Following the Breadcrumbs: Exploiting Artifacts Left Behind in the Frequency Domain to Detect Artificially Generated Faces

Jayanth Rao,[1] **Joshua Alfred Jayapal,**[1] **Adel del Valle**[1]

[1]New York University – Tandon School of Engineering
jr6594@nyu.edu, jj3811@nyu.edu, ad7082@nyu.edu

Abstract

The proliferation of generative adversarial networks (GANs) has revolutionized the field of image synthesis, enabling the creation of hyper-realistic deepfakes. While this progress introduces valuable applications, it also amplifies the risks of misinformation, fraud, and the erosion of trust in visual media. Existing detection methods relying on spatial domain features often falter against the sophistication of modern generative models. In this work, we chart an untraveled path, exploring the latent footprints left behind in the frequency domain by GAN-based upsampling mechanisms. Specifically, we propose a novel detection framework that leverages discrete cosine transform (DCT) preprocessing to amplify spectral artifacts inherent to deepfake generation. Our lightweight convolutional neural network (CNN), optimized for frequency domain features, achieves state-of-the-art performance with reduced computational overhead, surpassing traditional spatial-domain detection techniques. Comprehensive evaluations on a curated dataset of real and GAN-generated images demonstrate a significant improvement in detection accuracy and robustness across varying architectures. Our proposed network performs with a testing accuracy of 95.43% and an inference time of 2.5ms with the spectral image dataset. By unearthing the overlooked potential of frequency analysis, our method represents a paradigm shift in the fight against synthetic media proliferation, paving the way for future innovations in secure and trustworthy AI.

Introduction

Over the past half-decade, the rise and proliferation of generative adversarial networks (GANs) has raised significant concerns in the domains of security, the integrity in media, and digital forensics. This rapid technological innovation has led to a prevalence of artificially generated images, notably deepfakes, in our information-rich social landscape. Studies have shown how deepfakes erode trust in visual media; a notable example is the military coup in Gabon, inspired by a claim that the video of Gabon's president was a deepfake and in reality, the president was already dead (Hao 2019). Deepfake generation technology, democratized by tools like NVIDIA's **StyleGAN** (Karras, Laine, and Aila 2019), offer numerous educational and informational utilities, but also pose a threat if misused. Deepfake images can easily sow chaos and distrust, leading to the spread of misinformation and fraud. Detecting these fraudulent images has become a leading field of study and innovation to match the pace of improvement in deepfake technology.

Traditional deepfake detection methods rely on noticing unnatural developments in the spatial domain, such as inconsistencies in lighting, facial expressions, or strange visual artifacts introduced during the generation process. However, these approaches appear to struggle when generalizing their detection methodology across various GANs. As performance of these generative models continues to improve, it clearly highlights the need for alternative techniques to reliably detect deepfake images.

One such technique is to perform an analysis on the spectral representation of the input image, also known as **frequency domain analysis**. As opposed to spatial domain analysis, the frequency domain can expose information about the overall structure of the image, and can reveal artifacts introduced during the image synthesis process that may not be visible in the spatial domain. These artifacts appear during the upsampling process as GANs struggle to replicate the fine details in the images upon which they are trained. The existence of these artifacts lends credibility to the utilization of frequency domain analysis to detect deepfakes.

In this paper, we present a novel approach to deepfake detection that leverages frequency domain preprocessing combined with a shallow-architecture convolutional neural network (CNN). We extract the features from the frequency domain using a discrete cosine transformation, which we then use to train the CNN to accurately recognize real versus fake images. We compare our results to a traditional CNN trained on spatial domain features of the images in our dataset, and show how our results can underline how frequency domain analysis can overcome the limitations of detecting deepfake images when using traditional detection methods.

Ultimately, the work we accomplish in this paper is threefold:

1. We introduce a frequency domain preprocessing pipeline to detect deepfakes by transforming images into a spectral representation using a discrete cosine transformation.
2. We develop a shallow CNN architecture tailored to identify frequency domain features and classify images of faces as real or fake.
3. We evaluate our proposed detection model against the current state-of-the-art spatial domain detection algo-

rithms, to show the efficacy of a lighter-weight model.

Dataset

The dataset we utilized for this project was manually constructed by generating artificial fakes using NVIDIA's StyleGAN3 model, paired with real faces extracted from the Flickr-Faces-HQ (FFHQ) dataset (NVIDIA Corporation 2019). We assembled 40,000 of each type to create a balanced dataset of 80,000 total images. Note the distribution of our training, validation, and testing subsets in **Table 1**.

