Week 5 Report

Note: This week builds on the robustness evaluation framework established in Week 4, expanding the experimental pipeline with threshold-based attack analysis and higher-resolution image testing.

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, 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.

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., RivaGAN) 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

Paper Title: *DeepSigns: An End-to-End Watermarking Framework for Ownership Protection of Deep Neural Networks*
Authors: Bita Darvish Rouhani, Huili Chen, and Farinaz Koushanfar
Publication Year: 2019
Summary
Key Methods

- DeepSigns presents an end-to-end watermarking framework that embeds watermarks into the activation maps of a deep neural network's intermediate layers.
- Embedding is guided by a custom loss function that ensures both robustness and fidelity without requiring modifications to the original network architecture.
- The watermark can be embedded during model training or post-training via fine-tuning, and is extracted through a key-dependent decoding process.

Key Findings

- DeepSigns demonstrates strong robustness against common model modification attacks such as:
    - Fine-tuning
    - Model pruning
    - Compression
    - Transfer learning
- The authors show minimal impact on model accuracy and high watermark extraction fidelity even under aggressive tampering.
- Compared to prior works, DeepSigns performs better in black-box verification scenarios, where access to model internals is restricted.

Limitations

- Focused on model-level watermarking, not data/image-level — it protects the neural network's parameters, not media assets like images or videos.
- It does not address image-space transformations such as JPEG compression, resizing, or geometric attacks (which are central to your project).
- Watermark embedding requires training-time or fine-tuning access to the model, which may not apply to inference-only deployment contexts.

Relevance to My Project

Aspect	DeepSigns	Your Project
Watermark Target	Neural network model parameters (activations)	Image pixel space (DWT-DCT, RivaGAN, etc.)

Goal	Model ownership protection	Image authenticity + deepfake traceability
Evaluation Attacks	Pruning, fine-tuning, compression	Crop, rotate, resize, JPEG, noise, brightness, masking
Method Type	Embedded during training	Embedded pre-attack; tested for robustness post-attack
Output Measured	Watermark bit recovery from model	Watermark bit recovery from image

While both projects focus on imperceptible watermarking and robustness, ours targets media forensics and proactive content protection, aligning more with cybersecurity use cases around deepfake detection and image tampering.

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.

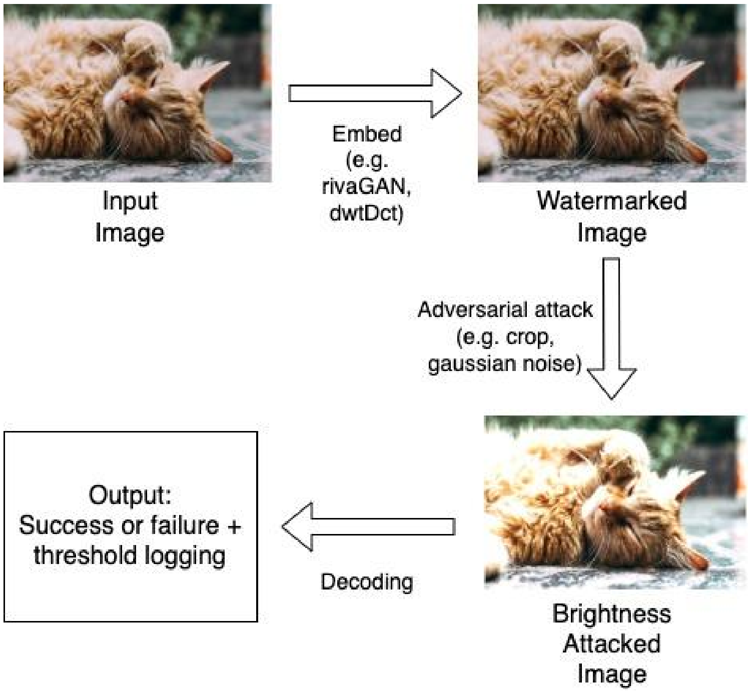
Limitations:

- 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.

