Watermark Robustness under Adversarial Attacks for Deepfake Detection*

*Working draft

Cat Lewin
School of Science and Engineering
University of Missouri-Kansas City
Kansas City, USA
clfkd@umsystem.edu

***Abstract*—As deepfakes and digital media tampering becomes increasingly sophisticated, invisible watermarking has emerged as a potential tool for ensuring image authenticity and traceability. This paper investigates the robustness of three invisible watermarking methods: DWT-DCT, DWT-DCT-SVD, and RivaGAN, under a variety of adversarial image perturbations. We developed a threshold-based attack evaluation framework to test the success of watermark decoding in a range of transformations. Experiments were conducted on a set of 512×512 face images from the Unsplash dataset, with some methods also tested on higher resolution originals. Our results show that the deep learning-based RivaGAN model exhibits superior robustness across most attack types, particularly under severe JPEG and crop distortions, while classical methods struggle under geometric and low-resolution conditions. Unexpectedly, some watermarks became decodable only after specific transformations, suggesting possible alignment effects that are worth exploring in future work. In addition, we introduce a perceptual similarity analysis using the LPIPS metric to identify attacks that degrade watermark performance while remaining visually inconspicuous. These findings underscore vulnerabilities in watermarking methods for proactive deepfake detection and motivate continued evaluation under generative and hybrid attack scenarios.**

***Index Terms*—adversarial image perturbations, invisible watermarking, deepfake detection, perceptual robustness**

I. Introduction

The proliferation of synthetic media, particularly deep fakes, has created serious challenges to the authentication of digital content, the protection of privacy, and the mitigation of misinformation. As generative AI models improve in realism and accessibility, malicious actors can now manipulate or fabricate human likenesses with minimal effort. In response, researchers have begun exploring proactive methods for verifying media integrity, with invisible watermarking emerging as a promising direction. By embedding imperceptible signals into image content, these techniques aim to verify authenticity even after downstream manipulations.

However, the robustness of invisible watermarks under adversarial image perturbations remains not sufficiently characterized, particularly in the context of real-world threats such as compression, geometric distortions, and low-visibility tampering. Prior work has primarily focused on video-level watermarking, model ownership protection, or evaluations under benign conditions. There remains a need for systematic testing frameworks that assess both watermark survivability and perceptual fidelity under adversarial attack scenarios, particularly as AI image generation and alteration become more sophisticated.

We evaluate three watermarking approaches: two classical signal processing methods DWT-DCT and DWT-DCT-SVD, which embed watermarks using combinations of Discrete Wavelet Transform, Discrete Cosine Transform, and Singular Value Decomposition, and one deep learning–based approach RivaGAN using a threshold-based robustness testing pipeline. We simulate a wide range of attacks, including JPEG compression, cropping, brightness adjustment, masking, and geometric transformations, to determine the conditions under which watermark decoding fails. Additionally we introduce a perceptual similarity analysis using LPIPS to identify which attacks are most effective at evading detection while preserving visual quality.

Our findings offer new insights into the trade-offs between watermark imperceptibility and robustness, and support the development of tamper-resistant, perceptually optimized watermarking systems for use in cybersecurity contexts such as deepfake detection, digital media authentication, and misinformation defense.

II. Related Work

Research on digital watermarking spans both model-level and image-level techniques. Model watermarking aims to embed ownership claims within neural networks themselves. DeepSigns [8], for example, embeds watermarks in the activation maps of intermediate layers using a custom loss function. This approach is resilient to model pruning, fine-tuning, and compression, and maintains performance even in black-box settings. However, it is tailored to model ownership verification and does not address image-space manipulations, which are central to visual media authentication.

In contrast, image-level watermarking techniques embed information directly into pixel data. A recent hybrid approach

Identify applicable funding agency here. If none, delete this.

by Patil and Patil [4] uses a combination of Discrete Wavelet Transform (DWT), Discrete Cosine Transform (DCT), and Singular Value Decomposition (SVD), coupled with matrix encryption, to secure medical images. Their method shows robustness to attacks such as noise, JPEG compression, and geometric distortions. While their use case differs, our work builds on this classical transform stack—DWT-DCT-SVD—as a baseline for watermark robustness without encryption.

