Watermark Robustness under Adversarial Attacks for Deepfake Detection

Cat Lewin

School of Science and Engineering
University of Missouri-Kansas City
Kansas City, USA

Introduction

As deepfakes and digital media tampering becomes increasingly sophisticated, invisible watermarking has emerged as a potential tool for ensuring image authenticity and traceability. 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.

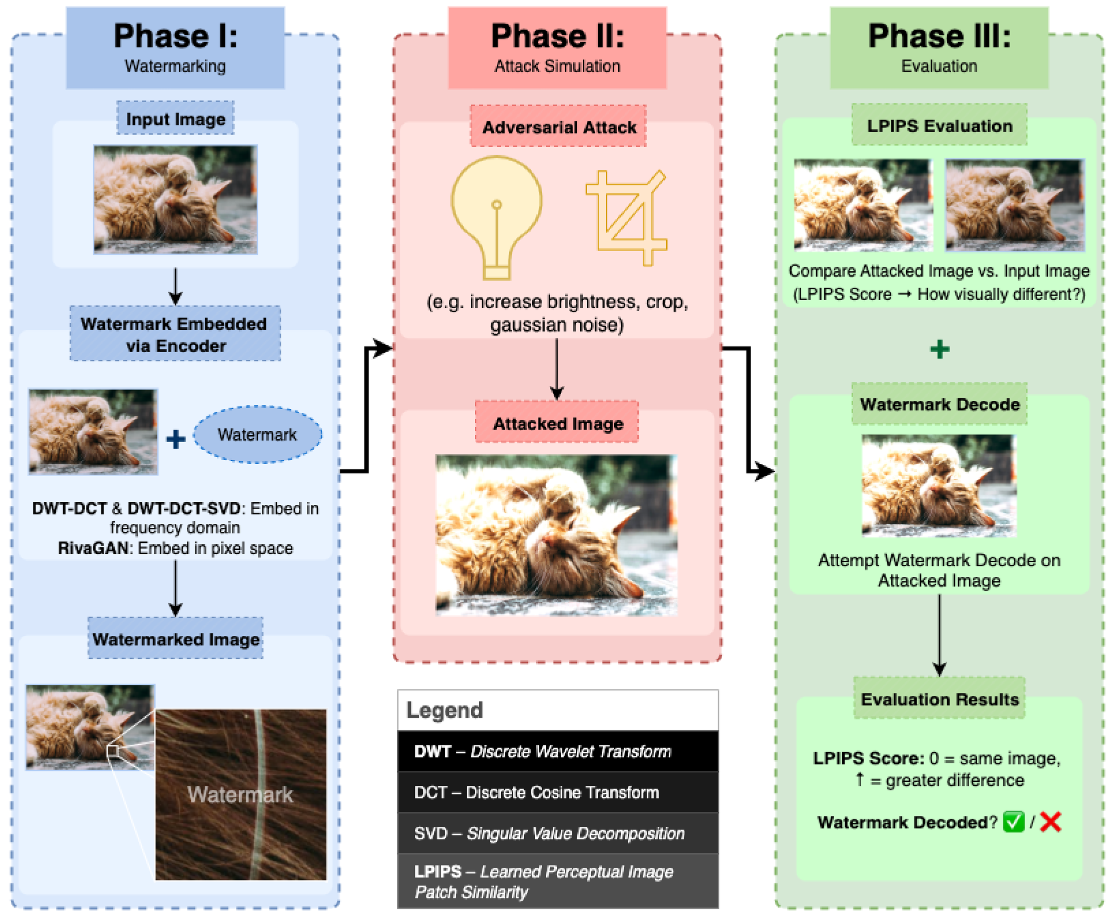

Figure 1:Watermarking evaluation pipeline. Phase I embeds an invisible watermark using classical or deep learning methods. Phase II applies adversarial perturbations. Phase III evaluates robustness using LPIPS (perceptual similarity) and decoding success.

Dataset

We use 15 diverse images from Unsplash, including animals, cityscapes, landscapes, objects, and portraits. All images are resized to 512×512 for consistency and fair comparison across methods. Classical watermarking approaches are also tested on the original high-resolution images to assess the impact of input quality on robustness.