Conceptual Diagram


Input Image
Embed (e.g. rivaGAN, dwtDct)
Watermarked Image
Adversarial attack (e.g. crop, gaussian noise)
Brightness Attacked Image
Decoding
Output: Success or failure + threshold logging

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)

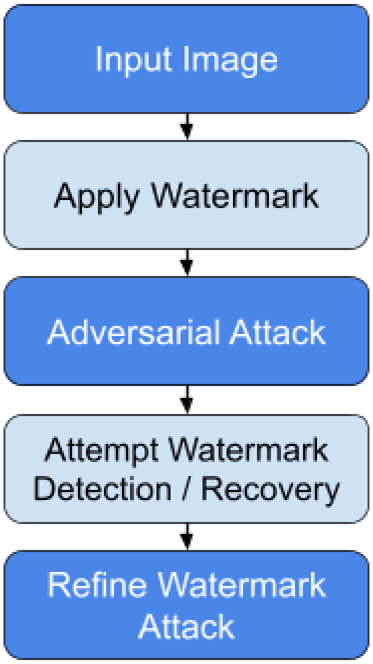
Tools/Libraries: Invisible Watermark (PyTorch), CelebA dataset (Kaggle)

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



Experimental Design Table

Experimental Design and Expectations

Experiment	**Model**	**Dataset**	**Metric**	**Baseline**	**Preliminary Results**
Exp 1	DWT-DCT	Unsplash (512×512)	Bitwise Decode Accuracy	Unattacked image	Fails on most clean 512×512 images
Exp 2	DWT-DCT -SVD	Unsplash (512×512)	Bitwise Decode Accuracy	Unattacked image	Slightly more robust than DWT-DCT
Exp 3	RivaGAN	Unsplash (512×512)	Bitwise Decode Accuracy	Unattacked image	Most robust overall, but image-sensitive
Exp 4	DWT-DCT	Unsplash (Original)	Bitwise Decode Accuracy	Unattacked image	Performs better than 512×512, but still weak to crop/rotate
Exp 5	DWT-DCT -SVD	Unsplash (Original)	Bitwise Decode Accuracy	Unattacked image	More stable at high-res, esp. resize/JPEG

Preliminary Results & Key Findings

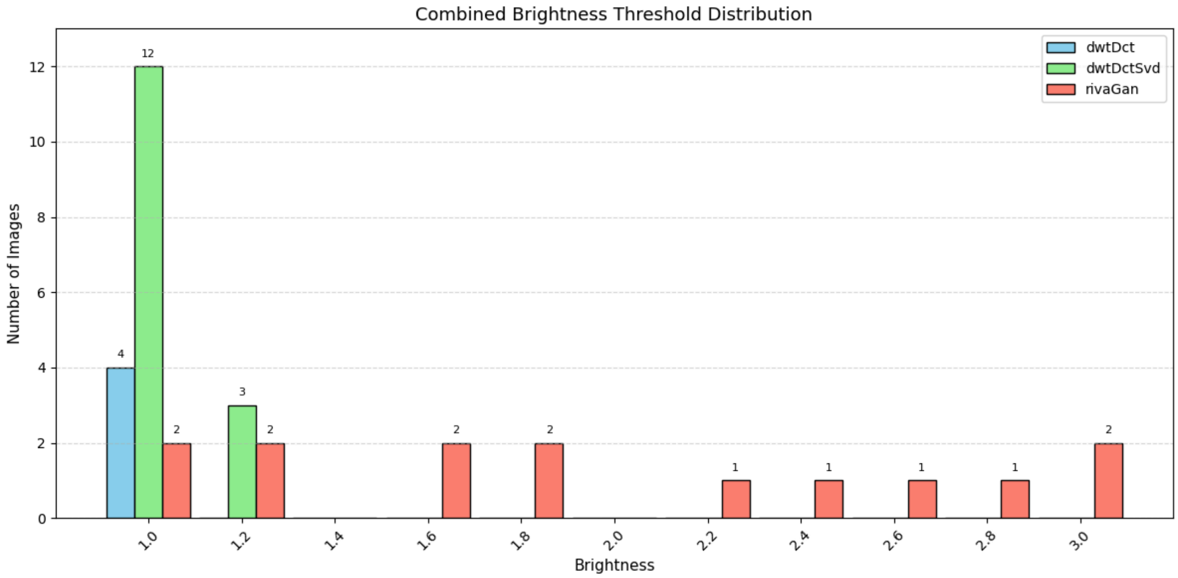
As of Week 5, I completed implementation of a threshold-based watermark robustness testing pipeline using three models — DWT-DCT, DWT-DCT-SVD, and RivaGAN — across 15 diverse 512×512 images from Unsplash. I tested 9 types of attacks at increasing intensities to determine decoding failure points for each image-method pair.

