# FIRST-PLACE SOLUTION TO NEURIPS 2024 INVISIBLE WATERMARK REMOVAL CHALLENGE

**Fahad Shamshad**[1], **Tameem Bakr**[1], **Yahia Shaaban**[1],
**Noor Hussein**[1,2], **Karthik Nandakumar**[1,2], **Nils Lukas**[1]
[1]Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), UAE
[2]Michigan State University (MSU), USA
{firstname.lastname}@mbzuai.ac.ae

## ABSTRACT

Content watermarking is an important tool for the authentication and copyright protection of digital media. However, it is unclear whether existing watermarks are robust against adversarial attacks. We present the **winning solution** to the NeurIPS 2024 *Erasing the Invisible* challenge, which stress-tests watermark robustness under varying degrees of adversary knowledge. The challenge consisted of two tracks: a black-box and beige-box track, depending on whether the adversary knows which watermarking method was used by the provider. For the **beige-box** track, we leverage an *adaptive* VAE-based evasion attack, with a test-time optimization and color-contrast restoration in CIELAB space to preserve the image's quality. For the **black-box** track, we first cluster images based on their artifacts in the spatial or frequency-domain. Then, we apply image-to-image diffusion models with controlled noise injection and semantic priors from ChatGPT-generated captions to each cluster with optimized parameter settings. Empirical evaluations demonstrate that our method successfully **achieves near-perfect watermark removal** (95.7%) with negligible impact on the residual image's quality. We hope that our attacks inspire the development of more robust image watermarking methods.

## 1 INTRODUCTION

Content watermarking is a widely used technique for embedding imperceptible information into digital media to ensure content authenticity, copyright protection, and traceability (Liu et al., 2024; Zhao et al., 2024). Given that generative AI services can produce unsafe or harmful content at scale, watermarking has become an essential tool for content owners and organizations to combat unauthorized distribution and forgery. The goal of a watermarking method is to hide a signal in generated content that can only be detected with a secret watermarking key, while remaining detectable under normal usage conditions. A robust watermarking scheme must ensure that evading detection requires significantly degrading content quality, making removal infeasible without introducing noticeable artifacts (Zhao et al., 2024; Lukas & Kerschbaum, 2023). However, despite its widespread deployment, watermarking systems remain vulnerable to both unintentional distortions and targeted adversarial attacks aimed at erasing embedded signals while preserving perceptual fidelity (An et al., 2024).

This paper presents our approach to the recent NeurIPS 2024 competition, *Erasing the Invisible: A Stress-Test Challenge for Image Watermarks* (Ding et al., 2024), which assessed the robustness of watermarking methods under two threat models: **beige-box**, where the watermarking methodology was known, and **black-box**, where no prior knowledge was available. Our team developed novel attacks for both settings, **securing first place in both tracks**. The proposed methods combine generative models, frequency-domain manipulations, and fine-tuned variational autoencoders to erase watermarks while preserving image quality. By exposing vulnerabilities in existing watermarking schemes, we aim to inspire the development of more robust defenses against such attacks.

Table 1: Black-box Track Final Leaderboard

| Black-box Track | | | | |
|---|---|---|---|---|
| Rank | Participant | Detection | Quality | Total |
| ① | **Team-MBZUAI** | **0.043** | **0.136** | **0.143** |
| ② | Team-SHARIF | 0.063 | 0.158 | 0.170 |
| ③ | Team-UFL | 0.087 | 0.177 | 0.197 |

Table 2: Beige-box Track Final Leaderboard

| Beige-box Track | | | | |
|---|---|---|---|---|
| Rank | Participant | Detection | Quality | Total |
| ① | **Team-MBZUAI** | **0.037** | **0.153** | **0.157** |
| ② | Team-SONY | 0.050 | 0.176 | 0.183 |
| ③ | Team-SHARIF | 0.127 | 0.222 | 0.256 |


Figure 1: *Top row*: Original watermarked images. *Bottom row*: Images after our attack, with minimal perceptual difference from the originals, showcasing the effectiveness of our method in preserving visual fidelity. Best viewed zoomed in.