More recent deep learning–based image watermarking frameworks aim to improve imperceptibility and robustness through learned representations. Zhang et al. [2] present RivaGAN: an encoder-decoder architecture that embeds watermarks in images, enabling recovery even after cropping, compression, or adversarial noise. While their work evaluates video and model watermarking, we adapt this method for static image protection and expand evaluation to include perceptual similarity metrics like LPIPS.

Several papers also explore watermarking in the context of deepfake detection. Wang et al. [1] introduce LampMark, which embeds watermark signals based on facial landmarks. This method can survive deepfake generation and enables watermark-based traceability. However, their implementation is dataset-constrained and was not reproducible in our experiments due to limited access. Our work draws inspiration from their proactive detection paradigm and extends it by evaluating robustness under controlled adversarial attacks.

Wang et al. [9] propose an alternative proactive deepfake detection framework that integrates watermarking directly into a learning model. Their method embeds watermark information using Quaternion Polar Harmonic Fourier Moments (QPHFMs) and enhances detection performance through a Dual-Task Mutual Learning (DTML) framework. This joint approach enables the system to both recover embedded watermarks and improve deepfake detection accuracy through mutual supervision. While their focus is on integrating watermarking into the feature space of learned representations, our work emphasizes resilience in pixel-space watermarking methods under attack, and introduces perceptual modeling to assess invisibility and robustness trade-offs. Together, these works reflect a growing interest in hybrid frameworks that blend imperceptible watermarking with proactive cybersecurity applications.

Across these prior works, there remains a gap in evaluating how different watermarking techniques respond to adversarial image perturbations in a controlled, perceptually grounded setting. Our project addresses this need by comparing classical and deep learning–based watermarking models under systematic attack simulations, with an emphasis on cybersecurity applications such as deepfake detection and media integrity verification.

III. Methodology

A. Overview

To evaluate the robustness of invisible watermarking methods against adversarial image attacks, we implemented a testing pipeline comparing two classical signal processing techniques (DWT-DCT and DWT-DCT-SVD) and one deep learning–based model (RivaGAN). Each method embeds a unique binary watermark into input images, which are then subjected to a series of distortion-based transformations. Watermark recovery is attempted post attack to assess decoding success and robustness.

In our implementation, the classical DWT-based methods embed a 64-bit message, while RivaGAN embeds a 32-bit message, reflecting architectural and training differences between hand-crafted and learned methods. Decoding success is evaluated based on full message recovery, with no partial credit assigned. In addition to raw decoding accuracy, we incorporate perceptual similarity analysis using the LPIPS metric to measure how visually noticeable each attack is relative to its effectiveness in disabling watermark decoding.

B. Watermarking Methods

DWT-DCT

This method combines Discrete Wavelet Transform (DWT) and Discrete Cosine Transform (DCT) to embed watermark information into frequency sub-bands of an image. DWT decomposes the image into multi-resolution components, and DCT is applied to select sub-bands for watermark embedding. The modified image is then reconstructed using inverse transforms. This approach balances watermark imperceptibility and robustness, but its performance varies under geometric distortions and resolution changes.

DWT-DCT-SVD

An extension of the above method, this variant introduces Singular Value Decomposition (SVD) into the embedding process. After applying DWT and DCT, SVD is performed on selected coefficients, and the watermark is embedded into the singular values before inverse transformations are applied. SVD adds an additional layer of stability, improving robustness to intensity-based attacks such as JPEG compression and brightness adjustments.

RivaGAN

RivaGAN is a deep learning–based framework that embeds watermarks using a convolutional encoder-decoder architecture. The encoder network imperceptibly modifies the input image to embed a binary watermark, while the decoder attempts to extract the watermark from potentially distorted images. The model is trained to optimize reconstruction loss and optionally includes perceptual losses to improve visual fidelity. RivaGAN is expected to be more resilient to complex transformations due to its learned embedding strategy.

