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. Problem Statement:

This project aims to examine the resilience and robustness of invisible image watermarks to adversarial manipulation, with the goal of enhancing media integrity verification in cybersecurity contexts. In particular, it supports proactive deepfake detection—an emerging privacy and misinformation threat—by identifying how image transformations degrade watermark integrity. It also informs design decisions for models that balance watermark imperceptibility with attack resistance, aiding in the development of watermark-based tamper detection tools.

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

III. Background and Related Work

A. Image Watermarking Mechanisms:

Digital image watermarking is the process of embedding imperceptible information into an image for purposes such as ownership verification, tamper detection, or content tracking. Techniques discussed in this paper fall into two categories: classical frequency-domain approaches and deep learning–based methods.

Classical approaches often operate in the frequency domain, using transformations like the Discrete Cosine Transform (DCT), Discrete Wavelet Transform (DWT), or Singular Value Decomposition (SVD). These methods manipulate transform coefficients to embed watermarks in regions less sensitive to human vision, allowing the watermark to remain robust under basic transformations such as compression or low-level noise. For example, the DWT-DCT-SVD combination has been used in medical image protection for its balance between invisibility and resilience to attacks like JPEG compression and geometric distortion [4].

In contrast, deep learning–based watermarking methods use encoder-decoder architectures to learn where and how to embed watermarks. One such method is RivaGAN [2], which learns to hide a watermark within an image and recover it after it has undergone transformations such as cropping or adversarial noise. These models aim to maximize robustness and invisibility by optimizing over large datasets and using perceptual loss functions. Another recent deep learning strategy includes embedding watermark signals in feature space representations, such as through facial landmarks in LampMark [1], particularly for applications like deepfake detection and provenance tracking.

B. Attacks on Image Watermarking:

Watermarking systems are vulnerable to a variety of attacks that seek to remove, distort, or make the watermark undetectable. These include:

- Noise addition (e.g., Gaussian noise),
- Compression artifacts (e.g., JPEG),
- Geometric distortions (e.g., cropping, rotation, resizing),
- Overlay or masking attacks, and
- Advanced adversarial attacks, including denoising or generative model–based reconstruction.

While many classical watermarking techniques can withstand simple perturbations, more advanced attacks are specifically designed to break robustness by manipulating image content while preserving perceptual quality.

To evaluate these challenges, our study simulates a wide variety of such attacks in a controlled setting. Importantly, we incorporate Learned Perceptual Image Patch Similarity (LPIPS): a deep feature–based similarity metric to evaluate how "invisible" or perceptually severe an attack is [10]. The LPIPS framework calculates differences between the original and attacked images across several layers, computes normalized squared differences, and weighs them using learned perceptual weights. Lower LPIPS scores indicate higher perceptual similarity. This allows us to quantify when decoding failures occur relative to perceptual changes, giving insight into the trade-off between watermark robustness and visual imperceptibility.

Together, this framework helps bridge a gap in existing literature: systematically evaluating how both classical and learned watermarking techniques hold up under diverse, perceptually grounded attacks.

C. Related Work:

Several recent works have applied these watermarking principles in practical systems. DeepSigns [8] embeds watermarks in the activation maps of deep neural networks to protect model ownership against pruning and fine-tuning. LampMark [1] proposes the embedding facial landmark–based watermarks to survive deepfake generation, although the implementation remains dataset dependent. RivaGAN [2], which we adapt in our own work, uses a GAN-based encoder-decoder to embed watermarks in images, allowing recovery under compression and adversarial distortion. Other hybrid approaches such as QPHFM with Dual-Task Mutual Learning (DTML)[9] integrate watermarking into learned features to jointly improve robustness and deepfake detection. These methods demonstrate increasing interest in combining imperceptibility, resilience, and proactive security in watermarking systems. Our work complements this direction by systematically comparing classical and deep learning–based image watermarking techniques under a diverse set of attacks, while also incorporating perceptual similarity (LPIPS) to ground our evaluation in human-visible distortion.

