Empirical Study on the Watermarking Methods for Diffusion Models

Alireza Ganjdanesh Jerome Mathew Zuzanna Rutkoswka Reza Shirkavand

Abstract

With the rise of generative AI models, especially text-to-image models, in recent years, establishing methods for detecting whether an image is AI-generated, protecting copyrighted materials, and tracing the origin of AI-generated content have gotten significant attention. Most of the proposed ideas focus on training detectors implemented as deep classifiers to detect whether an image is AI-generated. Yet, the trained models struggle to generalize to new generative models, which makes them impractical. Accordingly, watermarking has emerged as the prevailing approach for detecting and tracing AI-generated images recently. However, the proposed watermarking methods are prone to adversarial attacks that can 'wash away' the watermark. In this paper, we propose two ideas to recognize whether an image is AI-generated while aiming to address the limitations of the previous approaches: a) not training a model at all, and b) designing them to be less prone to adversarial attacks compared to watermark detection methods. In the first approach, we take an off-the-shelf pretrained diffusion model and employ it to reconstruct two noisy versions of images that are perturbed with different strengths. Our intuition is that the reconstruction error for the real and fake images should have different distributions. In our second idea, we propose a new watermarking technique in which we guide the sampling process of the diffusion model using a randomly initialized classifier. The hope is that the resulting watermarking method be resilient to adversarial attacks as the classifier is a randomly initialized one. While each variant excels in its setting, preliminary results show that both can be utilized for the recognition AI-images. Our code is available here.

1. Introduction

The recent surge in generative AI models has provided unprecedented opportunities in different applications. For example, powerful text-to-image generative models like Stable Diffusion [13], DALL.E3 [3], Midjourney, Imagen [17], and Adobe Firefly can enable users to edit images or create new ones with only using text prompts.

However, the widespread deployment of these models raises several challenges. First, as these models are usually trained on very large-scale datasets gathered from the internal data of the companies and/or from the internet, their training data may contain copyrighted materials. This can lead to cases in which the models re-generate the copyrighted material in their inference [19, 20]. Second, as these models may have been trained on sensitive contents, malicious users may employ them to generate harmful contents. Therefore, there is a need for a method to determine whether an image is AI-generated and to trace which model has generated an image and who has done so. We mainly focus on the former research question to determine whether an image is AI-generated.

The proposed methods in the literature to detect AI-generated images can be categorized into two groups: 1) classifier-based approaches, and 2) watermarking methods. Classifier-based methods usually train a deep classifier using a dataset containing real images and images generated by one or various generative AI model[s]. However, they mainly struggle to generalize to images that are generated by generative models other than their training data. This problem makes them impractical because of the rapid development in the generative AI area where a new model is introduced every few weeks and months, which requires training the classifier again for the new model. In another direction, watermarking methods have recently gained attention to detect AI-generated images. They embed an invisible signal into the generated images after [6, 15, 18] or during [24] the generation process. However, it has been shown that [16] these methods are vulnerable to adversarial attacks such that a malicious user can use adversarial attacks to 'wash away' the watermark included in the images, making them undetectable.

In this paper, we aim to propose new techniques to detect AI-generated images while trying to alleviate the shortcomings of the previous methods. In our first idea, We consider two main characteristics of our techniques: a) not training a new model to detect AI-generated images, and b) making our method more robust to adversarial attacks by design (at least intuitively). In more details, in our first idea, we perturb an image (real or generated) with two noise levels and employ an off-the-shelf diffusion model to reconstruct the noisy versions. Inspired by the binoculars paper [1], our intuition is that the difference between reconstruction errors for these two cases should be different for real *vs.* generated images. In our second idea, we propose a new watermarking technique in which we guide the generation process of a diffusion model using a classifier that is randomly initialized. The intuition is that as the guidance is coming from a random classifier that is unknown to the users, it will be hard for malicious users to adversarially attack it.

In summary, we propose two new techniques to detect whether an image is AI-generated and perform experiments to gauge how good they can perform to do so.

2. Related Work

Digital Watermarks Techniques to imprint watermarks on digital content have been widely used since the late 1990s [6, 15, 18]. The main technique that was used was to imprint the watermark in the frequency decomposition of the image. Some examples include DCT (Discrete Cosine Transform), DWT (Discrete Wavelet Transform), and the Fourier-Mellin transform . The main benefit to this class of transform techniques was that these transformations was invariant to simple manipulations like translations and rotations [24].