Figure 2:Sample Images from 15-Image Unsplash Subset.

Avg Thresholds and LPIPS Scores at First Decode Failure

Attack Type	DWT-DCT	DWT-DCT-SVD	RivaGAN
crop	Threshold: 0.9 ± 0.0 LPIPS: 0.164 ± 0.031	Threshold: 0.9 ± 0.0 LPIPS: 0.174 ± 0.039	Threshold: 0.729 ± 0.103 LPIPS: 0.373 ± 0.135
↓ brightness	Threshold: 0.7 ± 0.1 LPIPS: 0.066 ± 0.023	Threshold: 0.787 ± 0.05 LPIPS: 0.033 ± 0.022	Threshold: 0.386 ± 0.16 LPIPS: 0.249 ± 0.124
↑ brightness	Threshold: 1.2 ± 0.0 LPIPS: 0.028 ± 0.012	Threshold: 1.24 ± 0.08 LPIPS: 0.038 ± 0.025	Threshold: 1.9 ± 0.539 LPIPS: 0.23 ± 0.135
jpeg	Threshold: 100.0 ± 0.0 LPIPS: 0.001 ± 0.0	Threshold: 65.333 ± 18.571 LPIPS: 0.008 ± 0.007	Threshold: 51.429 ± 19.588 LPIPS: 0.013 ± 0.009
mask	Threshold: 0.25 ± 0.087 LPIPS: 0.324 ± 0.105	Threshold: 0.297 ± 0.012 LPIPS: 0.383 ± 0.067	Threshold: 0.657 ± 0.234 LPIPS: 0.7 ± 0.215
noise	Threshold: 7.5 ± 4.33 LPIPS: 0.021 ± 0.021	Threshold: 17.667 ± 3.091 LPIPS: 0.107 ± 0.073	Threshold: 24.286 ± 7.986 LPIPS: 0.152 ± 0.077
overlay	Threshold: 0.25 ± 0.15 LPIPS: 0.177 ± 0.146	Threshold: 0.14 ± 0.08 LPIPS: 0.071 ± 0.059	Threshold: 0.529 ± 0.144 LPIPS: 0.355 ± 0.159
resize	Threshold: 0.92 ± 0.04 LPIPS: 0.007 ± 0.005	Threshold: 0.28 ± 0.122 LPIPS: 0.176 ± 0.128	Threshold: 0.157 ± 0.082 LPIPS: 0.338 ± 0.153
rotate	Threshold: 2.0 ± 0.0 LPIPS: 0.056 ± 0.019	Threshold: 2.0 ± 0.0 LPIPS: 0.064 ± 0.021	Threshold: 13.286 ± 4.25 LPIPS: 0.318 ± 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 std. dev
Overlay:	% opacity
Resize:	Scale factor (1.0 = original)
Rotate:	Degrees rotated

Experimental Methodology

We evaluate the robustness of three invisible watermarking methods—DWT-DCT, DWT-DCT-SVD, and RivaGAN—by embedding binary watermarks into 15 diverse images and subjecting them to a series of adversarial image attacks. Classical methods embed in frequency space and use a 64-bit (8-character) watermark, while RivaGAN embeds in pixel space with a 32-bit (4-character) watermark due to model constraints. Each watermarked image undergoes 9 attacks (e.g., JPEG compression, brightness change, crop, rotation) applied in increasing severity until the watermark fails to decode. This threshold-based testing reveals the point at which robustness breaks down for each method. Decoding success is determined by full watermark recovery. We also compute LPIPS perceptual similarity to assess how visually noticeable each attack is relative to its impact on decoding.

