# Week 3 Report

**Note:** This past week included multiple shifts in our research focus. Our preliminary testing found that a speech-only copyright attack is currently not feasible using YouTube's detection system. We then shifted to the field of deepfake detection and adversarial attacks on watermarked images. The first framework we attempted to implement–LampMark–proved too challenging to realistically implement in the immediate future. Moving forward, we will implement the Invisible Watermark framework and create adversarial attacks on the watermarked images.

## Project Information

Project Type: Individual
Student Name: Cat Lewin
Mentor Name: Dr. Rui Duan
Research Title: Speech Copyright Detection and Deepfake Robustness via Adversarial Attacks on Watermarked Images

## Problem Statement 1: Speech Copyright Detection

This project examines the limitations of automated copyright detection systems by evaluating whether pure speech content can evade detection, thereby revealing potential vulnerabilities in existing enforcement mechanisms.

## Problem Statement 2: 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. **Speech Hypothesis:** Pure speech audio, without musical accompaniment or sampling, will not trigger YouTube's Content ID claims.
2. **Watermark Hypothesis:** Adversarial attacks will degrade watermark detection accuracy, particularly when designed to preserve perceptual fidelity.

## Research Questions

1. **Speech Questions:**
    - What types of non-musical audio (e.g. movie monologues, famous speeches) can trigger copyright claims?
    - What perturbations added to non-musical audio can successfully evade YouTube's Content ID copyright detection system?
2. **Watermark Questions**

- How robust are perceptual watermarks (e.g., Invisible Watermark) under adversarial image perturbations?
- What metrics best evaluate watermark visibility, resilience, and deepfake traceability?

**Literature Review**

**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.

**Title:** *Parrot-Trained Adversarial Examples: Pushing the Practicality of Black-Box Audio Attacks against Speaker Recognition Models*
**Authors:** Rui Duan, Yao Liu, Zhe Qu, Leah Ding, Zhou Lu
**Publication Year:** 2024

- **Summary:** This paper came up with a way to attack speaker recognition systems without needing to know anything about how the system works internally

(black-box). They used a technique that mimics a person’s voice (“parrot-trained”) to fool the system.
- **Contributions:** They introduced the idea of "parrot-trained" surrogate models, which can be built from just a short audio clip of someone's voice. These models make it possible to create adversarial audio that even works when played out loud into a microphone—no direct access to the target system needed.
- **Limitations:** This work focuses on attacking speaker recognition systems, so it doesn't really touch on general audio content or how people perceive the audio changes.
- **Expand/Improve:** I’m taking their idea further by applying it to broader detection systems like copyright enforcement, and adding models that better reflect how humans hear and interpret speech.

**Title:** *Perception-Aware Attack: Creating Adversarial Music via Reverse-Engineering Human Perception*
**Authors:** Rui Duan, Leah Ding, Zhe Qu, Yao Liu, Shangqing Zhao, Zhou Lu
**Publication Year:** 2022

- **Summary:** This paper researched how to make audio that tricks detection systems but still sounds normal to people. They trained a model to understand what kinds of distortions humans can’t easily notice, and used that to guide their attacks.
- **Contributions:** They built a regression model trained on real human feedback to measure how noticeable audio changes are. This model was then used to generate music adversarial examples that sneak past detection algorithms while sounding almost identical to the original.
- **Limitations:** The study did a great job with music, but it didn’t explore how these techniques might work for speech, which is different in how it's processed and understood by listeners.
- **Expand/Improve:** I'm adapting their perception-aware framework to focus on spoken content, looking specifically at how intelligible speech remains when adversarial noise is added.

**Title:** *Audio Adversarial Examples: Targeted Attacks on Speech-to-Text*
**Authors:** Nicholas Carlini, David Wagner
**Publication Year:** 2018

- **Summary:** This paper showed that it’s possible to subtly change an audio clip so that an automatic speech recognition (ASR) system hears something completely different—like hearing “okay google, search evil.com” instead of “play some music”—without humans noticing anything weird.