IV. 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.

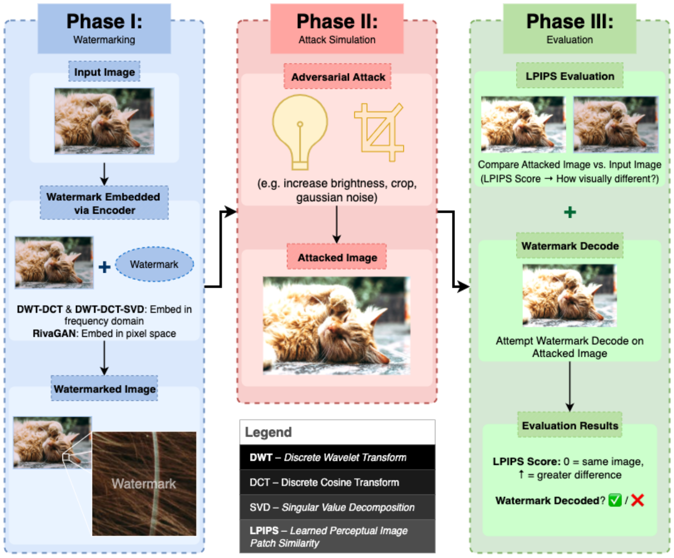

Fig. 1. Watermarking evaluation pipeline. Phase I embeds an invisible watermark using classical or deep learning methods. Phase II applies adversarial perturbations. Phase III evaluates robustness using LPIPS (perceptual similarity) and decoding success.

B. Watermarking Methods for Evaluation

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 the robustness of the watermark.

Fig. 2. Sample Images from 15-Image Unsplash Subset

D. Attack Simulation

We apply a suite of adversarial and perceptual attacks to evaluate watermark robustness. These include traditional transformations such as Gaussian noise (which adds random pixel-level variation), JPEG compression (which removes high-frequency detail), brightness adjustment, and rotation. We also test crop, resize, and mask operations, which modify image structure or remove key regions. Additionally, we include overlay attacks (which add distracting content) and perceptual transformations like denoising and rescaling using modern image restoration models. These attacks represent a broad spectrum of real-world image perturbations: from unintentional degradation to targeted tampering, and allow us to assess each watermarking method's resilience in diverse threat scenarios.

Our 9 adversarial image perturbations are 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.

V. 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.

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.

B. Decode Accuracy Under Attack

Figure 3 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.

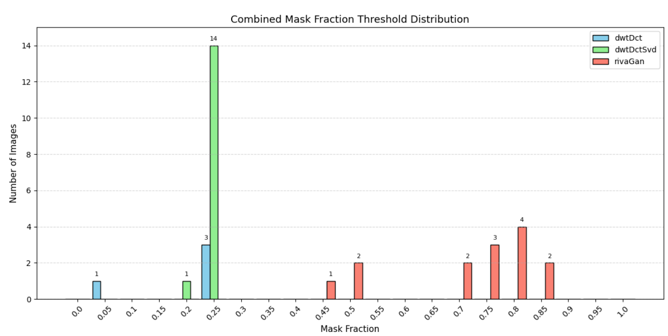

Fig. 3. Decode success rates for each method under increasing Mask. RivaGAN remains robust up to 85% of the image masked, while classical methods fail at lower mask thresholds.

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 remaining visually subtle. Figure 4 summarizes LPIPS scores at the point of first decoding failure for each method.

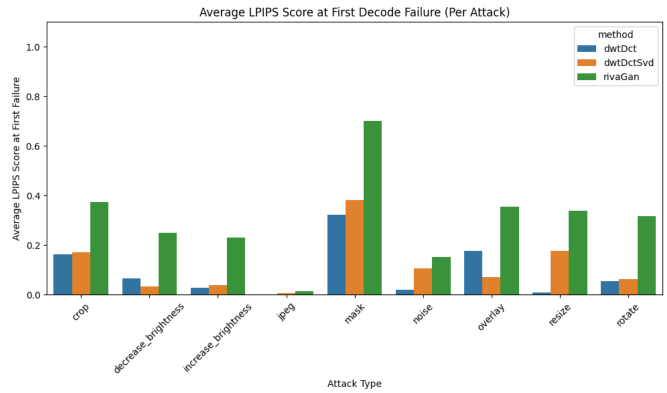

Fig. 4. Average LPIPS score at the first point of decoding failure per method. JPEG and Gaussian attacks degrade decoding at lower perceptual distortion levels than other attacks across all methods.