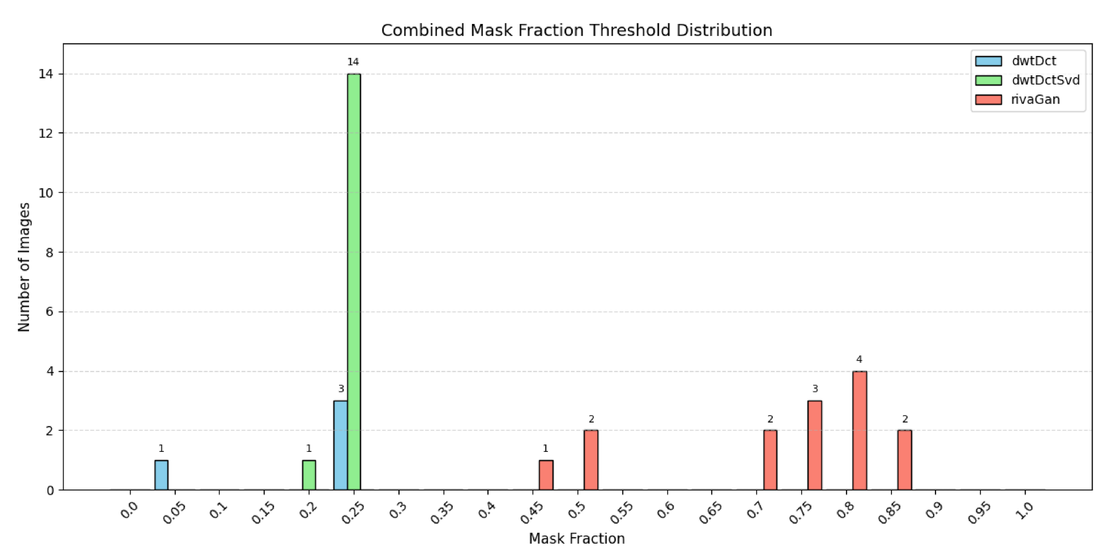

Figure 3:Decode success rates for each method under increasing levels of image masking. DWT-DCT-SVD remains robust up to 25% masked. RivaGAN varies across images, with thresholds between 45%-85%. DWT-DCT performs worst, often failing with just 5%–25% masked.

Results

Decode Success:

- **RivaGAN** consistently outperformed classical methods, surviving on average 66% masking, JPEG compression level 51, and 73% crop. 14/15 images successfully decoded clean.
- **DWT-DCT** frequently failed even without attack – only 4/15 clean images succeeded decoding.
- **DWT-DCT-SVD** was more stable under JPEG and resize, but still vulnerable to geometric attacks. All images decoded clean.

Perceptual Impact (LPIPS):

- JPEG and Gaussian noise caused decode failure at low LPIPS scores (
- Overlay, mask, and crop attacks had higher LPIPS.

See table above and figures for decoding thresholds and LPIPS comparisons.

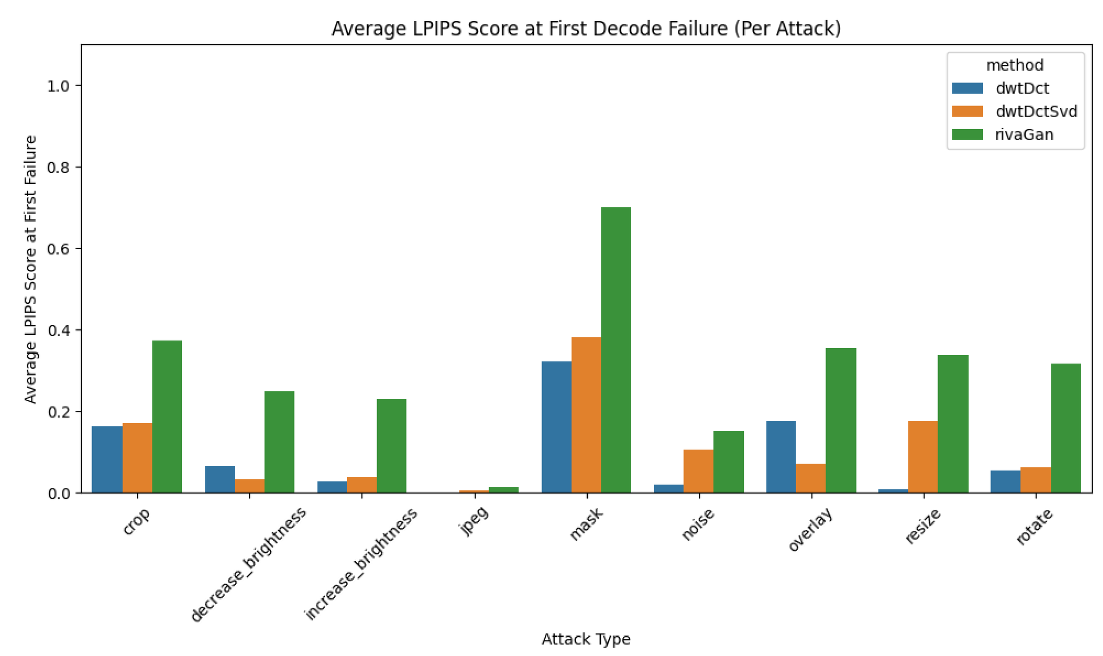

Figure 4:Average LPIPS score at the first point of decoding failure per method.