C. Dataset

We use a curated subset of 15 high-quality images from the Unsplash dataset, selected to ensure diversity in subject matter. The set includes 3 animals, 2 city scenes (day and night), 3 landscapes, 2 objects, and 5 portraits of people. All images are resized to 512×512 pixels to maintain consistency across models and to enable direct comparison between the computationally intensive RivaGAN method and the classical signal processing approaches. The classical methods are also evaluated on the original high-resolution versions of the images to examine how input quality affects watermark robustness.

D. Attack Simulation

To test robustness, we simulate nine types of adversarial image perturbations, each applied at incremental severity levels:

- Brightness adjustment (↑ / ↓)
- JPEG compression
- Gaussian noise
- Cropping
- Rotation
- Masking (with occlusion blocks)
- Overlay (with logos)
- Resizing (downscale + upscale)

For each attack, the watermarked image is distorted, then passed to the decoder for watermark extraction. The decoding is considered successful if the bitwise output matches the original embedded watermark.

E. Perceptual Similarity Evaluation

To complement binary decoding accuracy, we use the Learned Perceptual Image Patch Similarity (LPIPS) metric to assess how perceptually noticeable each attack is. LPIPS compares deep features extracted from pretrained CNNs (e.g., AlexNet) between original and attacked images. This allows us to identify attacks that are both visually imperceptible and watermark-destructive, which represent serious risks for content authentication systems.

F. Architecture Diagrams

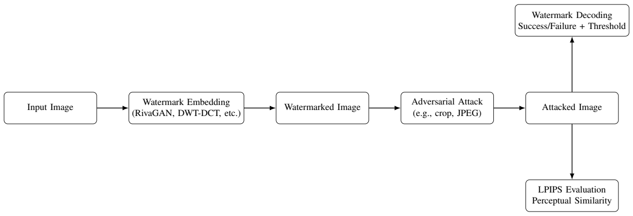
Input Image
Watermark Embedding (RivaGAN, DWT-DCT, etc.)
Watermarked Image
Adversarial Attack (e.g., crop, JPEG)
Attacked Image
Watermark Decoding Success/Failure + Threshold
LPIPS Evaluation Perceptual Similarity

Fig. 1. Watermarking robustness pipeline with parallel evaluation of decoding success and perceptual similarity (LPIPS).

IV. Experimental Results and Evaluation

We evaluated the robustness of three invisible watermarking techniques: DWT-DCT, DWT-DCT-SVD, and RivaGAN against a range of adversarial image perturbations. Experiments were conducted on a curated set of 15 512×512 images from the Unsplash dataset, with classical methods also tested on higher-resolution originals. Each image was embedded with a binary watermark and then subjected to one of nine transformation-based attacks applied incrementally to determine decoding failure thresholds.

Watermark recovery was recorded as a binary success/failure. In addition, we measured perceptual distortion using the LPIPS metric to evaluate whether attacks that successfully broke watermark decoding also introduced noticeable visual changes.

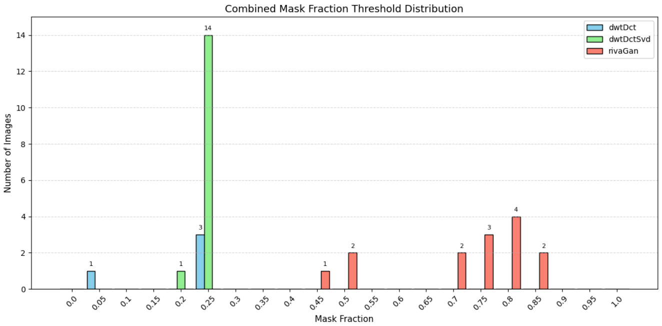

Fig. 2. Decode success rates for each method under increasing Mask. RivaGAN remains robust up to 85% of the image masked, while classical methods fail at lower mask thresholds.

A. Summary of Experimental Design

All methods were tested using the same 15-image subset, and each attack was applied at increasing severity levels (e.g., rotation 0°–20°, masking 0-100%). Table I summarizes the threshold ranges for each attack under which the embedded watermark remained decodable across the tested methods.