Notable Results

Attack Type	Robustness Observation
Crop	RivaGAN survives up to 20–30%; classical methods fail below 10%
Brightness ↓	RivaGAN can survive down to 0.2× for some images, others fail at 0.8×
JPEG	RivaGAN decodes down to Q20–35
Mask	RivaGAN survives up to 85% masked area; classical methods fail earlier around 25%
Rotate	All models fragile; classical methods instantly fail, rivaGAN decoding fails between 10°–16°

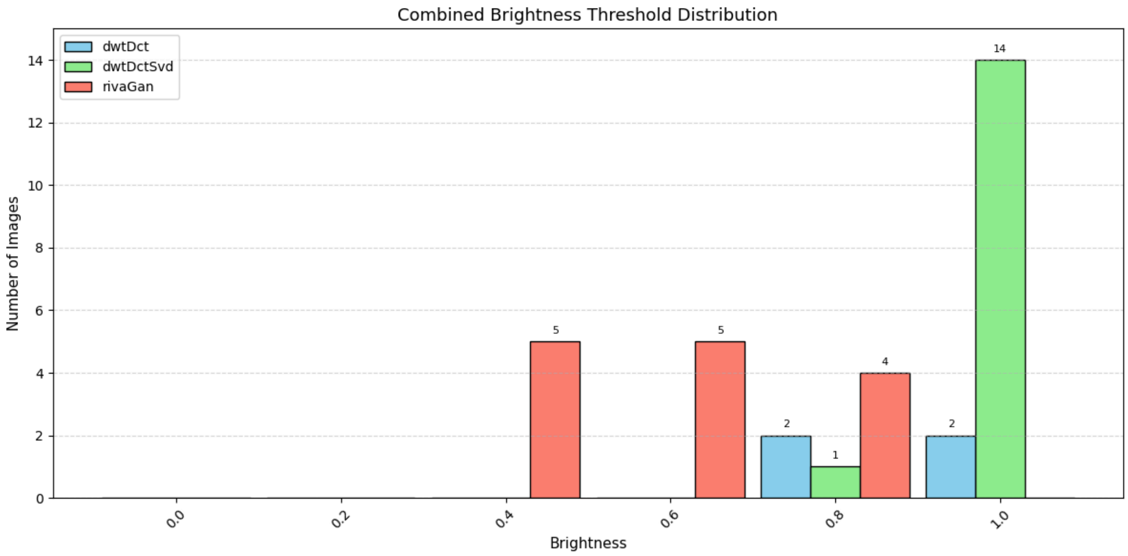
CSV summaries and visualizations for each attack (bar plots, threshold distributions) are included in the GitHub repo: https://github.com/catlewin/invisible-watermark-cat/tree/main/threshold_tests