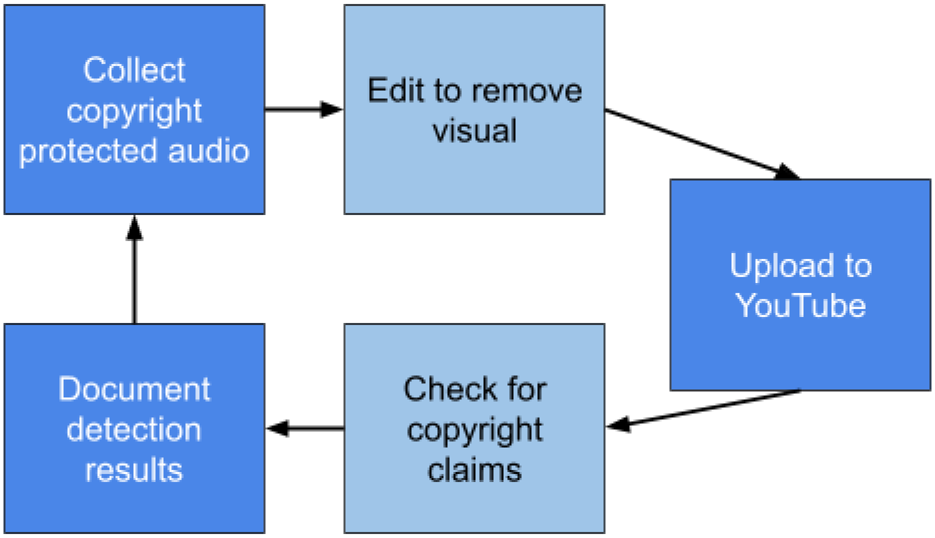
- **Contributions:** They created a white-box attack method that could reliably get ASR systems to output specific phrases, even with very small audio changes. It was one of the first to show how vulnerable these systems are to targeted manipulation.
- **Limitations:**Their approach is highly effective for speech-to-text, but it doesn't quantify how perceptible the sound disturbances are, and it doesn't look at other types of detection systems. Additionally, this is a white-box attack specific to Mozilla's DeepSpeech, and is not a particularly feasible attack in real time.
- **Expand/Improve:** I'm using models that try to align better with human perception and shifting the focus from changing transcriptions to evading systems like Content ID.

**Title:** *Adversarial attacks on Copyright Detection Systems*
**Authors:** Parsa Saadatpanah, Ali Shafahi, Tom Goldstein
**Publication Year:** 2019

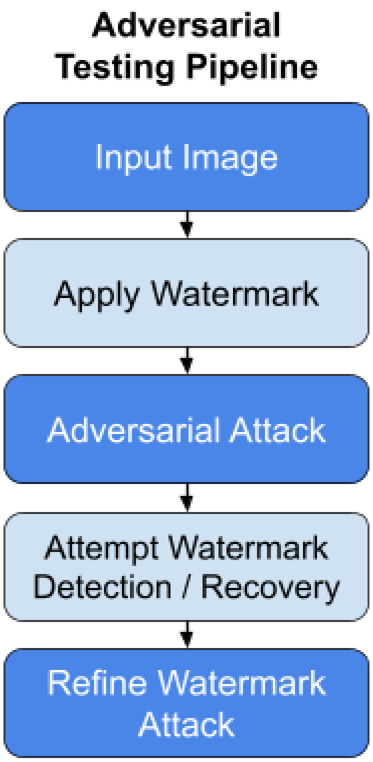
- **Summary:** This study looked at how to evade copyright detection systems, like YouTube's Content ID, with slight changes to the audio. They found that you can often bypass detection without making the clip sound any different to a human.
- **Contributions:** They built and tested attacks on both open and closed systems, showing that even very small perturbations could evade detection. They also explored whether attacks that worked on one system could transfer to another.
- **Limitations:** The paper shows that these systems can be tricked, but mostly with music and without formally measuring how noticeable the changes are to listeners.
- **Expand/Improve:** I'm testing similar attacks on speech content, and working to make sure the changes are less noticeable using perceptual models.

**High-Level Proposed Approach**

1. **Speech Copyright Project:**
    - Evaluating YouTube's Content ID system, an existing proprietary AI system:
        - Uses a combination of audio fingerprinting and machine learning (likely convolutional networks, but details are not public).
        - It performs closed-source classification to detect matches between uploaded content and a reference database.
    - Tools: YouTube Content ID, iMovie
    - Cybersecurity context: Identify system boundaries and blind spots of YouTube's copyright detection system, simulating adversarial copyright infringement.


Collect copyright protected audio
Edit to remove visual
Upload to YouTube
Check for copyright claims
Document detection results

2. **Watermarking Robustness Project:**
    - 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.



