Detecting AI-Generated Images Through Spatial-Frequency Analysis and Diffusion-Based Reconstruction

Syed Ahmad Shah
Stevens Institute of Technology
Hoboken, NJ
sshah6@stevens.edu

I. Introduction

The goal of this project is to create a system able to distinguish AI Generated images from real photographs using content based analysis rather than relying on metadata or hidden watermarks. This problem is increasingly important as modern generative models are creating images that have become visually convincing, whilst common methods of tagging these images (watermarks, metadata, patterns, etc) can be easily removed through methods such as screenshotting, compression, and cropping.

This project is based on the hypothesis that AI generated images and real photographs are inherently different. AI-generated images are created through statistical distributions dictated by the learned trajectories a model takes in the creation of images, leading to pixels being related by mathematical regularities learned during training. In contrast, real photographs are formed through physical interactions involving light, camera optics, sensor noise, and the geometry of the scene. Even when both types of images may appear similar, they differ in the way their pixels are arranged.

To test this we explore two different methodologies:

- **A hybrid spatial frequency classifier** that combines a pretrained CNN (spatial features) and a Fast Fourier Transform (Frequencies) features, to then be classified by a final neural network.
- **A diffusion based reconstruction method** that determines how well a diffusion model restores "broken" images.

II. Hypothesis

The core hypothesis is that the statistical existence in AI generated images can be distinguished from physical coherence in real images.

For the frequency method, the assumption is that the diffusion model may leave spectral artifacts, recurring structures, or unnatural regularity that would not occur in a real image.

For the diffusion method, the assumption is that a diffusion model should be able to reconstruct images from its own trained distribution more effectively than that from real images. The idea is to have the diffusion model try to "repair" a real image, and a generated image, of which it should be able to better reconstruct the generated image through the statistical consistency of the rest of the image.

III. Dataset

There are two separate datasets for each method. For the hybrid model, the dataset contains a total of 1965 images split by training, validation, and test (70/20/10).

TABLE I
Hybrid Model Dataset Split

Subset	Images
Training set	1,375
Validation set	391
Test set	199

These images were acquired from two datasets available on Kaggle:

- https://www.kaggle.com/datasets/cashbowman/ai-generated-images-vs-real-images
- https://www.kaggle.com/datasets/rhythmghai/ai-vs-real-images-dataset

Class distribution:

TABLE II
Class Distribution

Class	Train	Validation	Test
Real images	824	234	121
AI images	551	157	78

The image categories include people, nature, cities, animals, and art, to prevent the model from overfitting on any one category.

For the diffusion experiments, a separate dataset was used with three groups (20 images each):

- `Native_AI/`: images generated by Stable Diffusion v1.5
- `Cross_AI/`: images generated by other AI models
- `Real/`: real photographs

The images within the "Real" and "Cross_AI" were random images pulled from the available kaggle datasets.

IV. Methodology 1: Hybrid Spatial-Frequency Analysis

The spatial branch uses a pretrained ResNet50 backbone to extract the semantic and textural features, such as edges and patterns. This branch was implemented in an attempt to capture visual differences between real and generated images.

The frequency branch computed several FFT descriptors to capture recurring artifacts, grid patterns (from the generation process), and unnatural regularity.

The FFT spectrum on the entire image would result in 262k values (512x512), which is far too many to feed into the final neural network, so we compute 6 statistical features from the spectrum:

The frequency branch returns 6 features which are mapped through a MLP (multi-layer perceptron). ResNet50 returns a 2048 dimensional feature vector, which is then embedded and concatenated into a 2176 dimensional representation. This vector is passed through a fully connected classification head [2176 $\rightarrow$ 512 $\rightarrow$ 128 $\rightarrow$ 2] with ReLU and dropout, with two discrete outputs [Real, AI].

