Robustness of AI-Generated Image Detection Against Localized Inpainting Attacks

Oğuz AKIN
Saarland University
ogak00001@stud.uni-saarland.de
August 17, 2025

***Abstract*—The increase of high-fidelity AI-generated images (AIGI) makes it necessary for reliable detection methods to get ahead of misinformation waves. However, the robustness of these detectors against well-known post-processing manipulations like image inpainting remains significantly ignored. This thesis evaluates the robustness of three main AIGI detection paradigms: proactive watermarking, reactive passive detection, and generalized training-free methods. Using the state-of-the-art SEMI-TRUTHS benchmark, this work will systematically measure the performance degradation of six well-known detectors. The expected results will identify which detection methodologies are the most resilient and robust to inpainting in order to provide a clear benchmark to guide the development of more trustworthy detection systems in the future.**

I. Introduction

A shift in paradigm has been seen in the last five years in the domain of digital content creation, stemming from powerful generative models like Generative Adversarial Networks (GANs) and Diffusion Models. These models can produce AI-generated images (AIGI) of such quality and realism that they are sometimes indistinguishable from real photographs to the human eye. Such power, while game changing for creative industries, also comes with significant societal risks such as the spread of sophisticated disinformation and the eroding of public trust in digital media as with great power, comes great responsibility.

To address this, the research community came up with a very rich array of powerful AIGI detection methods. However, the adversarial setting of this domain means that as detectors improve themselves, so do the augmentation methods to evade getting detected. This turns the situation into a cat and mouse game where a surging and practical danger is the use of post-processing manipulations on images. An adversary that is aware that their generated image might be caught can apply simple modifications to disrupt the detector's underlying assumptions and avoid getting detected.

Image inpainting is defined as a technique for plausibly filling in missing or corrupted parts of an image and it can be a particularly potent attack vector. An adversary can use inpainting for two things. The first one would be to destroy a proactive signal like an embedded watermark. Furthermore, another would be to disrupt the subtle generative artifacts that reactive passive detectors depend on for detection. Despite its simplicity and effectiveness, Despite simplicity of inpainting on the robustness of different AIGI detection families, their impact has not been yet critically and systematically evaluated in a deeper sense. This prospective danger is particularly disturbing because, unlike complex adversarial attacks that require specialized knowledge, image inpainting is a widely accessible, "one-click" tool that can be used to completely 'sanitize' a generated image and to possibly evade a detector with minimal effort.

With this thesis, it is aimed to fill that gap by addressing the following research question: **How robust are different classes of AI-generated image detection methods to adversarial image inpainting?** To answer this, we pursue the following objectives:

1) To select and evaluate representative models from three main AIGI detection categories: watermarking, passive detection, and training-free methods.
2) To conduct experiments using a state-of-the-art benchmark dataset of inpainted images, SEMI-TRUTHS.
3) To systematically measure the degradation in detector performance when faced with inpainting attacks of varying magnitude.
4) To analyze and identify which detection methodologies are most, and least, resilient to this form of manipulation.

II. Related Work

The detection of AI-generated images is an exponentially evolving field, which makes it necessary to have a clear taxonomy of approaches. The methods can be categorized overall into three distinct paradigms, each of which will be evaluated in this thesis.

A. Watermarking Methods

Watermarking is a proactive defense mechanism where an invisible signature is sown into an image behind the scenes to allow for later provenance verification. These techniques are known as "proactive" as they require cooperation from the image generator. We investigate two modern systems representing distinct philosophies of a unified aim:

- **Stable Signature** [1] is an *in-processing* method that embeds the watermark directly into the model's latent space during the diffusion generation process. This approach aims to deeply integrate the signal into the image's fundamental structure which makes it theoretically more difficult to remove without degrading the image too much.
- **Tree-Ring Watermark** [2] is a *post-processing* method that applies a subtle pattern to the final pixel image

after it has been generated. This approach offers greater flexibility but it may also be more open to surface-level attacks that directly manipulate pixel values.

