

Week 4 Report

Note: Moving forward, my research focuses solely on testing the resilience of watermarking. Thus all work and literature from previous weeks relating to YouTube's copyright detection system has been removed.

Project Information

Project Type: Individual

Student Name: Cat Lewin

Mentor Name: Dr. Rui Duan

Research Title: Evaluating the Robustness of Invisible Watermarking Against Adversarial Attacks in Deepfake Detection

Problem Statement: Watermark Robustness in Deepfake Detection

This project aims to examine the resilience and robustness of invisible image watermarks to adversarial manipulation, and their reliability in support of proactive deepfake detection.

Hypotheses:

1. Adversarial attacks will degrade watermark detection accuracy, particularly when designed to preserve perceptual fidelity.
2. Hybrid watermarking techniques (e.g., DWT-DCT-SVD) will outperform simpler methods under distortion attacks.

Research Questions

- How robust are perceptual watermarks (e.g., Invisible Watermark) under adversarial image perturbations?
- What trade-offs exist between watermark imperceptibility and robustness?
- Can certain transformations be used to reliably weaken or remove embedded watermarks across different watermarking models?

Literature Review

Title: *Robust and Secure Watermarking Scheme Based on DWT-DCT-SVD with Matrix Encryption for Medical Images*

Authors: K. T. Patil and S. A. Patil

Publication Year: 2023

Summary: This paper presents a robust and secure digital image watermarking scheme tailored for medical imaging, utilizing a hybrid of Discrete Wavelet Transform (DWT), Discrete Cosine Transform (DCT), and Singular Value Decomposition (SVD).

The novelty lies in its integration of matrix-based encryption with the watermarking process to enhance security and prevent unauthorized extraction. The watermark is embedded in the singular values of the transformed image and encrypted using a key matrix to provide both robustness and confidentiality.

The approach is evaluated under several attacks—including Gaussian noise, salt & pepper noise, JPEG compression, rotation, and resizing—and demonstrates high watermark recovery accuracy with minimal perceptual degradation. Performance is assessed using Peak Signal-to-Noise Ratio (PSNR) and Normalized Correlation (NC), and results show consistent robustness even under intense distortions.

Connection to My Project: This paper directly supports the use of DWT-DCT-SVD in my implementation. While their application is in the medical domain, the core method is the same as in my evaluation framework. Their integration of encryption is out of scope for my current study, but the robustness benchmarks they use (e.g., JPEG compression, resizing, noise) align well with my attack simulations.

This paper reinforces the viability of DWT-DCT-SVD as a robust classical baseline to compare against newer deep learning-based watermarking methods like RivaGAN and Invisible Watermark. My project builds on this work by applying the same technique to face datasets and explicitly comparing robustness under both benign and adversarial image transformations.

Methods, Datasets, and Benchmarks:

Watermarking Method:

- DWT → DCT → SVD → Matrix Encryption → Watermark embedding

Evaluation Metrics:

- PSNR (Peak signal-to-noise ratio), NC (Normalized Correlation), BER (bit error rate)

Attack Benchmarks:

- JPEG Compression, Gaussian Noise, Rotation, Resizing, Salt & Pepper Noise

Planned Comparison in My Project:

- Similar attack types and metrics used to evaluate watermark recovery on CelebA
- No encryption layer in my current setup, but core transform stack (DWT-DCT-SVD) is the same

Title: *Invisible Watermarking of Deep Neural Networks for Intellectual Property Protection*

Authors: Huili Zhang, Yujie Liu, Chenhao Yu, Jinyuan Chen, and Yingying Chen

Publication Year: 2019

Summary: This paper proposes an end-to-end invisible watermarking framework designed to protect the intellectual property of deep neural networks. Unlike traditional

watermarking approaches that only embed binary patterns, this method hides a watermark (typically the owner's ID) within a neural network's parameters using a convolutional encoder-decoder architecture. The embedded watermark does not interfere with the model's functionality or accuracy and can be reliably decoded even after common model modifications such as fine-tuning or pruning. The watermark is encoded into images via an encoder network, then decoded after potential transformation by a decoder network. The authors evaluate robustness under a variety of attacks, including JPEG compression, noise addition, cropping, resizing, and adversarial attacks.

Connection to My Project: This framework forms the baseline of my implementation. I adapted its encoder-decoder architecture to watermark images in a perceptual way and test recovery after transformations — not in model weights, but in pixel space. Although the original focus is protecting models, their watermark embedding methodology is highly applicable to image-based perceptual watermarking as a proactive defense in deepfake detection. My experiments will extend their ideas by testing visibility (SSIM), recovery accuracy, and resilience under targeted adversarial image perturbations.

Methods, Datasets, and Benchmarks:

Methods Used:

- CNN-based encoder and decoder architecture
- Losses: Mean squared error for reconstruction, optional perceptual loss
- Training process simulates transformations to build robustness

Datasets Used:

- CIFAR-10
- MNIST
- ImageNet (subset)

Comparison in My Project:

- I reproduce and extend their image watermarking methodology. While their paper evaluates video watermark recovery, I evaluate image-level watermark recovery post-attack. I use similar robustness benchmarks: cropping, noise, blur, and compression, enabling direct comparison of recovery under attack conditions and will expand it to test watermark recovery against deepfake transformations.

