CS-488 Technical Report

Cayden Knight

September 2025

Abstract

As AI-generated images continue to close the gap between authentic and synthetic content, it becomes increasingly difficult to distinguish between the two with the naked eye. My project dives into the frequency domain to expose artifacts that aren't visible in the spatial domain. I tested a simple MLP that only looks at a 1D FFT profile and compared it to a deeper CNN that processes the full 2D FFT magnitude spectrum. The difference between the two was noticeable: the MLP topped out at 59%, while the CNN reached 88%. This indicates that the frequency domain does contain useful artifacts, but only a model with sufficient depth and spatial awareness can effectively learn and utilize them.

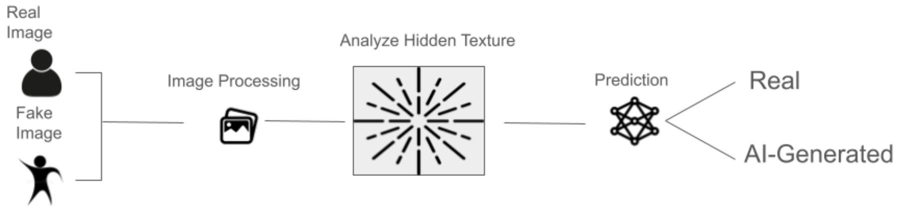

Figure 1: Visual abstract highlighting the process for predicting if images are real or AI-generated.

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 examines whether it can be used to distinguish the differences between synthetic images from the authentic images by comparing a simple MLP that uses 1D FFT to a deeper CNN module trained on full 2D FFT magnitude spectrum.

2 Literature Review

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 those of Wang et al. (2020), which target 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) demonstrated that CNN-based GANs fail to accurately replicate the natural spectral distributions of real images, resulting in 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 Methods

The overall system architecture for the CNN models, including the frequency transformation pipeline, is illustrated in Figure 2.

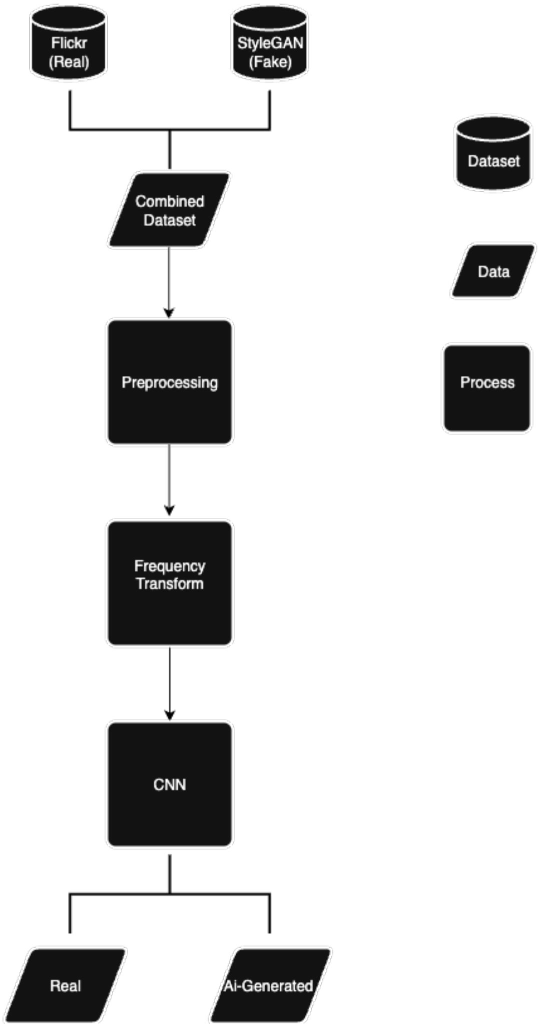

Figure 2: Data architecture diagram of the proposed AI-image detection system.

3.1 Datasets

I am using the 140k Real and Fake Faces dataset from Kaggle(xhlulu, 2020), a collection of authentic and AI-generated images. The dataset is sorted into training, validation, and test sets. The model was trained on 100,001 images, validated on 20,000 images, and will be tested on 20,001 images for final evaluation. All images were resized to 128 × 128 pixels.

