CS-388 Project Proposal: Detecting AI-Generated Images

Cayden Knight

April 2025

Abstract

The increasing realism of AI-generated images poses new challenges in distinguishing authentic content from synthetic visual media. This project proposes a lightweight, interpretable detection system leveraging frequency-based analysis to identify structural artifacts characteristic of AI-generated imagery. By transforming images into the frequency domain and extracting azimuthally averaged magnitude spectra, the system aims to detect statistical anomalies in both traditional photographs and AI-generated artwork. The methodology involves preprocessing, frequency transformation using 2D FFT, and classification through a multilayer perceptron (MLP). Evaluation will focus on generalization across multiple generative models, robustness to post-processing, and effectiveness on abstract content.

1 Introduction

Artificial intelligence has enabled the development of advanced image synthesis tools, including Generative Adversarial Networks (GANs) and diffusion models, which can produce highly realistic images of faces, landscapes, and digital artwork. Tools such as StyleGAN, DALL·E, Midjourney, and Stable Diffusion allow users to create detailed visual content with minimal effort. While this has advanced creative expression, it also raises critical concerns around misinformation, identity manipulation, and the erosion of trust in visual media.

AI-generated images often lack the nuanced imperfections found in genuine human-made content—such as natural noise, texture inconsistencies, or irregular brushstrokes. These inconsistencies may not always be apparent in the spatial domain but can often be revealed through frequency-based analysis. This project investigates whether such spectral differences are consistently present in AI-generated images and whether they can be used to reliably distinguish synthetic content from authentic images.

2 Related Work

Early approaches to deepfake detection leveraged spatial-domain features, using Convolutional Neural Networks (CNNs) to detect local pixel anomalies or unnatural patterns. Guarnera et al. (2020) proposed detecting deepfakes by analyzing convolutional traces left by generative networks. Their model performed well on face-centric deepfakes but was less generalizable across diverse content (Guarnera et al., 2020). Similarly, Li et al. (2020) introduced Face X-ray, which detects splicing artifacts by decomposing image layers using CNNs (Li et al., 2020).

Other CNN-based techniques include attention-based approaches such as that of Wang et al. (2020), which targeted facial regions to enhance detection accuracy (Wang et al., 2020). Marra et al. (2018) demonstrated the applicability of CNN models in detecting GAN-generated images on social media platforms (Marra et al., 2018).

However, spatial approaches face challenges when applied to non-facial or abstract content. As generative models diversify, there is a need for model-agnostic detection techniques. Frequency-based approaches offer a compelling alternative. Durall et al. (2020) showed that CNN-based GANs fail to replicate the natural spectral distributions of real images, leading to distinctive frequency-domain artifacts (Durall et al., 2020). Frank et al. (2020) proposed leveraging log-magnitude spectra and azimuthal averaging to identify consistent frequency anomalies across multiple GANs (Frank et al., 2020). These techniques have the advantage

of being generalizable, as they target fundamental structural properties rather than image semantics.

McCloskey and Albright (2019) further identified inconsistencies in color saturation cues resulting from the generative process, indicating that frequency-domain anomalies extend beyond luminance features (McCloskey & Albright, 2019). Compression-based techniques, as explored by Marra et al. (2019), analyze how real and synthetic images respond to JPEG compression, uncovering latent inconsistencies that persist even under heavy downsampling (Marra et al., 2019).

Tolosana et al. (2020) and Verdoliva (2020) offer comprehensive surveys of detection methods, highlighting the strengths and limitations of spatial, frequency-based, and hybrid approaches (Tolosana et al., 2020; Verdoliva, 2020). These studies provide a foundation for the proposed system, which builds on frequency-based methods due to their robustness and generalizability.

3 System Design

The proposed system will be implemented in Python using TensorFlow and Keras. It consists of several main stages: image preprocessing, frequency transformation, feature extraction, and classification.

First, all input images will be resized to 128×128 pixels, converted to grayscale, and normalized. This standardization facilitates consistent frequency analysis across diverse image types. Each image will then undergo a 2D Fast Fourier Transform (FFT), from which the magnitude spectrum is extracted. Log-scaling may be applied to enhance contrast in low-frequency components. Azimuthal averaging is used to condense the 2D spectrum into a 1D frequency profile, capturing the radial distribution of energy.

This 1D profile serves as the input to a multilayer perceptron (MLP) classifier, which will be trained to predict whether the image is real or AI-generated. Should the MLP un-

derperform, a 2D CNN classifier trained directly on raw magnitude spectra will be explored.