Title: *LampMark: Proactive Deepfake Detection via Training-Free Landmark Perceptual Watermarks*

Authors: Tianyi Wang, Shaofei Yang, and Yanzhi Wang

Publication Year: 2024

Summary: LampMark proposes a proactive approach to deepfake detection that embeds invisible watermarks into images based on facial landmarks, allowing the

system to trace deepfakes even after manipulation. Unlike passive detection systems—which analyze synthetic artifacts in already-manipulated images—LampMark embeds robust, traceable signals that survive deepfake generation processes. The watermark is generated by projecting facial landmarks into a binary space and then encrypting and embedding it into the image using a trained end-to-end Convolutional Neural Networks (CNN) framework.

Contributions:

- Introduces a landmark-perceptual watermarking system capable of resisting (many) deepfake transformations.
- Establishes that facial landmark structures are measurably and consistently altered by deepfake operations but not by benign manipulations.
- Provides an efficient method to trace the source of manipulated images by comparing recovered watermarks against expected landmark-based encodings.
- Utilizes CNNs to embed, encrypt, and recover watermarks in a manner that ensures discrimination (detectability), confidentiality (privacy), and robustness (resilience to attacks).

Limitations:

- The distinction between semi-fragile and robust watermarks is implied but not clearly defined or quantified in terms of resistance to different manipulation types.
- The framework is trained on the CelebA-HQ and Labelled Faces in the Wild (LFW) datasets, which may not generalize across all face distributions or deepfake generators.
- Due to limited access to the original training datasets, I was unable to reproduce or evaluate the reported performance of the system.

Expand/Improve:

- Test the framework on newer or more diverse datasets and with more varied generative models.
- Explore how landmark-based watermarking performs under adversarial attacks or intentional watermark removal strategies.

Title: *Progressive Growing of GANs for Improved Quality, Stability, and Variation*

Authors: Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen

Publication Year: 2018

Summary: This paper introduces a training strategy for generative adversarial networks (GANs) called progressive growing, where both the generator and discriminator are trained to produce increasingly higher-resolution images over time. The method starts with low-resolution outputs and gradually adds layers to reach full-resolution generation. This approach significantly stabilizes GAN training and improves both the quality and variation of generated images, particularly on high-resolution face datasets like CelebA-HQ.

Contributions:

- Proposes progressive layer expansion for GANs to improve stability and avoid mode collapse.
- Enables high-resolution (e.g., 1024×1024) image generation with much higher visual fidelity than prior GAN models.
- Publicly released CelebA-HQ, a high-quality aligned face dataset created using the progressive training pipeline from the CelebA dataset.

Limitations:

- Although the progressive strategy improves stability, training remains computationally intensive, requiring high-end GPUs and significant memory.
- Training from scratch remains slow despite improved convergence behavior.
- In my experience, either due to the computer I was using or the provided source data, the GitHub program consistently created corrupted files.

Expand/Improve:

- This paper serves as the foundational process for CelebA-HQ generation. In my work, I (attempted to) use their TensorFlow implementation to recreate the dataset from original CelebA images as part of watermark embedding experiments.

Proposed Approach

AI Model(s) used: Invisible Watermark Framework ([Zhang et al., 2019]):

- Uses Convolutional Neural Networks (CNNs) in an end-to-end architecture:
 - Encoder Network: embeds the watermark into the image.
 - Decoder Network: recovers the watermark from potentially manipulated images.
- The watermark embedding and recovery is trained using:
 - Reconstruction loss
 - Perceptual loss (optionally)

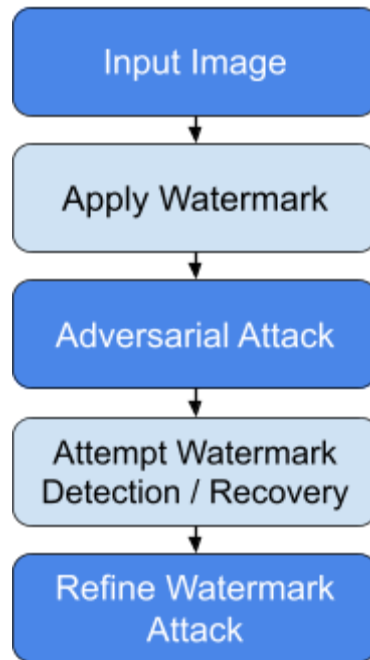
Tools/Libraries: Invisible Watermark (PyTorch), CelebA dataset (Kaggle)

Approach:

- Embed watermarks into face image dataset (potentially use a portion of CelebA dataset)
- Apply adversarial attacks to distort image
- Evaluate structural similarity index measure (SSIM), detection accuracy, and visibility of distortions

Cybersecurity context: evaluation of the limits of deepfake watermarking through adversarial perturbations of the images.