## 2 Related Work

**Watermarking.** Content watermarking embeds imperceptible information into images for authentication, copyright protection, and forensic tracking (Qi et al., 2022). Traditional methods rely on spatial or frequency domain manipulations, embedding data into pixel values or transformed coefficients such as DCT, DWT, or DFT (Kang et al., 2003), aiming to balance imperceptibility and robustness. Recent deep learning-based approaches leverage CNNs and generative models to embed watermarks via learned feature representations. Notable methods like StegaStamp (Tancik et al., 2020) and TreeRing (Wen et al., 2024) improve robustness against common distortions and are widely used to protect AI-generated content. However, these methods remain susceptible to adversarial and targeted removal attacks, highlighting the need for more secure watermarking techniques (An et al., 2024).

**Robustness of Watermarks.** Watermarks can be degraded by common distortions such as Gaussian noise, blurring, and compression, while adversarial attacks aim to deliberately exploit model vulnerabilities to remove or distort the watermark with minimal perceptual change (Hwang et al., 2024; Yang et al., 2024; Głuch et al., 2024). To counter these threats, recent work has explored adversarial training, where watermarking models are trained against a range of perturbations to enhance robustness (Huang et al., 2024; Thakkar et al., 2023). In addition, generative models such as autoencoders and diffusion models have been utilized to embed robust watermarks that preserve visual fidelity even under adversarial conditions. Despite recent advances, balancing robustness and imperceptibility remains challenging, particularly when attackers have partial or full knowledge of the watermarking method (Ma et al., 2025; Fairoze et al., 2025). The NeurIPS 2024 challenge (Ding et al., 2024) offers a benchmark to assess the resilience of state-of-the-art approaches.

## 3 Proposed Attack

Our approach to the NeurIPS 2024 *Erasing the Invisible* challenge employs a multi-stage attack strategy to effectively remove invisible watermarks while preserving high image quality. The effectiveness of our method is evaluated using two metrics: **(i)** an *attack metric* that quantifies the success of watermark removal and **(ii)** a set of *qualitative metrics* that assess the perceptual quality of the modified images. The attack metric $A$ is the **True Positive Rate (TPR) at 0.1% False Positive Rate (FPR)**, where watermark detection is determined based on the decoded message

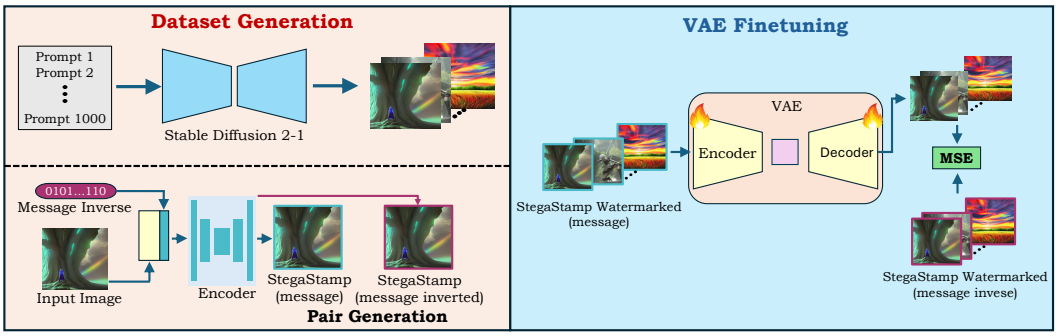


Figure 2: Overview of our dataset generation and VAE-based watermark removal pipeline for StegaStamp watermarks. We generate 1,000 images from Stable Diffusion 2-1 using publicly available prompts, embedding each with a random binary message and its inverse via a StegaStamp encoder to create paired watermarked images. A fine-tuned Variational Autoencoder (VAE) minimizes Mean Squared Error (MSE) loss between pairs, effectively suppressing watermark artifacts while preserving perceptual quality.

distance $d(m, m')$ falling below a threshold set by the 0.001st percentile of distances from 10,000 unwatermarked images. The quality of watermark-removed images is assessed using a combination of perceptual and fidelity metrics, including PSNR, SSIM, NMI, FID, CLIPFID (Kynkäänniemi et al., 2022), LPIPS (Zhang et al., 2018), Delta Aesthetics, and Delta Artifacts (Xu et al., 2024). *We can query the competition leaderboard up to five times per day to evaluate our submission performance*. Below, we provide details of our proposed attack for both the beige-box and black-box tracks.