Adversarial Attacks on AI detector A recent paper [16] was published that demonstrated two techniques for removing watermarks based on the size of the perturbation it caused to the image. In dealing with low-budget watermarks (watermarks that introduce subtle image perturbations), a diffusion purification attack was used that noised, and then subsequentely denoised the image. As purification attacks were only effective for low-budget watermarks, a different technique was engineered to address high-budget watermarks. A model substitution adversarial attack was used, where the substitute classifier was adversarially attack in order to deceive the watermark detector. Since the two techniques lead to a positive relationship between False Positive Rate (FPR) and True Positive Rate (TPR), choosing too low of a FPR can lead to these methods becoming impractical. However, with low time step value, these methods are quite efficient in attacking watermarks.

Integrated Watermarking in Generative Models
With the advent of deep learning, the watermarking community has had a shift in perspective. As opposed to post-hoc modification techniques (where a watermark was imprinted into he image after it was available), generative AI models have popularized the idea of ”deep” watermarking. In this technique, the watermark encoder and decoder are embedded into the generative model as learned models themselves. Examples of such watermarks include theThe Stable Signature [9] and the Tree-Ring watermark [24].

3. Approach

3.1. Detecting Diffusion Images

Diffusion models for image generation [10] are latent variable models of the form $p_\theta\left(\mathbf{x}_0\right) := \int p_\theta\left(\mathbf{x}_{0:T}\right) d\mathbf{x}_{1:T}$, where $\mathbf{x}_1, \ldots, \mathbf{x}_T$ are latents of the same dimensionality as the data $\mathbf{x}_0 \sim q\left(\mathbf{x}_0\right)$. The joint distribution $p_\theta\left(\mathbf{x}_{0:T}\right)$ is called the reverse process, and it is defined as a Markov chain with learned Gaussian transitions starting at $p\left(\mathbf{x}_T\right) = \mathcal{N}\left(\mathbf{x}_T; \mathbf{0}, \mathbf{I}\right)$:

$$p_\theta\left(\mathbf{x}_{0:T}\right) := p\left(\mathbf{x}_T\right) \prod_{t=1}^{T} p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right),$$

where $p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right) := \mathcal{N}\left(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta\left(\mathbf{x}_t, t\right), \boldsymbol{\Sigma}_\theta\left(\mathbf{x}_t, t\right)\right)$
The mean and the variance of the Gaussian transitions are learned by denoising the following forward process

$$q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t I\right),$$

with the loss function being the variational negative log likelihood:

$$L := \mathbb{E}_q\left[-\log p\left(\mathbf{x}_T\right) - \sum_{t>1} \log \frac{p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)}{q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right)}\right]$$

Importantly, the forward process allows for sampling x_t, given x_0, at arbitrary timestep t

$$q(x_t \mid x_0) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)I\right),$$

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. Given a pretrained model θ, the original image can be then approximately restored using t applications of the reverse process: $p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)$. Let $x_0(\theta, t)$ denote the result of the reverse process started at the timestep t using model θ. Our model assesses the likelihood of the image x_0 being generated by a diffusion model θ' by calculating the proportion

$$\frac{||x_0(\theta, t_1) - x_0||_F^2}{||x_0(\theta, t_2) - x_0||_F^2},$$

where $t_1 < t_2$ are different timestesps in the domain of all timesteps of the model θ. The obtained proportion is is a number in $\mathbf{R}_+$ where larger values give more likelihood of an image being from a real source while lower values give more likelihood of an image being generated by the other diffusion model θ'.

Other metrics. It is possible to avoid the inherently sequential and slow reverse process by parametrizing the diffusion model [21]. In this approach, the diffusion process is designed to preserve the marginal distribution

$$q(x_t \mid x_0) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)I\right),$$

at every timestep t, however, the diffusion is no longer Markovian. A model is then trained to predict x_0' given x_t at a single shot

$$f_\theta^{(t)}\left(\boldsymbol{x}_t\right) := \left(\boldsymbol{x}_t - \sqrt{1-\alpha_t} \cdot \epsilon_\theta^{(t)}\left(\boldsymbol{x}_t\right)\right) / \sqrt{\alpha_t}.$$