B. Passive Detection Methods

Passive detectors are reactive AI-based ”detectives” that are trained to find out the unintentional artifacts or ”fingerprints” left by generative models. These models analyze an image without any prior knowledge or modification. In this thesis, we evaluate two modern and contrasting approaches:

- **UniversalFakeDetect** [3] operates on a *high-level semantic* feature space. It uses a large pre-trained vision-language model (CLIP) to determine if the overall meaning, object relationships, and composition of an image are more consistent with the neighborhood of real images or fake images it has been shown in training.
- **Corvi et al. (2023)** [4] represents a modern and *low-level artifact-based* approach. It is a CNN-based classifier that is specifically trained to identify the uniquely subtle fingerprints and frequency imbalances characteristic of modern diffusion models, which makes it especially a relevant baseline for the current generative landscape.

C. Training-Free Methods

These new methods do not require training on large datasets of ”fake” images, instead leveraging fundamental properties of generative models. They are known to be more generalizable to novel generator architectures. We evaluate two contrasting philosophies based on reconstruction error:

- **DIRE** [5] uses a pre-trained *diffusion model* for reconstruction. It functions on the principle that diffusion models can reconstruct their ”native” generated images with very low error, while real-world photographs, with their inherent noise and complexity, will result in higher reconstruction error.
- **AEROBLADE** [6] uses a standard *autoencoder* for reconstruction. It works on the assumption that ”cleaner” and less complex AI images can be compressed and decompressed with lower error than noisy and high-entropy real images.

III. Methodology

This thesis will follow a quantitative experimental design to evaluate the robustness of various AI-Generated Image (AIGI) detection methods against localized image inpainting. The methodology is composed of four key areas: (1) Datasets and Preparation; (2) Selected Detection Methods; (3) Experimental Protocol; and (4) Evaluation Metrics.

A. Datasets and Preparation

To ensure an accurate and reproducible evaluation, this thesis will make use of two publicly available large-scale datasets. Each dataset serves a distinct and very important role in the experimental protocol.

1) Primary Benchmark: SEMI-TRUTHS: The center of the robustness evaluation will be conducted on the **SEMI-TRUTHS** dataset [7]. This benchmark is uniquely suited for our thesis objectives, as it was specifically designed to evaluate detector robustness against localized AI-driven modifications. Its key features include:

- **Diverse Content:** It contains more than 27,000 real images and 1.4 million AI-augmented counterparts, sourced from multiple domains, ensuring that our findings are generalizable and the subset we will use is very diverse.
- **Rich Metadata:** Each augmented image is accompanied by extensive metadata, most importantly the **Area Ratio** and the **Semantic Magnitude**. This metadata is the bedrock for a deep analysis of detector failure modes.

In our experiments, the original real images from SEMI-TRUTHS will serve as the ground-truth ”real” class, while the AI-augmented images will be used as the ”inpainted fake” class for the main robustness test.

2) Baseline Benchmark: DiffusionDB: To establish a performance baseline for each detector, we will use the **DiffusionDB** dataset [8]. This dataset contains millions of high-quality, fully AI-generated images. A handpicked subset of these images will be used as the ”complete fake” class in our baseline experiment, which will allow us to measure each detector’s maximum performance under ideal conditions. This will provide a bottom-line reference point from which we will measure the performance drop caused by the more subtle inpainting attacks.

B. Selected Detection Methods for Evaluation

We will evaluate six models, selecting representative methods from the three main AIGI detection paradigms. This selection enable us to evaluate a mix of classic baselines and state-of-the-art counterparts, with all models being publicly available and usable without retraining. The selected models are summarized in Table I. Specifically, Stable Signature was chosen as it is an example of a state-of-the-art latent-space watermarking technique, which is in contrast with the post-processing, pixel-space approach of Tree-Ring Watermark. For passive detection, UniversalFakeDetect represents a high-level semantic methodology, while the method by Corvi et al. was selected as a potent, low-level detector trained on the subtle artifacts of modern diffusion models. Finally, DIRE and AEROBLADE were chosen as they represent two separate, state-of-the-art philosophies in training-free detection. All models will be implemented using their official public codebases and pre-trained weights to ensure reproducibility and reflect their intended performance.