512x512 Results Bar Graphs



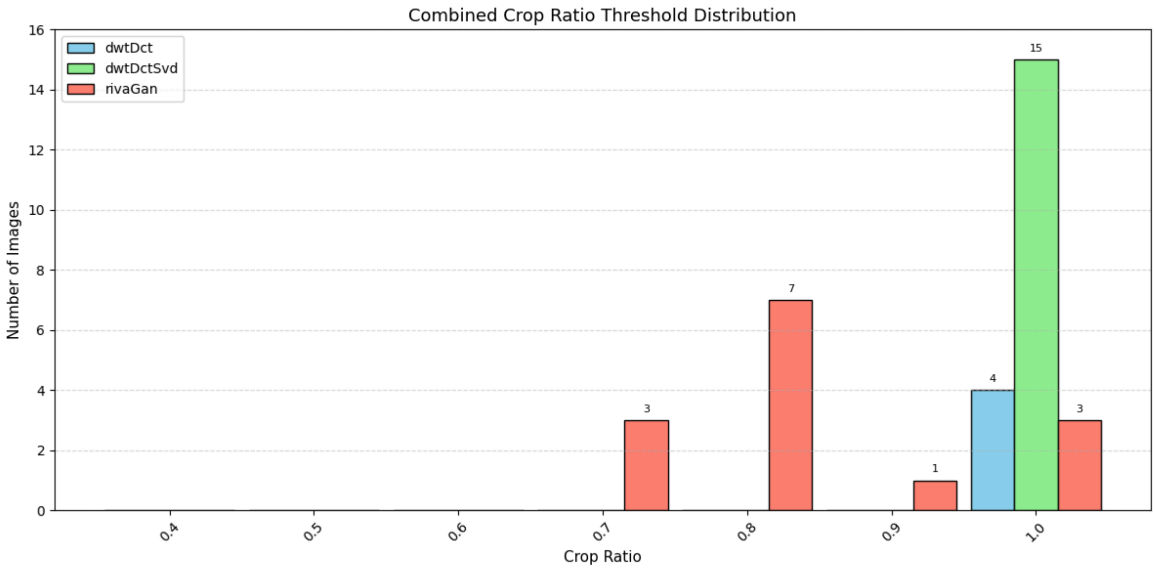
Combined Brightness Threshold Distribution

Method	Images	Clean Failures	Attack Failures	# Valid Thresholds	Avg Threshold	Median	Std Dev	Min	Max
dwtDct	15	11	11	4	1.00	1.00	0.00	1.00	1.00
dwtDctSvd	15	0	0	15	1.04	1.00	0.08	1.00	1.20
rivaGan	15	1	1	14	1.94	1.80	0.70	1.00	3.00