Adversarial Testing Pipeline








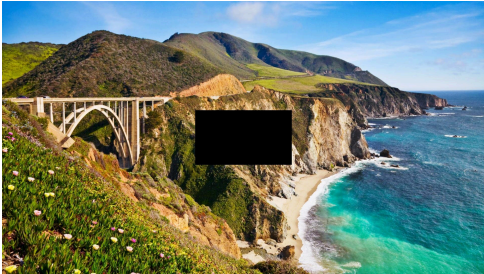

Preliminary Results

All three algorithms (dwtDct, dwtDctSvd, rivaGan) have been implemented and tested on a sample set.

Attacks tested: Gaussian noise, JPEG compression, brightness, overlay, mask, cropping, rotation, and resizing.

Attack	Image	dwtDct	dwtDctSvd	rivaGan
No Attack		✓	✓	✓

JPEG Compression		✗	✗	✓
Resize		✓	✓	✓
Gaussian Noise		✓	✓	✓
Crop		✗	✗	✓
Brightness		✓	✗	✓

Overlay		✓	✓	✓
Mask		✓	✓	✓
Rotate		✗	✗	✗

Legend:

✓: Exact match with expected watermark.

✗: Failure or decoding error.

Experimental Design Table

Experimental Design and Expectations

Experiment	Model	Dataset	Metric	Baseline	Expected Outcome
Exp1	DWT-DCT	CelebA	SSIM, Decode Accuracy	Unattacked	Degradation under blur, resize
Exp2	DWT-DCT-SVD	CelebA	SSIM, Decode Accuracy	Unattacked	Better robustness under distortion
Exp3	RivaGAN	CelebA	Decode Accuracy	Unattacked	Most resilient

Reproduction of SOTA

- LampMark (Paused): Attempted to implement LampMark, a perceptual watermarking method for proactive deepfake detection, using the CelebA-HQ dataset. I reached out to the creators of CelebA-HQ to see if they can directly share the dataset with me. I am waiting for a response from them.

- **Invisible Watermark** (Active): results above.

Model: Invisible Watermark (Zhang et al., 2019)

Details: Reproduced encoder-decoder architecture; tested watermark recovery post-transformation

Challenges: Some robustness issues observed under high-resize or aggressive cropping

Evaluation: I've implemented tests on the original test image to compare my results with the author's. Because the tested attacks are not specified in the repository, we had to estimate and recreate the attacks. Because of this, there is some room for error comparing the two implementations. Looking forward I will test the 3 watermarks on a larger dataset and add in non-benign image attack.

Watermark Robustness Comparison Table – Author vs. My Reproduction

Attack	Original Repo (Freq Method)	Original Repo (RivaGAN)	My Result (DWT-D CT)	My Result (DWT-DC T-SVD)	My Result (RivaGAN)	Notes
No Attack	✓ Pass	✓ Pass	✓	✓	✓	All methods pass, confirming correct baseline
JPEG Compression	✓ Pass	✓ Pass	✗ (decode error)	✗ (decode error)	✓	My RivaGAN matches SOTA; others struggle
Resize 50%	✗ Fail	✗ Fail	✓	✓	✓	My methods surprisingly resilient here
Gaussian Noise	✓ Pass	✓ Pass	✓	✓	✓	Fully consistent with original results

Crop 7×5	✗ Fail	✓ Pass	✗ (decode error)	✗ (decode error)	✓	Reproduction consistent; RivaGAN robust
Brightness	✓ Pass	✓ Pass	✓	✗ (decode error)	✓	DWT-DCT-SVD sensitive to brightness changes
Overlay	✓ Pass	✓ Pass	✓	✓	✓	All methods robust under overlays
Mask	✓ Pass	✓ Pass	✓	✓	✓	Confirmed robustness across methods
Rotate 30°	✗ Fail	✗ Fail	✗ (decode error)	✗ (decode error)	✗ (decode error)	Rotation remains challenging across the board

Next Steps

- Test DWT-DCT, DWT-DCT-SVD, and RivaGAN on a larger dataset (CelebA) with various attacks.
- Test the robustness of the watermarks against non-benign deepfake attacks.
- Measure perceptual quality degradation.
- (Once given access to CelebA-HQ) implement LampMark.

References

- [1] T. Wang, S. Yang, and Y. Wang, *LampMark: Proactive Deepfake Detection via Training-Free Landmark Perceptual Watermarks*, Proceedings of the ACM International Conference on Multimedia (MM), 2024. [Online]. Available: <https://dl.acm.org/doi/10.1145/3664647.3680869>
- [2] H. Zhang, Y. Liu, C. Yu, J. Chen, and L. Liu, *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*, Journal of King Saud University – Computer and Information Sciences, Elsevier, 2023. [Online]. Available: <https://doi.org/10.1016/j.jksuci.2022.10.020>
- [5] ShieldMnt Team, *Invisible Watermark GitHub Repository*, <https://github.com/ShieldMnt/invisible-watermark>, Accessed June 2025.
- [6] T. Wang, *LampMark GitHub Repository*, https://github.com/wangty1/LampMark/tree/main/image_data, Accessed June 2025.
- [7] T. Karras et al., *Progressive Growing of GANs (TensorFlow implementation)*, https://github.com/tkarras/progressive_growing_of_gans, Accessed June 2025.