Given the marginal distribution of the diffusion process let us define the model us

$$p_\theta^{(t)}\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t\right) = \begin{cases} \mathcal{N}\left(f_\theta^{(1)}\left(\boldsymbol{x}_1\right), \sigma_1^2 \boldsymbol{I}\right) & \text{if } t=1 \\ q\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, f_\theta^{(t)}\left(\boldsymbol{x}_t\right)\right) & \text{otherwise,} \end{cases}$$

for some constant σ_1^2. Therefore, at timestep t the model θ is trained to optimize the variational lower bound between distribution $q(x_{t-1} \mid x_t)$ and $p_\theta^{(t)}\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t\right)$. Regardless, the model allows to approximate x_0 at a single shot from any timestep t. Under the hypothesis that for a small timestep t the quality of this approximation is similar to the full-step reverse process, we propose another metric for detecting diffusion images

$$KL\left(\nabla f_\theta^{(t)}\left(\boldsymbol{x}_t\right); \nabla x_0\right).$$

Here, the operator ∇ denotes the set of coordinate-wise differences in the pixel space of an image. Then, the Kullback-Leibler divergence is applied to the obtained set of differences treated as distributions on the real axis. We expect real images to have sharper edges and thus be harder to predict for the reverse process. Thus we associate a higher value of this metric with real images while a lower should correspond to images originating from a diffusion model.

3.2. Guided Watermarking

Watermarking within a diffusion model involves embedding information while ensuring its imperceptibility to maintain data integrity. Our approach employs guided diffusion [2, 8], utilizing a frozen watermarking model $f_\psi(.)$ randomly initialized. During the denoising process, the noisy latent in each timestep z_t is used to estimate the initial denoised latents z_0 according to DDIM:

$$z_0^t = \frac{z_t - \sqrt{1-\alpha_t}\epsilon}{\sqrt{\alpha_t}} \tag{1}$$

Then we get the final image using the VQVAE decoder:

$$x_0^t = Decoder(z_0^t) \tag{2}$$

The image is input to the regressor, and the resuting gradients obtained from the regressor are utilized as guiding signals to modulate the diffusion steps:

$$\tilde{\epsilon}_\theta(z_t) = \epsilon_\theta(z_t) - \omega \nabla_{z_t} log(f_\psi(x_0)) \tag{3}$$

$$z_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \frac{z_t - \sqrt{1-\alpha_t}\tilde{\epsilon}}{\sqrt{\alpha_t}} + \sqrt{1-\bar{\alpha}_{t-1}}\tilde{\epsilon} \tag{4}$$

Our approach aims to confuse the watermarking technique by adopting a gradient-based adversarial strategy, inducing significant discrepancies between the output distributions of the watermarking model for generated images and authentic ones. To distinguish watermarked images from non-watermarked ones, a straightforward method involves setting a threshold τ and categorizing images where the model output surpasses τ as watermarked, while labeling other images as non-watermarked.

4. Results

4.1. Detecting Diffusion Images

Datasets. We used three datasets to evaluate our model. Each dataset consists of two sets: a set images generated by a diffusion model and a set of real images from some domain. The detailed description of each dataset is given below.

- Dataset A. It comprises: 1000 real images of quality 256×256 pixels presenting churches [11]; 1000 images generated by Dalle 3 model of quality 512×512 pixels [7].
- Dataset B. It comprises: 400 real images of quality 768×768 pixels presenting churches [5]. These images were chosen from the Wikipedia webpages [4]; 400 images generated by Stable Diffusion 2.1v of quality 768×768 pixels. [22],
- Dataset C. It comprises: 400 real personal images of quality at least 768×768 pixels presenting variety of things (landscapes, animals, food, etc.) [12]; 400 images generated by Stable Diffusion 2.1v of quality 768×768 pixels. Prompts were generated randomly by ChatGPT. [23],

Experiments. We tested both approaches on each of the datasets A, B, and C. We performed a grid search for deciding the best pair of timesteps, trying every pair of timestep $t_1, t_2 \in \{5, 20, 25, 75, 100, 200\}$. For an image x_0 the proportions

$$\frac{||x_0(\theta, t_1) - x_0||_F^2}{||x_0(\theta, t_2) - x_0||_F^2}$$

and

$$KL\left(\nabla f_\theta^{(t)}\left(\boldsymbol{x}_t\right); \nabla x_0\right)$$

were calculated using Stable-Diffusion v2.1-base as the pretrained model θ. This model has been natively trained on 512 × 512 images and for meaningful results, all images were transformed to this resolution before processing by our model. To avoid upscaling the input image, which might affect the signal of the image, we preferred center-cropping if possible. Following this approach, images from datasets B, C were only cropped before processing by the model. The real images from dataset A were additionally pre-processed by bilinear interpolation to match the model 512×512 resolution.