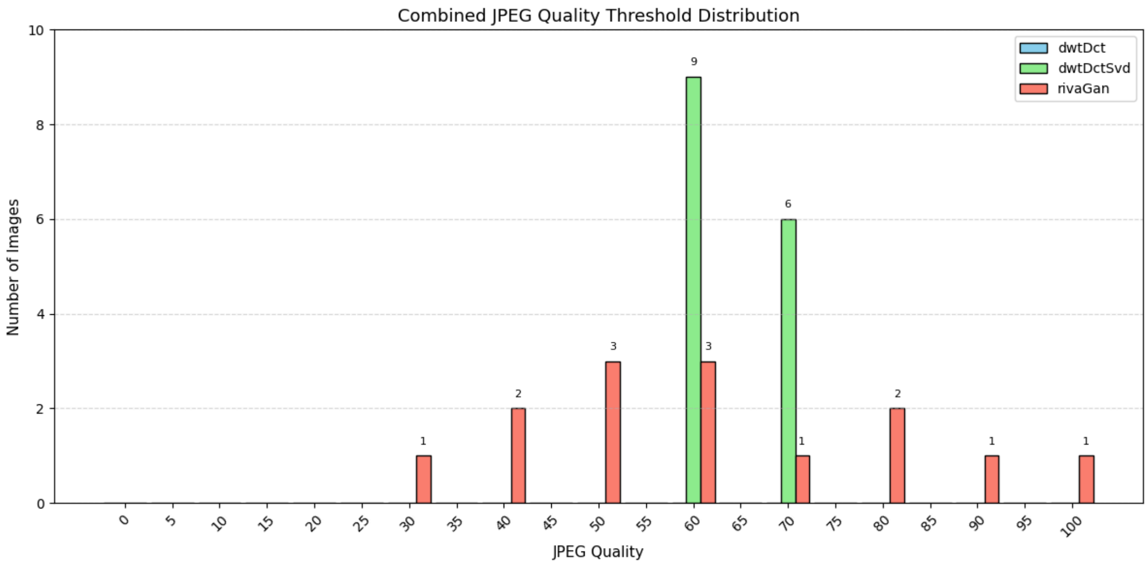


Combined Brightness Threshold Distribution

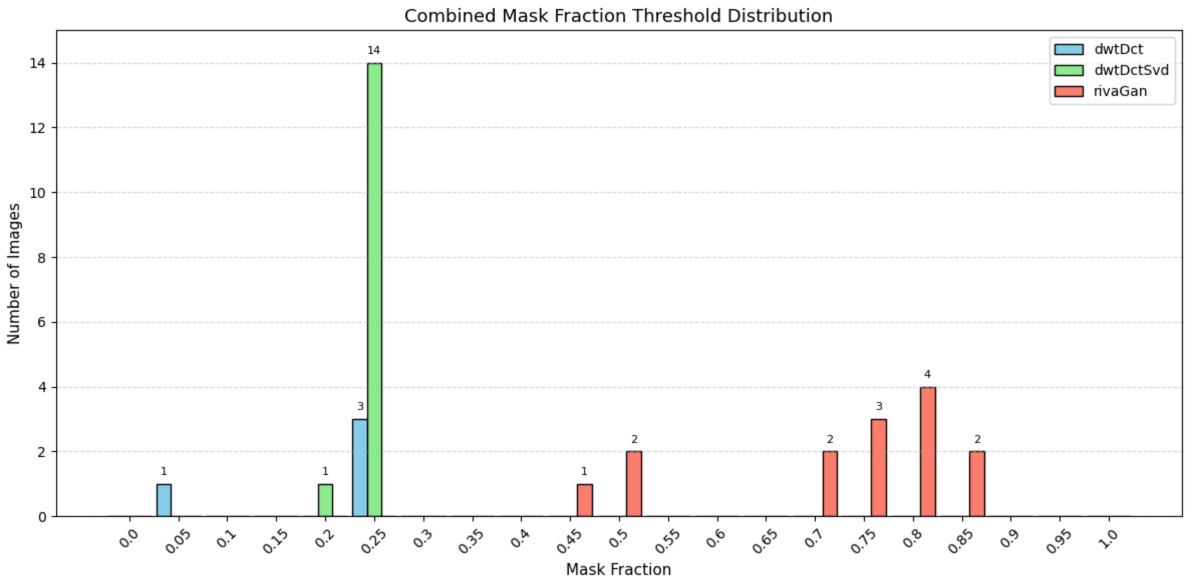
Method	Images	Clean Failures	Attack Failures	# Valid Thresholds	Avg Threshold	Median	Std Dev	Min	Max
dwtDct	15	11	11	4	0.90	0.90	0.10	0.80	1.00
dwtDctSvd	15	0	0	15	0.99	1.00	0.05	0.80	1.00
rivaGan	15	1	1	14	0.59	0.60	0.16	0.40	0.80