C. Experimental Protocol

The experiment is divided into two separate workflows.

1) Workflow for Passive and Training-Free Detectors: This workflow consists of two stages to specifically measure robustness against inpainting, following standard evaluation practices in the field [3], [6]:

TABLE I
THE SIX SELECTED DETECTOR MODELS, CATEGORIZED BY THEIR DETECTION PARADIGM, DATASET USAGE, AND ROLE IN THE THESIS. PERFORMANCE WILL BE EVALUATED USING AUC AND ATTACK SUCCESS RATE (ASR).

Category	Model Name	Dataset Usage	Description & Role in Thesis
Watermarking	Stable Signature & Tree-Ring Watermark	Generated images which are then attacked with the LaMa inpainting model.	Evaluating two distinct philosophies: an in-processing, latent-space system versus a post-processing, pixel-space system.
Passive Detection	UniversalFakeDetect & Corvi et al. (2023)	**Baseline:** SEMI-TRUTHS (Real) vs. DiffusionDB (Fake). **Robustness:** SEMI-TRUTHS (Real vs. Inpainted).	Comparing two modern passive philosophies: a high-level semantic detector vs. a low-level, artifact-based detector designed for diffusion models.
Training-Free	DIRE & AEROBLADE	**Baseline:** SEMI-TRUTHS (Real) vs. DiffusionDB (Fake). **Robustness:** SEMI-TRUTHS (Real vs. Inpainted).	Comparing two separate and state-of-the-art reconstruction-based detection philosophies (Diffusion vs. Autoencoder).

1) **Baseline Performance Test:** The four detectors (UniversalFakeDetect [3], Corvi et al. [4], DIRE [5], and AEROBLADE [6]) will be evaluated on distinguishing real images that are inside the real image subset of SEMI-TRUTHS from fully AI-generated images within the DiffusionDB dataset. This defines the upper bound of each detector's performance, referred to as the Baseline AUC.
2) **Inpainting Robustness Test:** The same four detectors will then be evaluated on differentiating the real images inside SEMI-TRUTHS from the AI-augmented (inpainted) images within SEMI-TRUTHS.
3) **Analysis:** The performance from the robustness test will be compared to the baseline. The change in performance will be correlated with the Area Ratio and Semantic Magnitude metadata to identify specific failure conditions.

2) Workflow for Watermarking Systems: This workflow follows a "generate-attack-detect" loop for the two watermarking systems (Stable Signature [1] and Tree-Ring [2]):

1) **Generate:** For the Stable Signature system, a watermarked image will be generated directly. For the Tree-Ring system, the watermark will be applied on a base image generated using a standard Stable Diffusion model [9].
2) **Attack:** The watermarked image will be attacked using a standard inpainting model, **LaMa** [10], to erase a semantically significant object.
3) **Detect:** The system's corresponding detector will be run on both the original and attacked images to determine if the watermark survived the inpainting attack.

D. Evaluation Metrics

To provide a comprehensive assessment, we will use a set of primary and secondary metrics tailored to each experimental workflow.

1) Metrics for Passive and Training-Free Detectors:

- **Primary Classification Metric (AUC):** The Area Under the ROC Curve will be the primary metric. AUC provides a comprehensive measure of a model's ability to distinguish between classes across all possible decision thresholds, making it robust and standard for academic benchmarks.
- **Primary Robustness Metric (ΔAUC):** The core measure of robustness will be the AUC Drop, calculated as: 'Baseline AUC - Robustness Test AUC'. A lower ΔAUC indicates a more robust detector.
- **Secondary Robustness Metric (ASR):** To directly measure the success of the inpainting attack, we will also calculate the Attack Success Rate. For this, ASR is defined as the **False Negative Rate**: the percentage of inpainted images that are incorrectly classified as "Real."