The calculated proportions were treated as the likelihood of this image being a real image. We assessed the quality of our model by presenting the receiver operating characteristic for each dataset and each model, see Figure 1 and Figure 3. Note, that the images of the dataset A have been affected by the upscaling operation which results in a much large similarity between generated and real images (figures (a) in the top and the bottom row).

4.2. Guided Watermarking

We use Stable Diffusion v2.1 [14]. We employ a ResNet-50 architecture as the core watermarking method. By replacing its classification head with a 10-way linear layer, we obtain $f_\psi(.) : \chi \rightarrow \mathbb{R}^n$. To reduce variance, we aggregate the outputs' values by averaging them. In our experimentation, we observed that a randomly initialized model shows slightly superior performance compared to a pre-trained model. As a result, we consistently utilize the randomly initialized model across all our experiments. We explore various watermarking guidance scales denoted as (ω in Equation (3)). Examples of generated images for each guidance scale and the resulting watermarking model distribution are illustrated in Figure 4

5. Conclusion

In summary, our exploration into detecting AI-generated images introduced two promising methods. The first leveraged pre-trained diffusion models, demonstrating potential in distinguishing between real and AI-generated images across various datasets. Meanwhile, our guided watermarking technique, employing randomly initialized classifiers, displayed divergence between watermarked and authentic images, offering a proactive approach to tracing AI-generated content. These methods show promise, but further refinement and broader evaluations are needed for practical application. Nonetheless, they could prove valuable toward detecting and tracing AI-generated content, crucial for combatting misinformation and protecting intellectual property in the era of generative AI.

References

[1] Anonymous. Spotting LLMs with binoculars: Zero-shot detection of machine-generated text. In *Submitted to The Twelfth International Conference on Learning Representations*, 2023. under review. 1

[2] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 - Workshops, Vancouver, BC, Canada, June 17-24, 2023*, pages 843–852. IEEE, 2023. 3

[3] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2023. 1

[4] Churches from Wikipedia Dataset. https://openreview.net/forum?id=QVeMBoRXAo_. 3

[5] Churches Wikipedia 768x768 Dataset. https://drive.google.com/drive/folders/1e08JqlfoxWK1_uiyqGPLadsfsaKP5Jbd?usp=share_link. 3

[6] Ingemar J. Cox, Joe Kilian, Tom Leighton, and Talal Shamoon. Secure spread spectrum watermarking for images, audio and video. In *Proceedings 1996 International Conference on Image Processing, Lausanne, Switzerland, September 16-19, 1996*, pages 243–246. IEEE Computer Society, 1996. 1, 2

[7] dalle 3 Dataset. ttps://huggingface.co/datasets/laion/dalle-3-datase. 3

[8] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 8780–8794, 2021. 3

[9] Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models, 2023. 2

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[11] LSUN Church Dataset. https://huggingface.co/datasets/tglcourse/lsun_church_train. 3

[12] Personal photos Dataset. https://drive.google.com/drive/folders/1q9scsj0mzq0wCL_kkc6NFugFk_TAvlt2?usp=share_link. 3