TABLE III
FFT Statistical Features

Feature	Explanation
Mean Magnitude	Average frequency strength *Higher for images with more structure* *Lower for simple images*
STD Magnitude	Variation in frequency distribution *Higher for complex images* *Lower for simple images*
High-Frequency Energy	Sum of the magnitudes in the outer regions *Real Images: Higher (due to sensor noise and finer textures)* *AI Images: Lower*
Frequency Ratio	High-Frequency Energy/Low-Frequency Energy *Measures distribution*
Spectral Entropy	Randomness of frequency distribution *Real images: Higher entropy (more random)* *AI Images: Lower entropy (structure within the image)*
Radial Variance	Variance of radial frequency profile *Compute average magnitude at each distance from the center* *To look for "grid artifacts" from diffusion*

Final Version: Images at 256 x 256

This was the best-performing classification model.

TABLE IV
Best-Performing Classification Model Results

Metric	Value
Test accuracy	78.89%
Precision	89.13%
Recall	52.56%
F1-score	66.13%

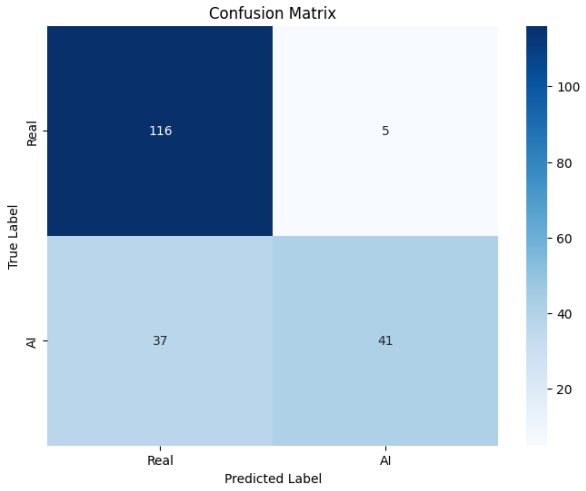

Fig. 1. Best-performing hybrid spatial-frequency classifier results

The model performed well, and achieve strong precision but moderate recall, as it missed nearly half of the AI images.

We can also examine the FFT plots across predictions.

Real Images:

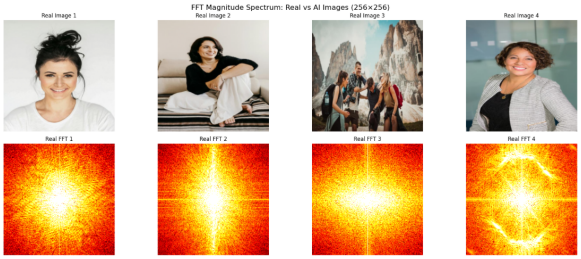

Fig. 2. Example FFT plots for real images.

AI Images:

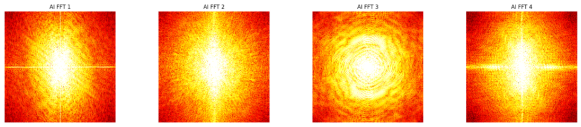

Fig. 3. Example FFT plots for AI-generated images.

TABLE V
How to Read FFT Images

Pattern	Meaning
Bright Center	Low frequencies [Large smooth structures and illumination in the image]
Outer dark regions	High frequencies [fine texture and detail]
Horizontal/Vertical lines	Strong edges or directional patterns
Rings or bands	periodic structures

In just observing the images, you can note that in "Real FFT1", the plot is kind of a blur, with the outer "orangish" color fading into the rest of the plot, whilst in "AI FFT1" you immediately notice very sharp vertical/horizontal lines and the bright center doesn't diffuse into the rest of the plot as prominently.

In “Real FFT3” there is a vertical line, however it is not as pronounced and is fading outwards, whilst in “AI FFT3” we notice a very clear “Ring” pattern in the plot.

Essentially, in real images and ai images, there are similar patterns, however the patterns within AI images are far more pronounced and explicit (Soft vertical/horzontal lines in “Real FFT2” and strongly defined vertical/horizontal lines in “AI FFT1”.

(Images are compressed, so it may be less visually clear, but still observable)

V. Methodology 2: Diffusion-Based Reconstruction

The intuition is that a model such as Stable Diffusion should be able to recreate images that are close to its learned distribution. So we purposely apply gaussian noise to the top-left quadrant of both images, then have the model denoise both images. We tested three on three different situations: images created by the same model, images created by other models (cross-validation), and real images.

