of Machine Learning Research*, pp. 51465–51483, PMLR.

[11] Fei Peng, Bo Long, and Min Long, “A semi-fragile reversible watermarking for authenticating 3d models based on virtual polygon projection and double modulation strategy,” *IEEE Transactions on Multimedia*, vol. 25, pp. 892–906, 2023.

[12] Bianca Jansen van Rensburg, Adrian G. Bors, William Puech, and Jean-Pierre Pedeboy, “Simultaneous watermarking and draco 3d object compression method,” in *2023 IEEE International Conference on Image Processing (ICIP)*, 2023, pp. 725–729.

[13] Innfarn Yoo, Huiwen Chang, Xiyang Luo, Ondrej Stava, Ce Liu, Peyman Milanfar, and Feng Yang, “Deep 3d-to-2d watermarking: Embedding messages in 3d meshes and extracting them from 2d renderings,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10021–10030.

[14] Yinan Li, Weiming Zhang, Han Fang, Xi Yang, Zehua Ma, and Nenghai Yu, “Font watermarking network for text images,” in *2022 IEEE International Conference on Image Processing (ICIP)*, 2022, pp. 486–490.

[15] Zhenhao Shi, Hongxia Wang, Heng Wang, and Xinyi Huang, “Robust screen-shooting document watermarking for multiple fonts,” *IEEE Signal Processing Letters*, vol. 31, pp. 2215–2219, 2024.

[16] Alex Krizhevsky, Geoffrey Hinton, et al., “Learning multiple layers of features from tiny images,” 2009.

[17] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, “Deep learning face attributes in the wild,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 3730–3738.

[18] Tero Karras, Samuli Laine, and Timo Aila, “A style-based generator architecture for generative adversarial networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 12, pp. 4217–4228, 2021.

[19] Mang Ning, Enver Sangineto, Angelo Porrello, Simone Calderara, and Rita Cucchiara, “Input perturbation reduces exposure bias in diffusion models,” in *Proceedings of the 40th International Conference on Machine Learning*. 23–29 Jul 2023, vol. 202 of *Proceedings of Machine Learning Research*, pp. 26245–26265, PMLR.

[20] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.