JPEG compression was particularly effective at disabling watermark decoding while maintaining low LPIPS scores (typically between 0.005 and 0.015), 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.

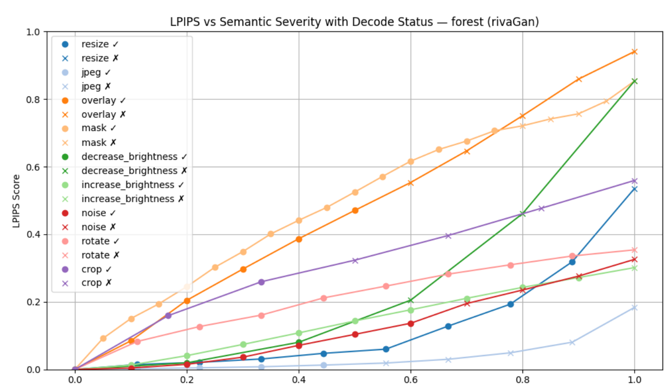

Fig. 5. LPIPS vs. decode status for RivaGAN on a forest image. Despite high perceptual distortion (LPIPS ≈ 0.7), decoding succeeded under overlay but failed under JPEG with LPIPS < 0.01.

D. Preliminary vs. Baseline Comparison

Baseline: In the baseline (unattacked) condition using 512×512 images, the DWT-DCT method successfully decoded

TABLE I

AVERAGE THRESHOLDS AND LPIPS SCORES (± STD) AT FIRST DECODE FAILURE

Attack Type	DWT-DCT	DWT-DCT-SVD	RivaGAN
crop	Threshold: 0.9 $\pm$ 0.0 LPIPS: 0.164 $\pm$ 0.031	Threshold: 0.9 $\pm$ 0.0 LPIPS: 0.174 $\pm$ 0.039	Threshold: 0.729 $\pm$ 0.103 LPIPS: 0.373 $\pm$ 0.135
$\downarrow$ brightness	Threshold: 0.7 $\pm$ 0.1 LPIPS: 0.066 $\pm$ 0.023	Threshold: 0.787 $\pm$ 0.05 LPIPS: 0.033 $\pm$ 0.022	Threshold: 0.386 $\pm$ 0.16 LPIPS: 0.249 $\pm$ 0.124
$\uparrow$ brightness	Threshold: 1.2 $\pm$ 0.0 LPIPS: 0.028 $\pm$ 0.012	Threshold: 1.24 $\pm$ 0.08 LPIPS: 0.038 $\pm$ 0.025	Threshold: 1.9 $\pm$ 0.539 LPIPS: 0.23 $\pm$ 0.135
jpeg	Threshold: 100.0 $\pm$ 0.0 LPIPS: 0.001 $\pm$ 0.0	Threshold: 65.333 $\pm$ 18.571 LPIPS: 0.008 $\pm$ 0.007	Threshold: 51.429 $\pm$ 19.588 LPIPS: 0.013 $\pm$ 0.009
mask	Threshold: 0.25 $\pm$ 0.087 LPIPS: 0.324 $\pm$ 0.105	Threshold: 0.297 $\pm$ 0.012 LPIPS: 0.383 $\pm$ 0.067	Threshold: 0.657 $\pm$ 0.234 LPIPS: 0.7 $\pm$ 0.215
noise	Threshold: 7.5 $\pm$ 4.33 LPIPS: 0.021 $\pm$ 0.021	Threshold: 17.667 $\pm$ 3.091 LPIPS: 0.107 $\pm$ 0.073	Threshold: 24.286 $\pm$ 7.986 LPIPS: 0.152 $\pm$ 0.077
overlay	Threshold: 0.25 $\pm$ 0.15 LPIPS: 0.177 $\pm$ 0.146	Threshold: 0.14 $\pm$ 0.08 LPIPS: 0.071 $\pm$ 0.059	Threshold: 0.529 $\pm$ 0.144 LPIPS: 0.355 $\pm$ 0.159
resize	Threshold: 0.92 $\pm$ 0.04 LPIPS: 0.007 $\pm$ 0.005	Threshold: 0.28 $\pm$ 0.122 LPIPS: 0.176 $\pm$ 0.128	Threshold: 0.157 $\pm$ 0.082 LPIPS: 0.338 $\pm$ 0.153
rotate	Threshold: 2.0 $\pm$ 0.0 LPIPS: 0.056 $\pm$ 0.019	Threshold: 2.0 $\pm$ 0.0 LPIPS: 0.064 $\pm$ 0.021	Threshold: 13.286 $\pm$ 4.25 LPIPS: 0.318 $\pm$ 0.093