Instead of a final prediction neural network, we analyze the reconstruction quality using metrics and rankings across the image types.

- **Stable Diffusion v1.5**
- **VAE** for image-to-latent encoding and decoding
- **UNet2DConditionModel** for denoising
- **CLIP tokenizer and text encoder**
- **DDIMScheduler** for timestep control

A. Metrics

TABLE VI
Metrics

Metric	Description
SSIM	Measures how structurally similar two images are Instead of comparing pixels, it compares patches [luminance, contrast, structure] Score is between 0-1, where 1 means identical structure
LPIPS	Measure perceptual similarity, how similar two images “look” Two images can have small pixel differences, but look entirely different, this method tries to capture human visual perception Images are passed to a pretrained N, where features are then extracted and the distance between those features are calculated. A lower distance means the images are perceptually similar

B. Attempt 1: Full Image Denoising

In the first attempt, noise was added to the top-left quadrant, but the entire 512x512 image was then encoded and denoised. Where the reconstruction quality was measured on the full image reconstruction. The noisy images were injected at timestep 30 of the diffusion process [if there were 50 iterations for a typical diffusion process] meaning it did 20 iterations to denoise the image.

Results:

TABLE VII
Attempt 1 Results

Image Type	SSIM
Native AI images	about 0.47 to 0.60
Cross-model AI images	about 0.49 to 0.56
Real images	about 0.29 to 0.35

Native AI reconstructed the best, Cross_model came second, and real images were reconstructed the worst.

Native_AI Reconstruction:

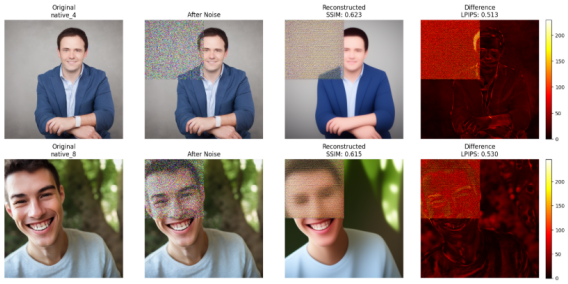

Fig. 4. Native AI reconstruction examples.

Cross_AI Reconstruction:

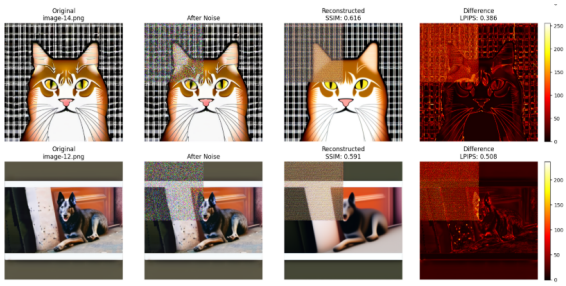

Fig. 5. Cross-model AI reconstruction examples.

Real Reconstruction:

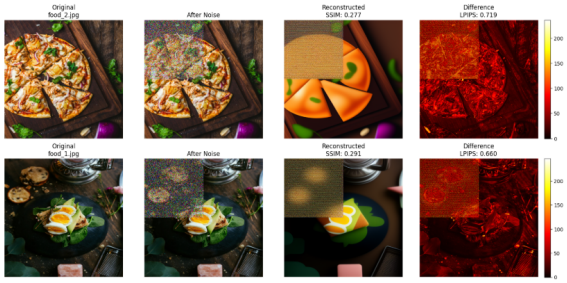

Fig. 6. Real-image reconstruction examples.

These are some of the reconstructions from each type of image. This directly supports the hypothesis that global reconstruction quality reflects the difference between statistical and physical relationships within images.

One of the noticeable differences, is that within the AI_generated images, the model is able to understand the meaning of the pixels, relating the final image to represent the same topic, In contrast, while trying to denoise the real images, it does not understand the content of the images, but instead notices the dominating colors and patterns, interpreting

the finer details as “noise” and in the process of denoising, simplifies the images [Pizza with toppings become a yellow circle with green, while the dog besides the door still clearly represents a dog besides a door].