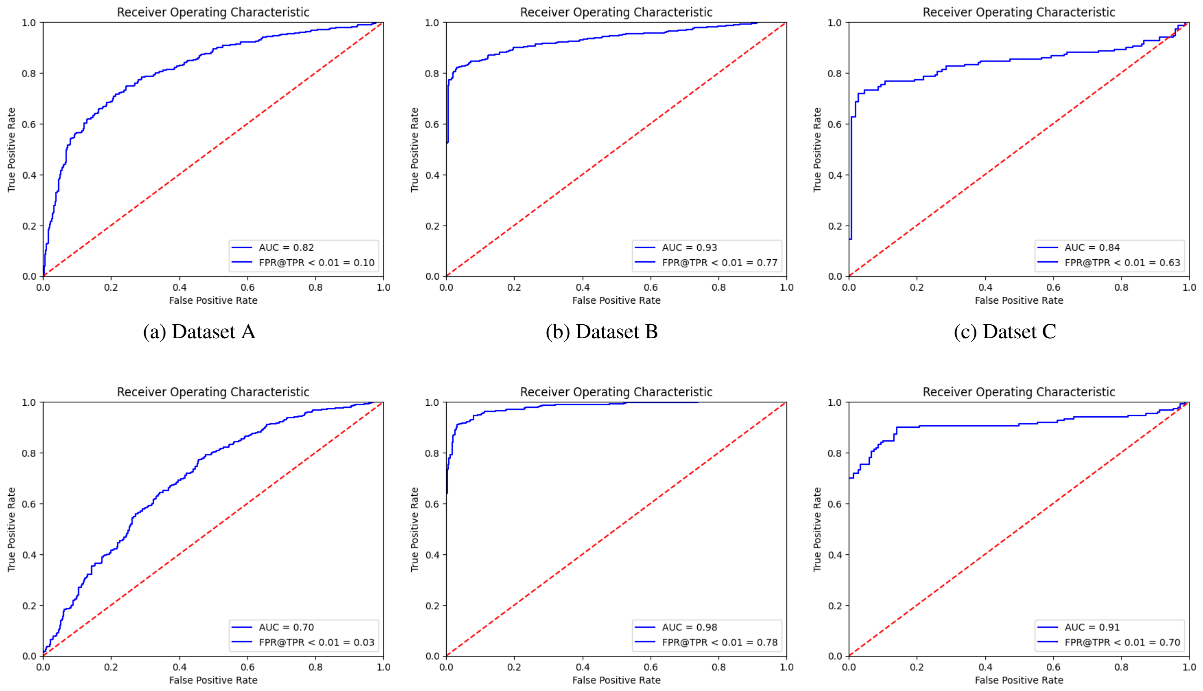
[13] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1

[14] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 4

[15] Joseph Ó Ruanaidh and Thierry Pun. Rotation, translation and scale invariant digital image watermarking. In *Proceedings 1997 International Conference on Image Processing, ICIP '97, Santa Barbara, California, USA, October 26-29, 1997*, pages 536–539. IEEE Computer Society, 1997. 1, 2

[16] Mehrdad Saberi, Vinu Sankar Sadasivan, Keivan Rezaei, Aounon Kumar, Atoosa Chegini, Wenxiao Wang, and Soheil Feizi. Robustness of ai-image detectors: Fundamental limits and practical attacks, 2023. 1, 2

[17] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans,



(a) Dataset A (b) Dataset B (c) Datset C

(d) Dataset A (e) Dataset B (f) Dataset C

Figure 1. Receiver operating characteristic plots for three datasets A, B, C for different pairs of timesteps. Each dataset consists of 300 images of each category: generated and real. The metric FPR@TPR$<$ 0.01 is given in right bottom corner of each plot. **Top row:** the timesteps of the diffusion process applied to the input image are $t_1 = 5$, $t_2 = 20$ (possible timesteps are from the range $[0, \ldots, 1000]$). **Bottom:** the timesteps are set to $t_1 = 5$, $t_2 = 50$.

et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1

[18] Jin S. Seo, Jaap Haitsma, Ton Kalker, and Chang Dong Yoo. A robust image fingerprinting system using the radon transform. *Signal Process. Image Commun.*, 19(4):325–339, 2004. 1, 2