TABLE I
Robustness Summary Across Attacks for Each Watermarking Method

Attack Type	**DWT-DCT**	**DWT-DCT-SVD**	**RivaGAN**
JPEG Compression	Fails Instantly	60-70 Compression	30-100 Compression
Crop	Fails Instantly	Fails Instantly	0–30% Crop
Brightness ↑	Fails Instantly	0-20% Increase	0-300% Increase
Brightness ↓	0-20% Decrease	0-20% Decrease	20-60% Decrease
Rotation	0°	0°	2–16°
Resize	Fails Instantly	Scale 0.2-0.5	Scale 0.2-0.4
Mask	5-25%	20-25%	45-85%
Overlay	0-30% Opacity	0-30% Opacity	10-60% Opacity
Gaussian Noise	0-10 std dev	5-15 std dev	5-35 std dev

B. Decode Accuracy Under Attack

Figure 2 shows decoding success rates across attack threshold ranges. RivaGAN consistently outperformed classical methods, especially under crop, JPEG compression, and brightness reduction. Classical models often failed immediately under geometric transformations.

Notable results:

- RivaGAN survived up to 85% masking, down to 30 compressions, and moderate cropping
- DWT-DCT failed on most resized images, even without attack
- DWT-DCT-SVD performed more stably on JPEG and resize, but still weak to rotation

C. Perceptual Impact (LPIPS Scores)

We used the LPIPS metric to quantify the perceptual similarity between original and attacked images. This allowed us to evaluate which attacks could break watermark decoding while

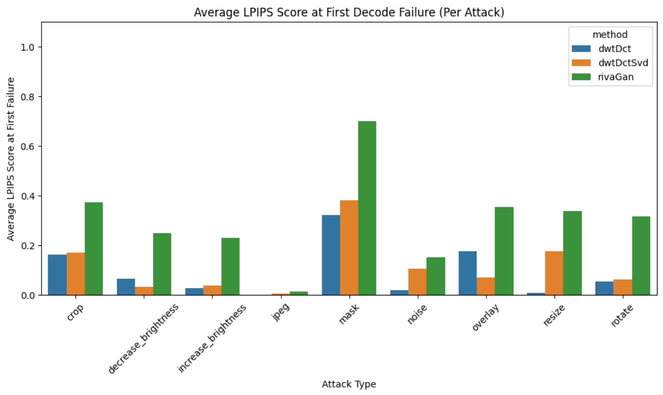

Fig. 3. Average LPIPS score at the first point of decoding failure per method. JPEG and Gaussian attacks degrade decoding at lower perceptual distortion levels than other attacks across all methods.

remaining visually subtle. Figure 3 summarizes LPIPS scores at the point of first decoding failure for each method.

JPEG compression was particularly effective at disabling watermark decoding while maintaining low LPIPS scores (typically between 0.05 and 0.15), indicating that these distortions were often imperceptible to human observers but highly disruptive to watermark integrity. In contrast, attacks such as overlay, masking, and cropping produced significantly higher LPIPS values, reflecting more noticeable visual degradation. Notably, failure cases in RivaGAN were often image-specific and did not consistently correspond to attack severity. This suggests that its decoder may be sensitive to content-dependent structural features rather than simple distortion magnitude.

D. Preliminary vs. Baseline Comparison

Baseline: In the baseline (unattacked) condition using 512×512 images, the DWT-DCT method successfully decoded the watermark in only 4 out of 15 cases, whereas DWT-DCT-SVD achieved a 100% success rate, and RivaGAN decoded 14 out of 15 images successfully. These results establish an upper bound for clean decoding performance and reveal that the classical DWT-DCT method struggles even without perturbations.

Improvements at higher resolution: When tested on higher-resolution originals, the classical methods showed notable improvement in decode performance, particularly under resizing and compression-based attacks. Interestingly, several images that failed to decode in the clean state were successfully decoded after undergoing mild transformations such as brightness adjustments or masking. This suggests the presence of potential alignment or feature-activation effects that warrant further exploration, especially for classical frequency-based methods.