Combined Crop Ratio Threshold Distribution

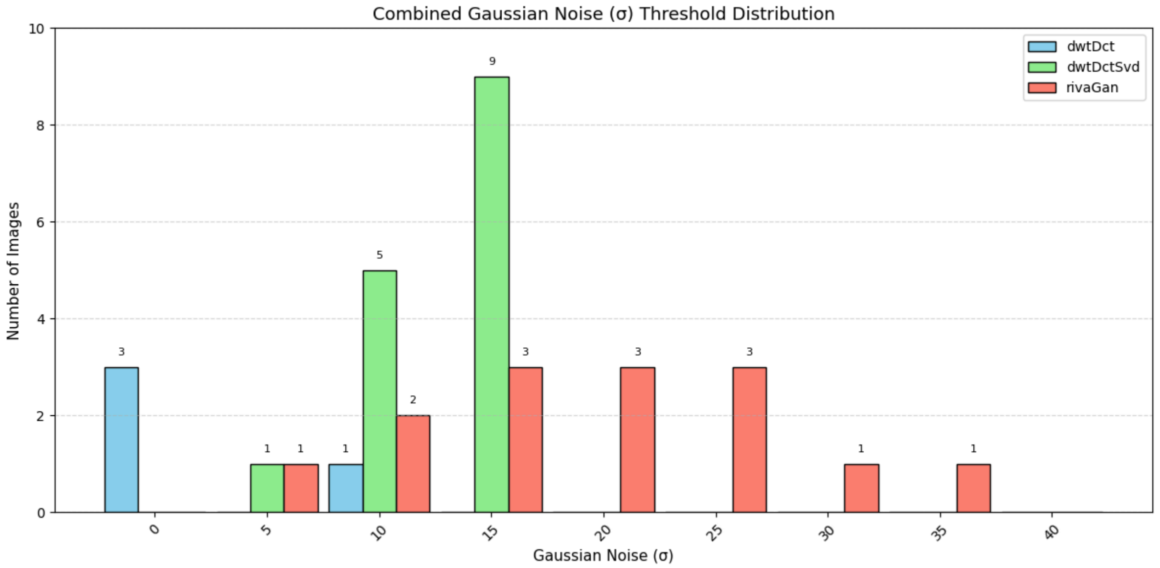
Method	Images	Clean Failures	Attack Failures	# Valid Thresholds	Avg Threshold	Median	Std Dev	Min	Max
dwtDct	15	11	11	4	1.00	1.00	0.00	1.00	1.00
dwtDctSvd	15	0	0	15	1.00	1.00	0.00	1.00	1.00
rivaGan	15	1	1	14	0.83	0.80	0.10	0.70	1.00


Combined JPEG Quality Threshold Distribution

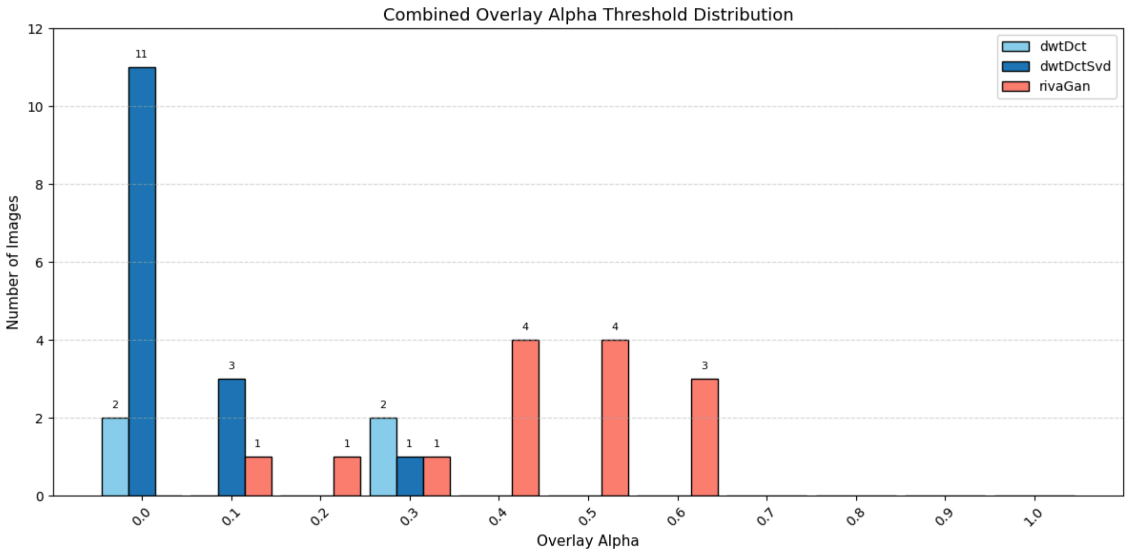
Method	Images	Clean Failures	Attack Failures	# Valid Thresholds	Avg Threshold	Median	Std Dev	Min	Max
dwtDct	15	11	15	0	--	--	--	--	--
dwtDctSvd	15	0	0	15	64.00	60.00	4.90	60.00	70.00
rivaGan	15	1	1	14	61.43	60.00	19.59	30.00	100.00