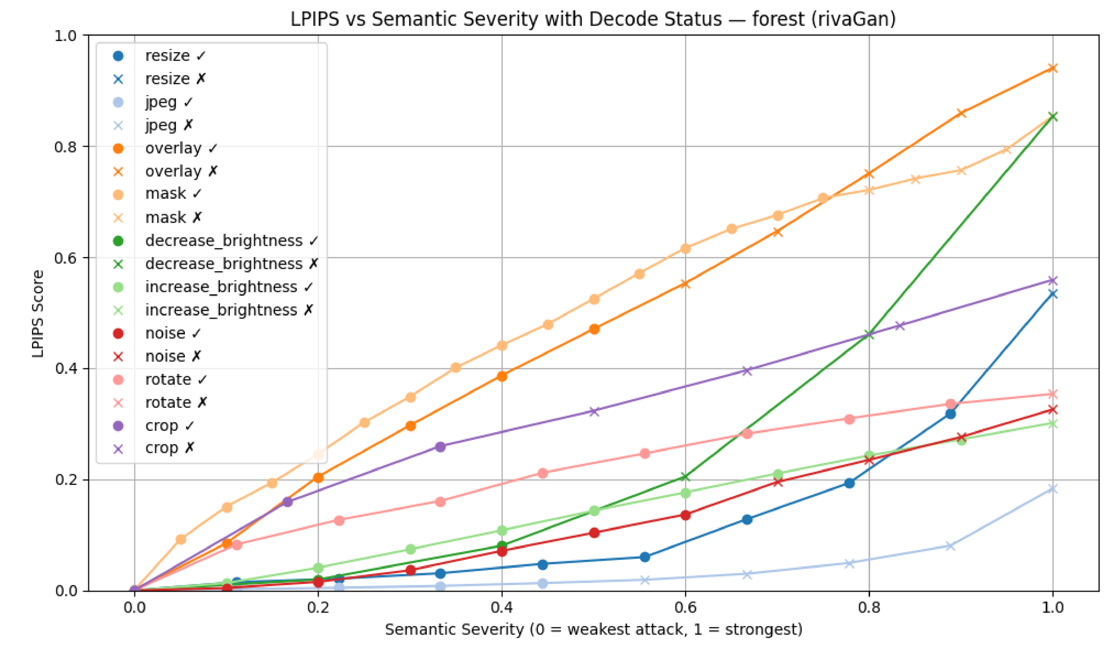

Figure 5:LPIPS vs. decode status for RivaGAN on a forest image. Despite high perceptual distortion (LPIPS ≈ 0.7), decoding succeeded under overlay but failed under JPEG with LPIPS < 0.01.

Conclusion & Future Work

Key Takeaways:

- RivaGAN is more robust than classical methods but still vulnerable to subtle or geometric attacks.
- Some attacks (e.g., JPEG) break decoding while remaining visually imperceptible.
- Decode success is sometimes content-dependent, especially in deep learning methods.

Future Work:

- Explore deepfake-specific and hybrid attacks.
- Integrate generative models (e.g., diffusion, transformers).
- Investigate explainable AI and multi-model embedding strategies.

GitHub: `https://github.com/catlewin/invisible-watermark-cat`

References

[1] H. Zhang *et al.*, "Robust Invisible Video Watermarking with Attention," *arXiv:1909.01285*, 2019.

[2] ShieldMnt Team, "Invisible Watermark GitHub Repository," GitHub, 2025. `https://github.com/ShieldMnt/invisible-watermark`

[3] R. Zhang *et al.*, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018. `https://github.com/idealo/image-super-resolution`

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant Number CNS-2349236. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.