### 3.1 BEIGE-BOX TRACK

In the beige-box track, we had access to the algorithmic description of the watermarking method used to embed invisible watermarks in the provided images, but not to its hyper-parameters or the generating model. This prior knowledge allowed us to design an adaptive attack strategy designed against the specific watermarking method, similar to (Lukas et al., 2024; Diaa et al., 2024). The organizers released 300 watermarked images, divided into two categories: Half of the images are watermarked using a modified StegaStamp algorithm, while the remaining images are watermarked using a variant of the TreeRing (Wen et al., 2024) watermarking method.

Table 3: Performance comparison of our watermark removal pipeline. **Det.** refers to the detection metric (lower is better), **Qual.** measures image quality. Progressive enhancements through test-time optimization and color/contrast adjustment improve quality.

| Method | Det. | Qual. | Overall |
|---|---|---|---|
| VAE Finetune | **0.023** | 0.192 | 0.193 |
| + Test-Time Opt. | 0.033 | 0.161 | 0.165 |
| + Color/Contrast | 0.037 | **0.153** | **0.157** |

#### 3.1.1 STEGASTAMP WATERMARK

For the StegaStamp-based watermark removal, we employed a three-step approach consisting of **dataset generation**, **VAE-based watermark removal**, and **post-processing** for quality enhancement (see Fig. 2). This structured methodology enabled effective suppression of the embedded watermark while preserving the perceptual quality of the images.

***Dataset Generation:*** We first curated a comprehensive training dataset leveraging 1 000 text prompts from the Hugging Face Stable-Diffusion-Prompts dataset (Gustavosta, 2024). Using these prompts, we generated corresponding images via Stable Diffusion 2-1 with a guidance scale of 7.5 and 50 inference steps. Each generated $512^2$ image was resized to $400^2$ pixels using bilinear interpolation before being processed through a pretrained StegaStamp model from the WAVES repository. The key aspect of our dataset preparation involved creating image pairs where each original image was encoded with both a 100-bit binary message $m$ sampled uniformly at random and its inverse $1 - m$, resulting in a dataset of 1,000 paired examples that captured watermarking artifacts.

***VAE Finetuning:*** The core of our attack framework centers on a Variational Autoencoder (VAE) that was adaptively tuned against a specific watermarking method. Let $x_w$ denote a watermarked image containing the original binary message and $x_i$ represent the corresponding image with the inverted message. The VAE consists of an encoder $E_\theta$ and decoder $D_\phi$, where we leverage the pretrained architecture from the SDXL diffusion model. During fine-tuning, we optimize both components to minimize the mean squared error loss:

$$\mathcal{L}(\theta, \phi) = \|D_\phi(E_\theta(x_w)) - x_i\|_2^2, \tag{1}$$

where $D_\phi(E_\theta(x_w))$ represents the reconstructed image from the watermarked input, and $x_i$ is the target image containing the inverted message. We optimize this objective using Adam optimizer with learning rate $\alpha = 1 \times 10^{-5}$ for 10 epochs with a batch size of 16. To stabilize training, we employ gradient clipping with a maximum norm of 1.0. Model training was performed on an NVIDIA A6000 GPU (48GB VRAM) and completed in under two GPU hours.

***Post-Processing:*** To address the slight quality degradation in the VAE outputs in the ***VAE Finetuning*** stage, we implemented a two-stage post-processing pipeline with the aim to enhance the image quality without re-introducing the removed watermark. The first stage involved test-time VAE optimization using a pretrained VAE from the Stable Diffusion Refiner model. Let $x_r$ denote the output from the VAE Finetuning stage and $x_w$ represent the original watermarked image. During test-time optimization, we fine-tune a Refiner VAE parameters $\{\theta, \phi\}$ for each image to minimize:

$$\mathcal{L}_{total} = \underbrace{\|D_\phi(E_\theta(x_r)) - x_w\|^2}_{\text{MSE Loss}} + \underbrace{\mathcal{L}_{\text{LPIPS}}(D_\phi(E_\theta(x_r)), x_w)}_{\text{Perceptual Loss}} + \underbrace{0.5(1 - SSIM(D_\phi(E_\theta(x_r)), x_w))}_{\text{Structural Similarity Loss}} \tag{2}$$