3.2 Software

The project was implemented in Python using several libraries for machine learning and data processing. TensorFlow and Keras were used to build and train both the MLP and CNN models. The scipy.fft library was used to compute the Fast Fourier Transform (FFT), both for the offline preprocessing pipeline that generated the CNN's input data and within the MLP model itself. NumPy was used for numerical operations. Data augmentation and normalization were implemented through Keras preprocessing layers, and Matplotlib was used only for generating training and validation graphs.

3.3 Analysis Plan for MLP

The analysis followed a pipeline consisting of preprocessing, frequency transformation, feature extraction, and classification.

1. **Preprocessing**: Images were loaded in RGB format, resized to 128 × 128, and normalized to [0, 1].

2. **Frequency Transformation**: A 2D FFT was applied to each RGB channel. The log-magnitude spectra were computed and averaged together to form a single 128 × 128 frequency map.

3. **Feature Extraction**: The 2D spectrum was reduced to a 1D vector using azimuthal averaging. Each vector was padded or truncated to a fixed length of 91 values.

4. **Classification**: A simple MLP with layers of 128 and 64 units (ReLU activation) and a final sigmoid output was used for binary classification. The model was trained with Adam and binary cross-entropy loss.

3.4 Analysis Plan for CNN

The CNN model works on frequency-domain images rather than spatial images. Each RGB image in the dataset was converted into a $128 \times 128 \times 3$ log-magnitude FFT representation, which served as the input to the CNN model.

1. **Preprocessing**: All FFT magnitude images were stored and loaded as $128 \times 128 \times 3$ images. Normalization and CPU-based data augmentation (horizontal flips and small rotations) were applied during training.

2. **Model Architecture**: The final model was an 8-block residual CNN. Each block consisted of two 3×3 convolution layers with ReLU activations and batch normalization. Max pooling layers were inserted to downsize feature maps, and a Global Average Pooling layer reduced the spatial dimensions before classification. A dense layer with 512 units and a dropout rate of 0.35 was used before the sigmoid output.

3. **Training**: The model was trained using the AdamW optimizer with a tuned learning rate and weight decay. Early stopping and learning rate reduction were used to stabilize training. The model reached a test accuracy of approximately 88%.

4 Results

The MLP baseline struggled to learn frequency-domain differences and reached a final accuracy of approximately 59%. In contrast, the CNN performed much better when trained on the full 2D log-magnitude FFT representation. The final model achieved a test accuracy of roughly 88%, showing a clear improvement over the 1D MLP approach.

The training and validation curves (Figure 3) show that the CNN was able to consistently reduce validation loss and converge to a high-performing solution, while the MLP plateaued early and failed to capture the necessary spectral structure. These results indicate that detecting AI-generated images in the frequency domain benefits immensely from a deeper model with spatial awareness, such as the residual CNN used in this project.

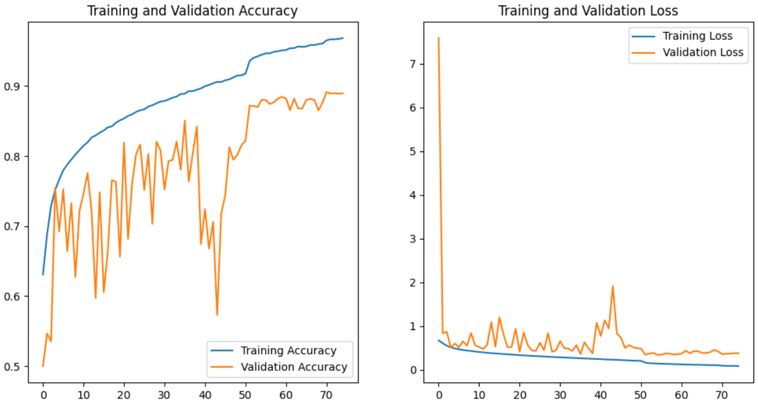

Figure 3: Training and validation accuracy and loss curves for the final CNN model.

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.

xhlulu. (2020). 140k real and fake faces [Accessed: 2025-03-01].