[19] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6048–6058, 2023. 1

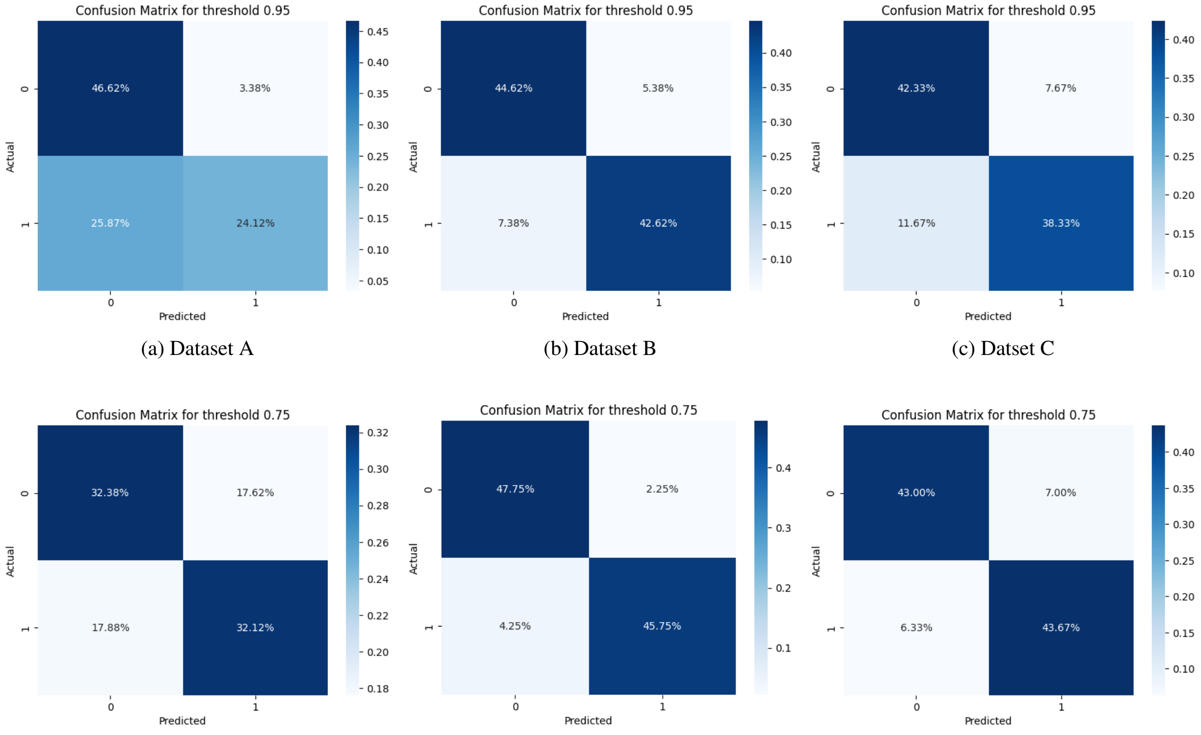
[20] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. *arXiv preprint arXiv:2305.20086*, 2023. 1

[21] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[22] Stable Diffusion v2 generated churches Dataset. https://drive.google.com/drive/folders/1d6TWa5hL2SKjSQszODCR6uBS_6de1H-7?usp=share_link. 3

[23] Stable Diffusion v2 generated, prompt chatGPT Dataset. https://drive.google.com/drive/folders/1d6TWa5hL2SKjSQszODCR6uBS_6de1H-7?usp=share_link. 3