The second stage performs color and contrast transfer in CIELAB space. Let $x_{opt}$ denote the output from test-time optimization, and $\{L_{opt}, a_{opt}, b_{opt}\}$ and $\{L_w, a_w, b_w\}$ represent the CIELAB components of $x_{opt}$ and the watermarked image $x_w$ respectively. For color transfer, we preserve the luminance channel from the optimized image while adopting the chrominance components from the watermarked image as $x_c = \mathcal{F}_{RGB}(L_{opt}, a_w, b_w)$ where $\mathcal{F}_{RGB}$ denotes conversion from CIELAB to RGB space. Subsequently, we perform contrast transfer by matching the statistical moments of the luminance channel as $L_{\text{final}} = \frac{\sigma_w}{\sigma_c}(L_c - \mu_c) + \mu_w$, where $\{\mu_c, \sigma_c\}$ and $\{\mu_w, \sigma_w\}$ are the mean and standard deviation of the luminance channels of $x_c$ and $x_w$ respectively. The final image is constructed as $x_{\text{final}} = \mathcal{F}_{RGB}(L_{\text{final}}, a_w, b_w)$, ensuring high perceptual quality without reintroducing the watermark. For impact of different components in our pipeline, see Table 3.

### 3.1.2 TREERING WATERMARK

For images embedded with TreeRing watermarks, we identified a significant vulnerability to phase attacks in the frequency domain. Notably, modifying the Fourier phase spectrum corresponds to simple translation in the image spatial domain. Leveraging this insight, our approach implements a straightforward yet effective spatial domain transformation as $x_{\text{shifted}} = \mathcal{T}(x_w, \Delta x)$ where $x_w$ is the watermarked image, $\mathcal{T}$ denotes a horizontal translation operator, and $\Delta x = 7$ pixels is the empirically determined optimal shift distance that balances watermark removal effectiveness with image quality preservation. To further mitigate translation-induced artifacts, we selectively restore the leftmost $\Delta x$ pixels from the original image:

$$x_{\text{final}}(i, j) = \begin{cases} x_w(i, j) & \text{if } j < \Delta x \\ x_{\text{shifted}}(i, j) & \text{otherwise,} \end{cases} \tag{3}$$

where the leftmost $\Delta x$ pixels are restored from the original watermarked image. This approach effectively removes TreeRing watermarks while preserving visual quality.

## 3.2 BLACK-BOX TRACK

For the black-box track, where no prior knowledge of the specific watermarking method was available, we adopted a data-driven approach to identify the most likely watermarking method

based on their spatial and frequency domain characteristics. We partitioned the 300 watermarked images into four distinct clusters (see Fig. 3): (**1**) images without noticeable artifacts, (**2**) images exhibiting border-like artifacts in the spatial domain, (**3**) images with circular patterns in the Fourier magnitude spectrum, and (**4**) images containing square patterns in the Fourier magnitude spectrum. This clustering enabled us to develop targeted removal strategies for each group.

**Image-to-Image Diffusion Model:** To remove the watermarks, we leveraged the **image-to-image** capabilities of the *Stable Diffusion Refiner model* (Meng et al., 2021) as a core component of our strategy. The diffusion process consists of two stages: (i) forward diffusion, which gradually corrupts an image by adding noise, and (ii) reverse diffusion, where a denoising network reconstructs the clean image by iteratively removing noise. Given a watermarked image $x_w$, the forward process perturbs it using a *predefined noise schedule* $\alpha_t$, where the *strength parameter* $s \in [0, 1]$ determines the *starting point* of the noise injection as $x_t = \sqrt{\alpha_t} x_w + \sqrt{1 - \alpha_t}\epsilon_t$, where $\epsilon_t \sim \mathcal{N}(0, I)$. Here, $s$ controls the initial noise level via $\alpha_1 = 1 - s$, determining how much of the original image is retained (see Fig. 4). A *smaller* $s$ preserves more of the original structure, enabling *subtle artifact suppression*, while a *larger* $s$ introduces greater noise, allowing for *aggressive watermark removal* at the cost of potential content modification. The *reverse diffusion process* iteratively removes noise through a learned denoising function $\epsilon_\theta$, reconstructing the final refined image as $x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \sqrt{1 - \alpha_t}\epsilon_\theta(x_t, t)\right)$.