Threshold Units by Attack Type. Crop: % of image remaining; ↓ Brightness: % decrease in brightness; ↑ Brightness: % increase in brightness; JPEG: JPEG quality level (0–100); Mask: % of image masked; Noise: Gaussian noise std. dev; Overlay: % opacity; Resize: Scale factor (1.0 = original); Rotate: Rotation angle in degrees.

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.

VI. Conclusion and Future Work

A. 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.

B. Future Work

Building on this foundation, several directions remain for exploration:

- **Perceptual Analysis:** Compute and visualize LPIPS scores for denoising and rescaling attacks to complete the perceptual robustness analysis.
- **Model Expansion:** Evaluate OpenStego (a traditional watermarking tool) and test the WatermarkAttacker framework, which targets the embedding space using adversarial noise and regeneration.
- **Framework Generalization:** Design a summary table or visualization framework to accommodate binary (non-threshold) attacks like denoising and rescaling.
- **Advanced Threats:** Explore deepfake-oriented attack scenarios, hybrid transformation pipelines, and attacks generated by diffusion or transformer models.
- **AI Explainability:** Investigate whether explainable AI techniques can help identify when a watermarked image has been tampered with.
- **Multi-model Embedding:** Consider integrating multiple watermarking models to improve robustness or recoverability in the face of unknown transformations.

This work offers a reproducible benchmark for evaluating invisible watermarking systems under adversarial pressure, while highlighting the need for perceptually grounded evaluation metrics and broader threat modeling to protect digital media integrity.

C. What's Completed vs. Remaining

NOTE: This section is for REU mentors only; remove for publication version

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

WEEK 7:

- Evaluated two new robustness scenarios: Denoising attack using NAFNet and Image rescaling using ISR (idealo/image-super-resolution)
- Drafted poster and updated project diagrams
- Created summary table of first decode failure LPIPS scores and threshold values by method
- Refined Related Work section with clearer background on watermarking mechanisms and attacks (suggested by dr. duan)

Remaining Tasks and Questions:

- **Perceptual Evaluation:** Compute LPIPS scores for denoising and rescaling results and visualize decode outcomes
- **Result Integration:** Develop comparative table or visualization to include binary (non-threshold) attacks alongside threshold-based results
- **Model Expansion:**
- Evaluate OpenStego — a traditional watermarking tool
- Test advanced attack from WatermarkAttacker (embedding-based noise + image regeneration)
- **Advanced/Exploratory Ideas:**
- Explore hybrid attacks and combined transformations
- Investigate robustness against deepfake generation models (e.g., GAN face swaps)
- Consider integration of Stable Diffusion or Transformer-based architectures for attack or watermarking
- Add AI explainability tools to detect whether a watermarked image has been tampered with
- Investigate multi-model embedding strategies
- **Logistics + Reproducibility:**
- Add final visualizations and summaries to REU GitHub
- Resolve GitHub size limits (use Google Drive)
- Document test pipeline and upload relevant code from invisible-watermark-cat repo to REU repo

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

[10] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018. [Online]. Available: https://github.com/idealo/image-super-resolution

[11] F. Cardinale *et al.*, "ISR: Image Super-Resolution," 2018. [Online]. Available: https://github.com/idealo/image-super-resolution