### Experimental Design Table

Table 1: Copyright Detection

| Clip Type | Clip Name | Claim? | Source | Notes |
|---|---|---|---|---|
| Speech | MLK Jr. – I Have a Dream | Yes | To the Left by DjeefSound | Likely false positive |
| Speech | Greta Thunberg – How Dare You | Yes | Eines Tages by MUSA & Oga Beats | Speech sampled in music |
| Speech | Barack Obama – A More Perfect Union | No | - | |
| Speech | JFK – Inaugural Address | No | - | |
| Speech | Wendy Suzuki TEDxTalk | No | - | |
| Podcast | Radiolab - Everybody's Got One Episode | No | - | |

| | | | | |
|---|---|---|---|---|
| Podcast | Joe Rogan #2312 | No | - | |
| Movie | Star Wars V - “I am your father" | No | - | |
| Movie | How to Train Your Dragon | No | - | |
| Movie | The Dark Knight – Joker Monologue | Yes | Intermezzo by UanmNess | Background audio matched |

Preliminary Experiment Table – Watermarking Robustness Project

| Component | Description |
|---|---|
| **Dataset** | Likely a random sample from CelebA |
| **Baseline Model** | Invisible Watermark (Zhang et al., 2019) – CNN-based encoder/decoder |
| **Adversarial Methods** | Still deciding, will solidify once Invisible Watermark is successfully implemented. |
| **Perturbation Targets** | Watermarked images (perturbations applied post-embedding) |
| **Evaluation Metrics** | - SSIM (Structural Similarity Index)<br>- Watermark recovery accuracy<br>- Visual imperceptibility |

### Reproduction of SOTA

- LampMark (Paused): Attempted to implement LampMark, a perceptual watermarking method for proactive deepfake detection, using the CelebA-HQ dataset. However, due to technical challenges with dataset reconstruction, Dr. Duan and I decided to shift to an easier to implement framework for now.
- Invisible Watermark (Active): This framework will be the reproducible baseline going forward.

### Reflections & Planning

**Reflections:** This week highlighted the fragility of current audio detection systems, particularly regarding non-musical speech content. I was surprised to find that YouTube seemingly either does not include non-musical audio within their database of copyright material or their ContentID system fails to identify non-musical copyrighted audio. This week also reinforced the importance of accessibility and feasibility when selecting research tools. Although implementing LampMark proved technically infeasible, this experience underscored the importance of flexibility and responsiveness in research planning.

By shifting to the Invisible Watermark project, I aim to stay aligned with the program timeline and broader goals of perceptual robustness and media integrity, while continuing to develop skills in adversarial testing.

### Next Steps

- Implement the Invisible Watermark method in Python (from GitHub repo).
- Design adversarial attack methods to degrade watermarked images.
- Measure detection success rate under attack and perceptual quality degradation.

## 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, *Invisible Watermarking of Deep Neural Networks for Intellectual Property Protection*, arXiv preprint arXiv:1909.01285, 2019.

[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.

[4] R. Duan, Y. Liu, Z. Qu, L. Ding, and Z. Lu, *Parrot-Trained Adversarial Examples: Pushing the Practicality of Black-Box Audio Attacks against Speaker Recognition Models*, arXiv preprint arXiv:2402.11290, 2024.

[5] R. Duan, L. Ding, Z. Qu, Y. Liu, S. Zhao, and Z. Lu, *Perception-Aware Attack: Creating Adversarial Music via Reverse-Engineering Human Perception*, Proceedings of the ACM International Conference on Multimedia (MM), 2022, pp. 5535–5543. doi: 10.1145/3503161.3548378.

[6] N. Carlini and D. Wagner, *Audio Adversarial Examples: Targeted Attacks on Speech-to-Text*, 2018. [Online]. Available: https://nicholas.carlini.com/writing/2018/audio-adversarial.html

[7] P. Saadatpanah, A. Shafahi, and T. Goldstein, *Adversarial Attacks on Copyright Detection Systems*, 2019. [Online]. Available: https://arxiv.org/abs/1908.05238

[8] T. Wang, *LampMark GitHub Repository*, https://github.com/wangty1/LampMark/tree/main/image_data, Accessed June 2025.

[9] T. Karras et al., *Progressive Growing of GANs (TensorFlow implementation)*, https://github.com/tkarras/progressive_growing_of_gans, Accessed June 2025.

[10] ShieldMnt Team, *Invisible Watermark GitHub Repository*, https://github.com/ShieldMnt/invisible-watermark, Accessed June 2025.