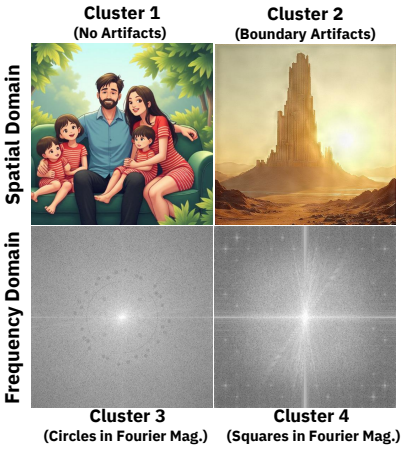


Figure 3: Spatial and frequency-based clustering for blackbox track.

We configured the diffusion process with $T = 500$ inference steps, using **image captions generated by ChatGPT** as text prompts and setting the **guidance scale to 1**. This allowed the model to incorporate *semantic priors* alongside the original image structure, facilitating more effective watermark suppression while preserving content fidelity. By optimizing $s$ across clusters, we achieved a robust trade-off between artifact removal and content preservation, ensuring that the refined images maintained their perceptual quality while successfully evading watermark detection.

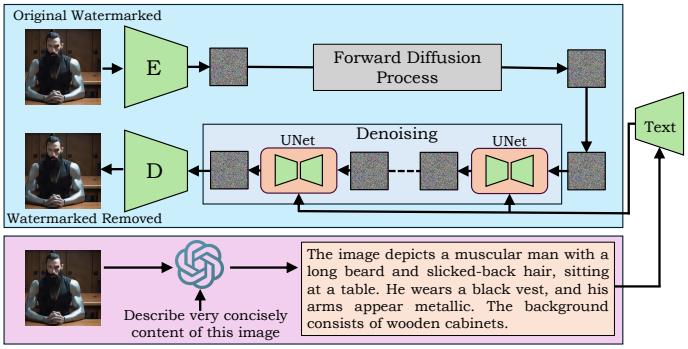


Figure 4: Proposed image-to-image diffusion-based pipeline for watermark removal, ensuring effective watermark suppression while preserving perceptual quality.

**Cluster-Specific Solutions:** Based on empirical analysis, we developed adaptive attacks for each cluster. **For Cluster 1 (no noticeable artifacts**, we employed the *image-to-image diffusion model* with a relatively high strength parameter ($s = 0.16$) to enable more aggressive watermark removal. **For clusters 2 and 3**, exhibiting structured patterns, we developed a three-stage pipeline as detailed in Sec. 3.1.1: (i) dataset generation with synthetic watermarks, (ii) VAE fine-tuning minimizing reconstruction loss between watermarked inputs and clean targets, and (iii) post-processing enhancement through color and contrast transfer. **For cluster 4 (square patterns)**, we implemented a hybrid approach combining *image-to-image diffusion* with lower strength ($s = 0.04$) followed by a 7-pixel horizontal translation and boundary pixel restoration. We found this cluster-specific strategy achieved good results, with significant improvements in the qualitative metrics compared to uniform parameter settings across clusters.

## REFERENCES


Bang An, Mucong Ding, Tahseen Rabbani, Aakriti Agrawal, Yuancheng Xu, Chenghao Deng, Sicheng Zhu, Abdirisak Mohamed, Yuxin Wen, Tom Goldstein, et al. Waves: Benchmarking the robustness of image watermarks. In *Forty-first International Conference on Machine Learning*, 2024.

Abdulrahman Diaa, Toluwani Aremu, and Nils Lukas. Optimizing adaptive attacks against content watermarks for language models. *arXiv preprint arXiv:2410.02440*, 2024.