E. What's Completed vs. Remaining

Completed:

- Threshold-based decode testing for 9 attacks
- LPIPS-based perceptual evaluation
- Pipeline automation and CSV + Markdown summaries
- Bar graphs and combined plots by method and image for threshold and LPIPS results

Work Remaining (For Internal Review)

Remaining Tasks and Questions:

- Add specific results for the tests on the original images for the DWT-DCT methods
- Find solution for GitHub space constraints — Google Drive?
- Get AI image resolution tool running
- Deepfake-specific transformation tests (e.g., GAN face swaps)
- Consider how to include Stable Diffusion and/or transformers — ask Dr. Duan and Dr. Lee

V. Conclusion

This research highlights the limitations of current invisible watermarking techniques when faced with adversarial image perturbations. While deep learning–based approaches like RivaGAN showed greater resilience than classical methods such as DWT-DCT and DWT-DCT-SVD, all methods exhibited vulnerabilities to specific transformations—particularly geometric distortions and perceptually subtle attacks like JPEG compression. Notably, some attacks were able to disable watermark decoding while maintaining low LPIPS scores, revealing a critical trade-off between visual imperceptibility and semantic integrity.

These findings underscore the need for continued development of robust, adaptive watermarking systems that can withstand real-world manipulations while remaining invisible to the human eye. Our testing pipeline provides a reusable and extensible framework for evaluating watermark robustness under both traditional attacks and perceptually grounded metrics, making it a valuable tool for future benchmarking and analysis.

Looking ahead, future work will explore deepfake-specific transformations, hybrid attack scenarios, and the integration of modern generative models such as transformers and diffusion-based architectures. As generative AI technologies continue to accelerate, the ability to proactively embed tamper-resistant signals into digital media will become increasingly important for real-world deployment in cybersecurity, media authentication, and misinformation defense. This work contributes toward that goal by helping to define the current limitations—and the critical paths forward—for invisible watermarking as a proactive defense against deepfakes.

References

[1] T. Wang, S. Yang, and Y. Wang, “LampMark: Proactive Deepfake Detection via Training-Free Landmark Perceptual Watermarks,” in *Proc. ACM Int. Conf. Multimedia (MM)*, 2024. [Online]. Available: https://dl.acm.org/doi/10.1145/3664647.3680869

[2] H. Zhang, Y. Liu, C. Yu, J. Chen, and Y. Chen, “Robust Invisible Video Watermarking with Attention,” *arXiv preprint arXiv:1909.01285*, 2019. [Online]. Available: https://arxiv.org/pdf/1909.01285

[3] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive Growing of GANs for Improved Quality, Stability, and Variation,” *arXiv preprint arXiv:1710.10196*, 2018. [Online]. Available: https://arxiv.org/abs/1710.10196

[4] K. T. Patil and S. A. Patil, “Robust and Secure Watermarking Scheme Based on DWT-DCT-SVD with Matrix Encryption for Medical Images,” *J. King Saud Univ. Comput. Inf. Sci.*, Elsevier, 2023. [Online]. Available: https://doi.org/10.1016/j.jksuci.2022.10.020

[5] ShieldMnt Team, “Invisible Watermark GitHub Repository,” GitHub, 2025. [Online]. Available: https://github.com/ShieldMnt/invisible-watermark

[6] T. Wang, “LampMark GitHub Repository,” GitHub, 2025. [Online]. Available: https://github.com/wangty1/LampMark/tree/main/image_data

[7] T. Karras *et al.*, “Progressive Growing of GANs (TensorFlow Implementation),” GitHub, 2025. [Online]. Available: https://github.com/tkarras/progressive_growing_of_gans

[8] B. D. Rouhani, H. Chen, and F. Koushanfar, “DeepSigns: An End-to-End Watermarking Framework for Ownership Protection of Deep Neural Networks,” in *Proc. ACM Int. Conf. Multimedia (MM)*, 2019. [Online]. Available: https://dl.acm.org/doi/10.1145/3297858.3304051

[9] C. Wang, C. Shi, S. Wang, Z. Xia, and B. Ma, “Dual-Task Mutual Learning With QPHFM Watermarking for Deepfake Detection,” *IEEE Signal Process. Lett.*, vol. 31, pp. 2740–2744, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10623297