2) Metrics for Watermarking Systems:

- **Primary Robustness Metric (ASR):** For the watermarking systems, the primary metric will be the Attack Success Rate. ASR is the percentage of attacked images in which the watermark **cannot be successfully detected**. A higher ASR indicates a more effective attack and a less robust watermark.

IV. CONCLUSION AND FUTURE WORK

This thesis aims to form a comprehensive analysis of AI-generated image detector robustness against the adversarial image inpainting attacks. Using state-of-the-art benchmarks and a carefully selected group of detectors covering watermarking, passive, and training-free paradigms, this work aims to provide a clear hierarchy of robustness among the dominant detection philosophies.

A. Expected Contributions

The expected results of this research will provide many important contributions to the field of media forensics. From this thesis, it is expected that training-free methods, especially those that are based on reconstruction error, will demonstrate the highest resilience to inpainting, as they do not rely on specific, editable artifacts. Furthermore, we expect watermarking methods to be the most fragile, especially post-processing techniques like Tree-Ring, as their signals can be directly deleted by a spatial attack like inpainting.

The analysis, correlated with the SEMI-TRUTHS metadata, will quantify these vulnerabilities and underscore the specific conditions (e.g., inpainting area size $> 30\%$) under

which different detectors fail. At the end of it all, the findings will serve as an important benchmark for the field, guiding the development of more secure and trustworthy detection systems in an era of rising synthetic media.

B. Future Work

Building on the findings of this thesis, many different perspectives for future research can be explored. This includes going deeper into the evaluation of a wider range of post-processing attacks, such as compression, noise addition, and geometric transformations. Furthermore, the insights gained could inform the development of a new and *inpainting-aware* detection model that is explicitly designed to be resistant in detection to such localized manipulations.

REFERENCES

[1] P. Fernandez, G. Couairon, H. Jégou, M. Douze, and T. Furon, "The stable signature: Rooting watermarks in latent diffusion models," *arXiv preprint arXiv:2303.15435*, 2023.

[2] Y. Wen, J. Kirchenbauer, J. Geiping, and T. Goldstein, "Tree-Ring Watermarks: Fingerprints for Diffusion Images that are Invisible and Robust," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[3] U. Ojha, Y. Li, and Y. J. Lee, "Towards Universal Fake Image Detectors that Generalize Across Generative Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 24 480–24 489.

[4] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, "On the detection of synthetic images generated by diffusion models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[5] Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hu, H. Chen, and H. Li, "DIRE for Diffusion-Generated Image Detection," *arXiv preprint arXiv:2303.09295*, 2023.

[6] J. Ricker, D. Lukovnikov, and A. Fischer, "AEROBLADE: Training-Free Detection of Latent Diffusion Images Using Autoencoder Reconstruction Error," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9130–9140.

[7] A. Pal, J. Kruk, M. Phute, M. Bhattaram, D. Yang, D. H. Chau, and J. Hoffman, "Semi-Truths: A Large-Scale Dataset of AI-Augmented Images for Evaluating Robustness of AI-Generated Image detectors," in *Thirty-eighth Conference on Neural Information Processing Systems*, 2024. [Online]. Available: https://openreview.net/forum?id=g3l4C45VnS

[8] Z. J. Wang, E. Montoya, D. Munechika, H. Yang, B. Hoover, and D. H. Chau, "DiffusionDB: A Large-scale Prompt Gallery Dataset for Text-to-Image Generative Models," *arXiv preprint arXiv:2210.14896*, 2022.

[9] 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)*, 2022, pp. 10 684–10 695.

[10] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky, "Resolution-robust large mask inpainting with fourier convolutions," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 2149–2159.