Mucong Ding, Tahseen Rabbani, Bang An, Souradip Chakraborty, Chenghao Deng, Mehrdad Saberi, Yuxin Wen, Xuandong Zhao, Mo Zhou, Anirudh Satheesh, et al. Erasing the invisible: A stress-test challenge for image watermarks. In *NeurIPS 2024 Competition Track*, 2024.

Jaiden Fairoze, Guillermo Ortiz-JimÃŠnez, Mel Vecerik, Somesh Jha, and Sven Gowal. On the difficulty of constructing a robust and publicly-detectable watermark. *arXiv preprint arXiv:2502.04901*, 2025.

Grzegorz Głuch, Berkant Turan, Sai Ganesh Nagarajan, and Sebastian Pokutta. The good, the bad and the ugly: Watermarks, transferable attacks and adversarial defenses. *arXiv preprint arXiv:2410.08864*, 2024.

Gustavosta. Stable diffusion prompts dataset. https://huggingface.co/datasets/Gustavosta/Stable-Diffusion-Prompts, 2024. Accessed: February 2025.

Huayang Huang, Yu Wu, and Qian Wang. Robin: Robust and invisible watermarks for diffusion models with adversarial optimization. *Advances in Neural Information Processing Systems*, 37: 3937–3963, 2024.

Dongjun Hwang, Sungwon Woo, Tom Gao, Raymond Luo, and Sunghwan Baek. Invisible watermarks: Attacks and robustness. *arXiv preprint arXiv:2412.12511*, 2024.

Xiangui Kang, Jiwu Huang, Yun Q Shi, and Yan Lin. A dwt-dft composite watermarking scheme robust to both affine transform and jpeg compression. *IEEE transactions on circuits and systems for video technology*, 13(8):776–786, 2003.

Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of imagenet classes in fr\'echet inception distance. *arXiv preprint arXiv:2203.06026*, 2022.

Aiwei Liu, Leyi Pan, Yijian Lu, Jingjing Li, Xuming Hu, Xi Zhang, Lijie Wen, Irwin King, Hui Xiong, and Philip Yu. A survey of text watermarking in the era of large language models. *ACM Computing Surveys*, 57(2):1–36, 2024.

Nils Lukas and Florian Kerschbaum. {PTW}: Pivotal tuning watermarking for {Pre-Trained} image generators. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 2241–2258, 2023.

Nils Lukas, Abdulrahman Diaa, Lucas Fenaux, and Florian Kerschbaum. Leveraging optimization for adaptive attacks on image watermarks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=O9PArxKLe1.

Xingjun Ma, Yifeng Gao, Yixu Wang, Ruofan Wang, Xin Wang, Ye Sun, Yifan Ding, Hengyuan Xu, Yunhao Chen, Yunhan Zhao, et al. Safety at scale: A comprehensive survey of large model safety. *arXiv preprint arXiv:2502.05206*, 2025.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

Wang Qi, Bei Yue, Chen Wangdu, Pan Xinghao, Cheng Zhipeng, Wang Shaokang, Wang Yizhao, and Wang Chenwei. An overview on digital content watermarking. In *Signal and Information Processing, Networking and Computers: Proceedings of the 8th International Conference on Signal and Information Processing, Networking and Computers (ICSINC)*, pp. 1311–1318. Springer, 2022.

Matthew Tancik, Ben Mildenhall, and Ren Ng. Stegastamp: Invisible hyperlinks in physical photographs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2117–2126, 2020.

Janvi Thakkar, Giulio Zizzo, and Sergio Maffeis. Elevating defenses: Bridging adversarial training and watermarking for model resilience. *arXiv preprint arXiv:2312.14260*, 2023.

Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-rings watermarks: Invisible fingerprints for diffusion images. *Advances in Neural Information Processing Systems*, 36, 2024.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024.

Pei Yang, Hai Ci, Yiren Song, and Mike Zheng Shou. Can simple averaging defeat modern watermarks? *Advances in Neural Information Processing Systems*, 37:56644–56673, 2024.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Xuandong Zhao, Sam Gunn, Miranda Christ, Jaiden Fairoze, Andres Fabrega, Nicholas Carlini, Sanjam Garg, Sanghyun Hong, Milad Nasr, Florian Tramer, et al. Sok: Watermarking for ai-generated content. *arXiv preprint arXiv:2411.18479*, 2024.