[24] Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust, 2023. 1, 2



(a) Dataset A

(b) Dataset B

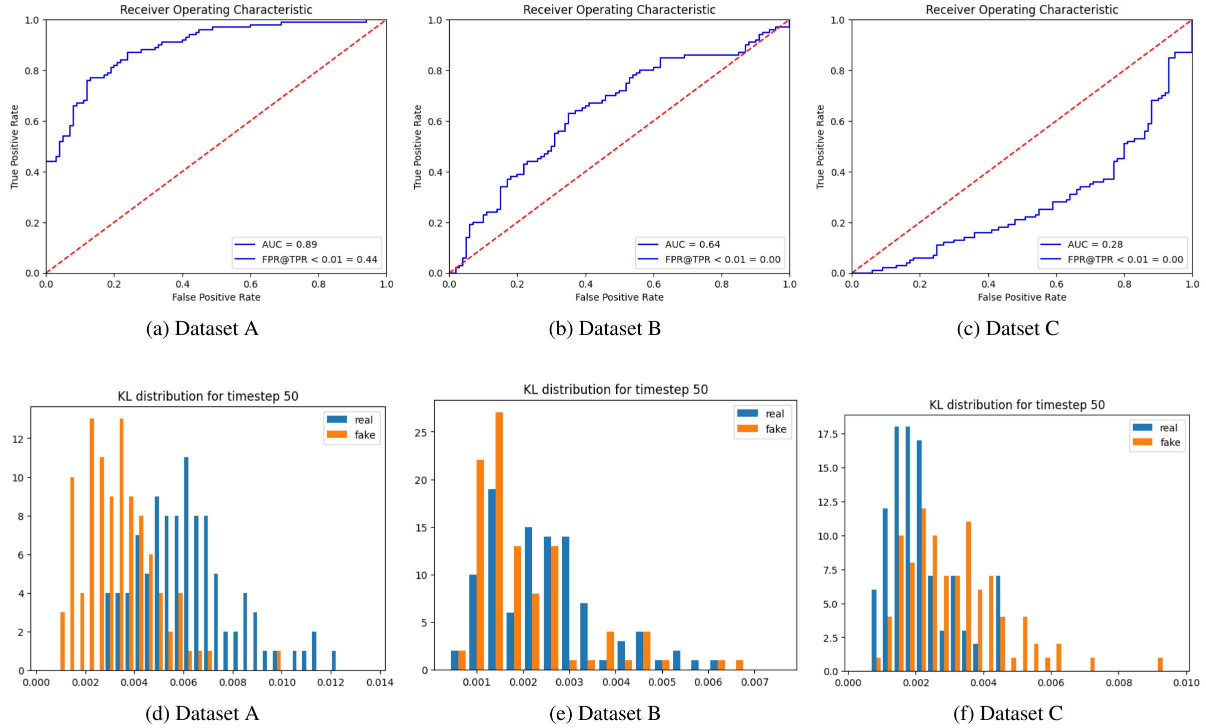
(c) Datset C

(d) Dataset A

(e) Dataset B

(f) Dataset C

Figure 2. The confusion matrices for three datasets A, B, C for different pairs of timesteps. Each dataset consists of 300 images of each category: AI-generated and real. **Top row:** the timesteps of the diffusion process applied to the input image are $t_1 = 5$, $t_2 = 20$ (possible timesteps are from the range $[0, \ldots, 1000]$). **Bottom:** the timesteps are set to $t_1 = 5$, $t_2 = 50$.



(a) Dataset A

(b) Dataset B

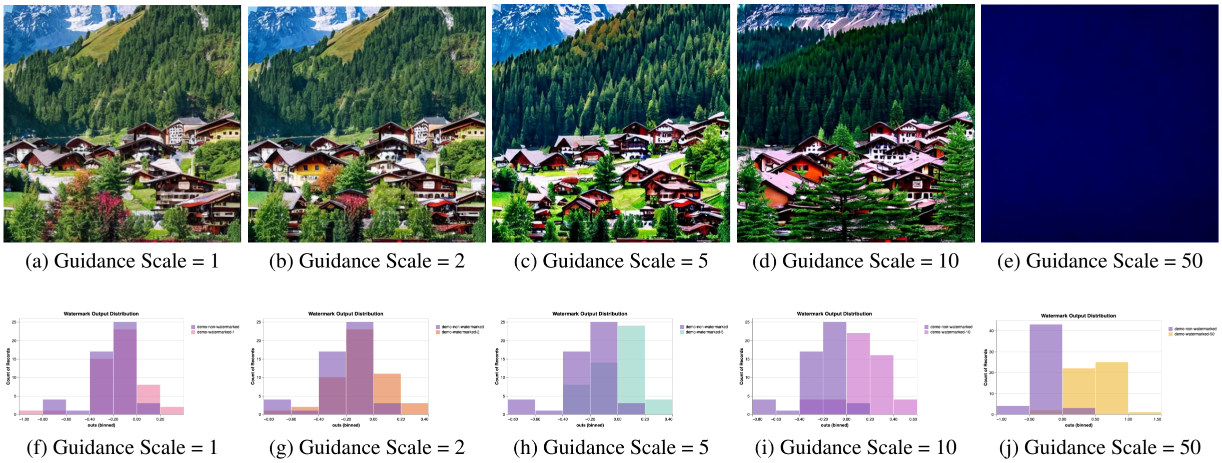
(c) Datset C

(d) Dataset A

(e) Dataset B

(f) Dataset C

Figure 3. Receiver operating characteristic plots (top row) and schematically plotted distributions (bottom row) of the gradient metric evaluated on three datasets A, B, and C, for the timestep $t = 50$. Each dataset consists of 100 randomly chosen images of each category: AI-generated and real. The metric FPR@TPR< 0.01 is given in right bottom corner of each plot.



(a) Guidance Scale = 1

(b) Guidance Scale = 2

(c) Guidance Scale = 5

(d) Guidance Scale = 10

(e) Guidance Scale = 50

(f) Guidance Scale = 1

(g) Guidance Scale = 2

(h) Guidance Scale = 5

(i) Guidance Scale = 10

(j) Guidance Scale = 50

Figure 4. The prompt is ’A peaceful village in the Swiss Alps.’ **Top:** Generated images obtained from 500 steps of DDIM across varied watermarking guidance scales within our guided watermarking method. As the guidance scale increases, image quality diminishes. **Bottom:** The histogram displaying averaged outputs from the watermarking model for watermaked generated images and unwatermarked authentic images. Evidently, with an increase in guidance scale, the divergence between the distributions of the two sets of images grows, aligning with expected outcomes.