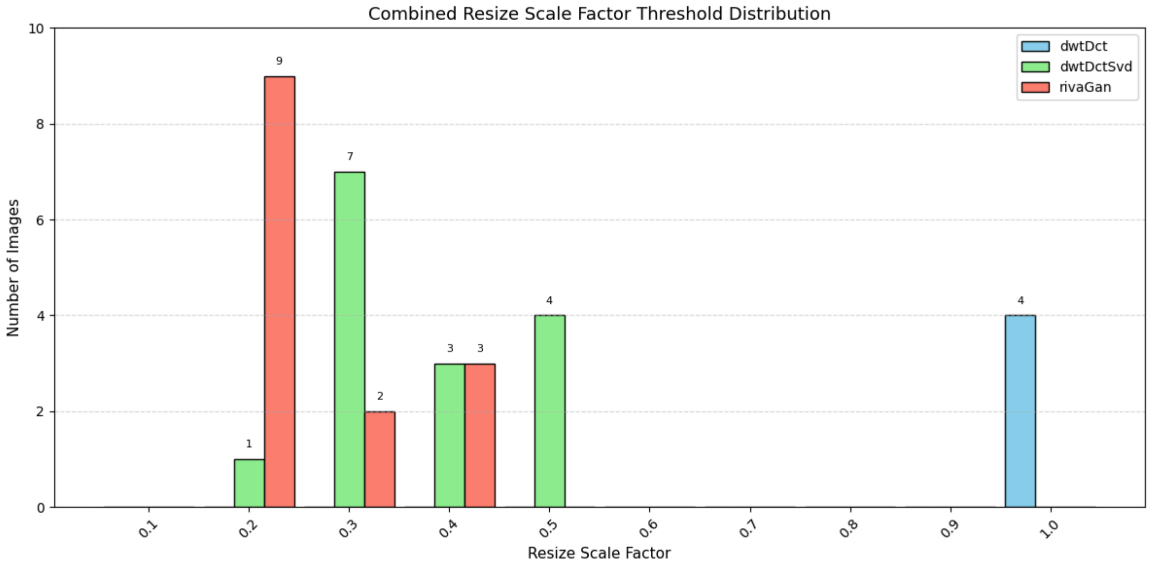
Method	Images	Clean Failures	Attack Failures	# Valid Thresholds	Avg Threshold	Median	Std Dev	Min	Max
dwtDct	15	11	11	4	0.20	0.25	0.09	0.05	0.25
dwtDctSvd	15	0	0	15	0.25	0.25	0.01	0.20	0.25
rivaGan	15	1	1	14	0.71	0.75	0.13	0.45	0.85



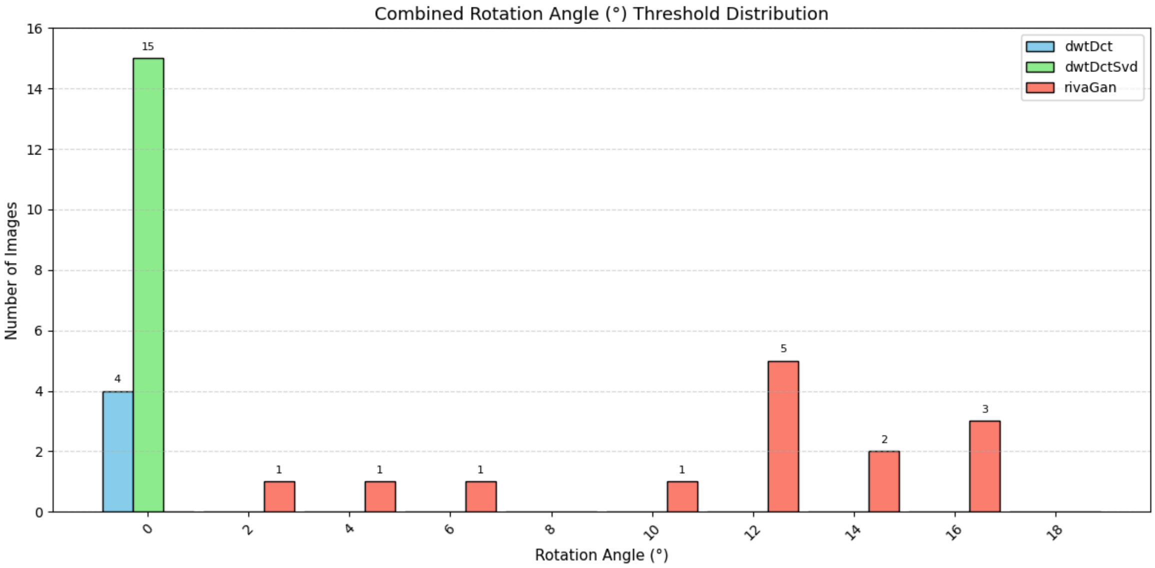
Method	Images	Clean Failures	Attack Failures	# Valid Thresholds	Avg Threshold	Median	Std Dev	Min	Max
dwtDct	15	11	11	4	2.50	0.00	4.33	0.00	10.00
dwtDctSvd	15	0	0	15	12.67	15.00	3.09	5.00	15.00
rivaGan	15	1	1	14	19.29	20.00	7.99	5.00	35.00