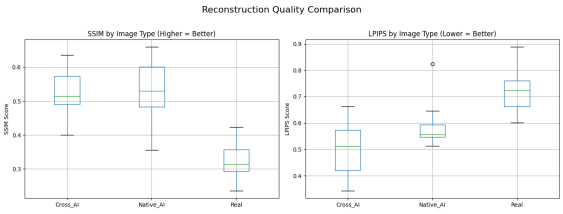

Fig. 7. Box plot of reconstruction metric distributions for Attempt 1.

Above depicts a box plot of the distribution of the reconstructions, the AI reconstruction distributions overlap substantially, while the real image distribution is more separated and includes outliers.

C. Attempt 2: Selective Denoising

In the second iteration, we again have the model denoise the entire image, but only the noisy quadrant from the denoised output was overlaid onto the original clean image. The idea is to evaluate the model on its ability to recreate only the section that was polluted. Similarity metrics were only conducted on the noisy portion.

Results:

TABLE VIII
ATTEMPT 2 RESULTS

Image Type	SSIM
Cross-model AI images	about 0.76 to 0.79
Native AI images	about 0.75
Real images	about 0.75

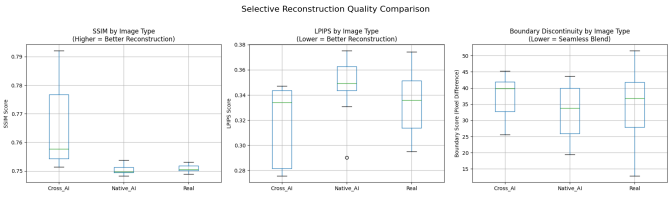

Fig. 8. Selective denoising results for Attempt 2

This diffusion attempt failed as a detection strategy, even though the absolute reconstruction quality was higher. The reason for these results is due to the task becoming too easy and far too local, as 75% of the image remained clean throughout the process, the model had strong guidance during denoising.

VI. Findings

Together these two methodologies support the claim that AI generated images do contain detectable signals, however those signals are sensitive to scale, framing, and implementation.

The hybrid classifier showed that supervised detection is a possible solution, reaching 78.89% test accuracy and 89.13% precision. Where further improvements to the architecture, such as testing various kernel sizes, and number of convolutional layers.

The diffusion reconstruction showed a conceptually important result, when evaluated globally, the reconstruction can be distinguished between Native AI, Cross Model AI, and real photographs.

Throughout this project, I had also tested several different variations, where several patterns emerged:

- **Simpler Baselines** often generalized better (especially with such a small dataset)
- **Global Analysis** was more informative than local patch analysis
- **Resolution** heavily affected the usefulness of FFT features and Spatial Detection

As we have more real photos than AI images, I had implemented class balancing by weighting the images’ influence on the training differently to balance the effect on training. However, this resulted in worse performance as the weighting was too aggressive.

Overfitting and generalization are the main barrier to stronger performance.

VII. Limitations and Further Improvements

- The frequency based classifier still has a low AI recall at 52.56%, as many AI images are being misclassified as being real.
- The dataset is still very small for a deep learning approach, this had contributed to the rapid overfitting I had noticed in my tests, as a result, I further simplified my approach to ease this effect.
- The current FFT features are still very broad and simple, and may not be the best at capturing artifacts and periodic structures.
- The diffusion model only evaluated Stable Diffusion on 20 and 50 iterations, the effect that a different number of iterations may have on the resulting image should be studied.
- The images within the dataset consisted of different sizes (which were normalized) meaning that information was lost when resized or low quality images were also present.
- Currently, I pass an empty prompt to the diffusion model when denoising, it should also be analyzed when the subject of the image is provided to the model.
- Testing of additional backbones such as ResNet18, EfficientNet, etc, should be investigated.
- Currently, in the FFT, the image is grayscale, and FFT is performed on a single channel, exploration of multi-scale frequency analysis would be a great avenue to research.
- For the diffusion I only apply a gaussian noise value of 0.6, different noise strengths should be tested, along with experimenting with different regions where noise should also be present.