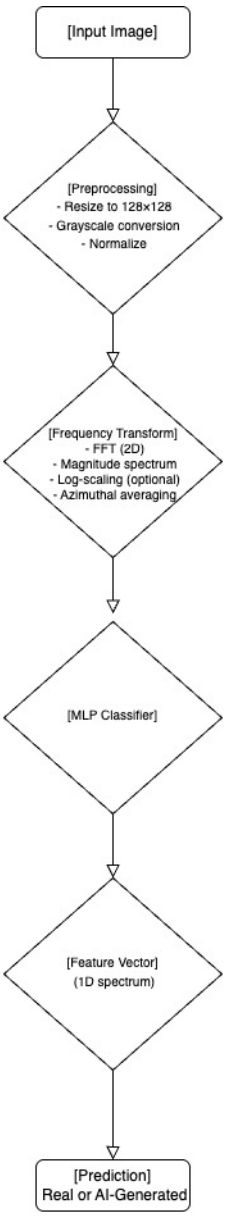

Figure 1: Overview of the proposed pipeline: images are preprocessed, transformed into the frequency domain, reduced via azimuthal averaging, and classified as real or synthetic.

4 Evaluation Plan

The system will be evaluated using a combination of real and synthetic datasets. Real images will be sourced from datasets such as FFHQ and CelebA-HQ. Synthetic counterparts will be generated using models such as StyleGAN2, Midjourney, and Stable Diffusion. The evaluation will include both conventional photographs and artistic content.

Performance will be measured using standard classification metrics: accuracy, precision, recall, F1-score, and ROC-AUC. Robustness will be tested by introducing common image post-processing techniques such as JPEG compression, Gaussian blurring, and resolution downscaling. Cross-model generalization will be assessed by training on one generator's outputs and testing on another.

5 Contributions

The heart of this project is a frequency-based detection system that can differentiate between real and AI-generated images, especially within both general photography and digital artwork. The core contributions will include:

- A clean, repeatable pipeline for converting images into frequency-domain profiles.
- A trained multilayer perceptron (MLP) classifier to detect AI-generated content from frequency profiles.
- A comparative evaluation across different GANs and diffusion models.
- Robustness testing against common transformations like compression and downsampling.

Auxiliary contributions may include experiments with alternative classifiers (like CNNs on spectral maps), analysis of cross-domain generalization (e.g., models trained on faces tested on art), and possible visualizations for interpretability.

6 Risk Analysis

Key risks include:

- **Spectral Artifacts Weakening:** Advanced generative models may better replicate natural frequency distributions. Contingency: Incorporate larger datasets and explore alternative descriptors.
- **Overfitting to Dataset Bias:** Classifier may learn dataset-specific quirks. Contingency: Use diverse sources and apply data augmentation.
- **Model Underperformance:** MLP may fail to capture complex patterns. Contingency: Train 2D CNN models on full spectra.

7 Timeline

- **Week 1:** Finalize dataset and generation pipeline.
- **Week 3:** Implement preprocessing and FFT pipeline.
- **Week 5:** Train MLP on frequency features; evaluate baseline performance.
- **Week 7:** Add post-processing robustness and cross-model testing.
- **Week 9:** Explore CNN alternative; conduct comparative analysis.
- **Week 12:** Compile results and prepare final report.

References

Durall, R., Keuper, M., & Keuper, J. (2020). Watch your up-convolution: Cnn-based generative deep neural networks are failing to reproduce spectral distributions. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 1–10.

Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., & Holz, T. (2020). Leveraging frequency analysis for deep fake image recognition. *Proceedings of the International Conference on Machine Learning (ICML) Workshops*.

Guarnera, L., Giudice, O., & Battiato, S. (2020). Deepfake detection by analyzing convolutional traces. *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 5061–5065.

Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Face x-ray for more general face forgery detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5001–5010.

Marra, F., Gragnaniello, D., Cozzolino, D., & Verdoliva, L. (2018). Detection of gan-generated fake images over social networks. *Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 384–389.

Marra, F., Gragnaniello, D., Cozzolino, D., & Verdoliva, L. (2019). Do gans leave artificial fingerprints? *Proceedings of the IEEE Conference on Multimedia Signal Processing (MMSP)*, 564–569.

McCloskey, S., & Albright, M. (2019). Detecting gan-generated imagery using saturation cues. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8301–8305.

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, *64*, 131–148. https://doi.org/https://doi.org/10.1016/j.inffus.2020.07.007

Verdoliva, L. (2020). Media forensics and deepfakes: An overview. *IEEE Journal of Selected Topics in Signal Processing*, *14*(5), 910–932. https://doi.org/https://doi.org/10.1109/JSTSP.2020.2998604

Wang, S., Xie, X., & Qiao, Y. (2020). Region attention based cnn for deepfake detection. *Proceedings of the ACM International Conference on Multimedia*, 3331–3339.