Method	Images	Clean Failures	Attack Failures	# Valid Thresholds	Avg Threshold	Median	Std Dev	Min	Max
dwtDct	15	11	11	4	0.15	0.15	0.15	0.00	0.30
dwtDctSvd	15	0	0	15	0.04	0.00	0.08	0.00	0.30
rivaGan	15	1	1	14	0.43	0.45	0.14	0.10	0.60



Method	Images	Clean Failures	Attack Failures	# Valid Thresholds	Avg Threshold	Median	Std Dev	Min	Max
dwtDct	15	11	11	4	1.00	1.00	0.00	1.00	1.00
dwtDctSvd	15	0	0	15	0.37	0.30	0.09	0.20	0.50
rivaGan	15	1	1	14	0.26	0.20	0.08	0.20	0.40



Method	Images	Clean Failures	Attack Failures	# Valid Thresholds	Avg Threshold	Median	Std Dev	Min	Max
dwtDct	15	11	11	4	0.00	0.00	0.00	0.00	0.00
dwtDctSvd	15	0	0	15	0.00	0.00	0.00	0.00	0.00
rivaGan	15	1	1	14	11.29	12.00	4.25	2.00	16.00

Reflections

What's Working Well

- The threshold-testing pipeline runs automatically across attack types and is scalable for more images or hybrid attacks.
- RivaGAN consistently outperforms classical methods across most attacks, especially crop, JPEG compression, and brightness variations.
- Classical methods show moderate performance under resize and mask attacks, with DWT-DCT-SVD slightly outperforming DWT-DCT overall.
- Testing on original high-resolution images (in addition to resized) revealed that image quality has a clear impact on watermark survival, especially for DWT-DCT-based methods.

Unexpected Observations / Challenges

- DWT-DCT failed to decode most clean 512×512 images (only 4/15 successful), but did better on high-res originals.
- RivaGAN failed on one unattacked image, suggesting some content sensitivity.

- Some images decoded only after being attacked (e.g., after brightness increase or masking), which may indicate that certain alterations unintentionally enhance decoder alignment — a possible direction for perceptual analysis.

Alignment with Hypothesis

- Hypothesis: RivaGAN would be more robust → Supported
- Hypothesis: High-res images would greatly improve classical model robustness → Partially supported (some improvement, but crop/rotate remain weak points)
- Hypothesis: Watermark failure would scale predictably with attack strength → Partially true; results show image-specific threshold variance, indicating a role for perceptual modeling in future work.

Next Steps

- Integrate perceptual modeling (e.g., SSIM, LPIPS) to evaluate decode variance
- Introduce generative transformations (e.g., GAN-based face swaps) to assess post-synthesis watermark survival

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.

[8] B. D. Rouhani, H. Chen, and F. Koushanfar, *DeepSigns: An End-to-End Watermarking Framework for Ownership Protection of Deep Neural Networks*, Proceedings of the ACM International Conference on Multimedia (ACM MM), 2019. [Online]. Available: https://dl.acm.org/doi/10.1145/3297858.3304051. Accessed June 2025.