For the sake of minimizing storage overhead, we opted to render these images in a 256x256 resolution. The fake faces were generated using the pretrained `stylegan3-r-ffhqu-256x256.pkl` model (NVIDIA Corporation 2021), and the real faces from FFHQ were downscaled from the original 1024x1024 resolution.

Finally, as part of preprocessing the data, we performed a discrete cosine transformation on all of the images in the dataset.

Split	Real (FFHQ)	Fake (StyleGAN3)
Train	28,004	27,996
Val	8,022	7,978
Test	3,974	4,026

Table 1: Data distribution of the developed dataset.

Related Work

In this section, we present a selection of papers that discuss the limitations and shortcomings of how GANs generate images with accurate spectral distributions. Prior studies focus on detecting inconsistencies in the image by analyzing the spatial domain, such as visible artifacting, unnatural lighting, or irregular textures. While effective, these methods lack a general predictive robustness that can be applied to multiple GAN architectures. Durall et al. (2020) extend the artificial image detection space by investigating the frequency domain, noting that spectral distortions appear to be caused by the upsampling methods implemented by GANs. Their work shows a near-perfect accuracy on detection benchmarks, showing the potential of leveraging these frequency-domain anomalies.

Similarly, Corvi et al. (2022) investigate the artifacts left behind by diffusion models (DMs), and how they compare to GANs. They show how DMs demonstrate different spatial forensic traces, but similar artifacts in the frequency domain compared to GANs. However, they note that more mature DMs, such as Stable Diffusion and DALL-E 2, produce weaker spectral artifacts, complicating detection efforts. The authors note the techniques used by these developed DMs, such as regularization and compression during post-processing, that can effectively fool state-of-the-art detection efforts. The work underscores the need for more robust detection methods, lending credence to the inclusion of frequency domain analysis in detection efforts.

Martin-Rodriguez et al. (2023) investigate AI-generated image detection by leveraging two pixel-wise feature extraction techniques: Photo Response Non-Uniformity (PRNU) and Error Level Analysis (ELA). PRNU determines the specific noise patterns formed by physical camera sensors, which are absent in photo-realistic AI-generated images. ELA finds areas of images that may have been manipulated or edited by analyzing the compression levels compared to the rest of the image. These features are then fed into a CNN, which provides an extremely high accuracy rate, and a robustness towards a variety of GAN models.

Finally, we examine the work done by Amoroso et al. (2024), who focus on distinguishing multimodal deepfake images generated by text-to-image DMs. They propose a framework built upon perceptual and semantic components to discern synthetic images. The authors also discuss the utilization of higher-level semantic information to help classify authenticity paired with lower-level artifacts from the frequency domain; this research complements the prior literature by addressing challenges specific to multimodal and text-driven generative models, and how frequency-domain analysis can help innovate the deepfake detection space.

Methodology

Most research works focused on extracting features from images, to find subtle differences in the spatial and luminance changes within neighboring pixels of images to detect deepfake content. Although they provided promising results, there are significant differences between a fake and real face image within the frequency domain. We intend to leverage these features or noisy artifacts in the frequency spectrum to make efficient deepfake detections. Also, our results proved that deepfake detection in the frequency domain is lightweight, requiring less compute, memory, and inference time. Fig. 2 details the architecture of our proposed convolutional neural network used for this purpose.

Our proposed network contains 3 blocks of jointed convolutional, batch normalization and max pooling layers. The convolutional layers are solely responsible in feature extraction and learning in our model, through patchwise convolution with different filters on the input images. The filter sizes of the convolutional layers gradually reduce from (7, 7) to (3, 3) so that the network captures the large-scaled features in the spectrograms first before drilling down to the low-level granular features. Batch normalization sets all the pixels in the feature map output to a new mean and standard deviation (using z-score and learnable arbitrary parameters). Batch normalization will help to make the updates independent of the current batch's distribution properties. In other words, learned features are not skewed in one direction, but rather presented in a uniform scale for the next layer. MaxPooling at the end of every two convolutional layers would help to reduce the feature map size while keeping only the relevant, highly-weighted information within the map.

Then, the feature maps are transformed into a 1D representation and feed through a series of dense layers and dropout layers. Dropout is a regularization technique to prevent overfitting and to enhance the model's generalizing ability. By setting a proportion of neurons in the layer (given by the dropout ratio of 0.5) to zero, the model trains on a partially connected network and learns features. This process is iteratively done on multiple combinations of dropped

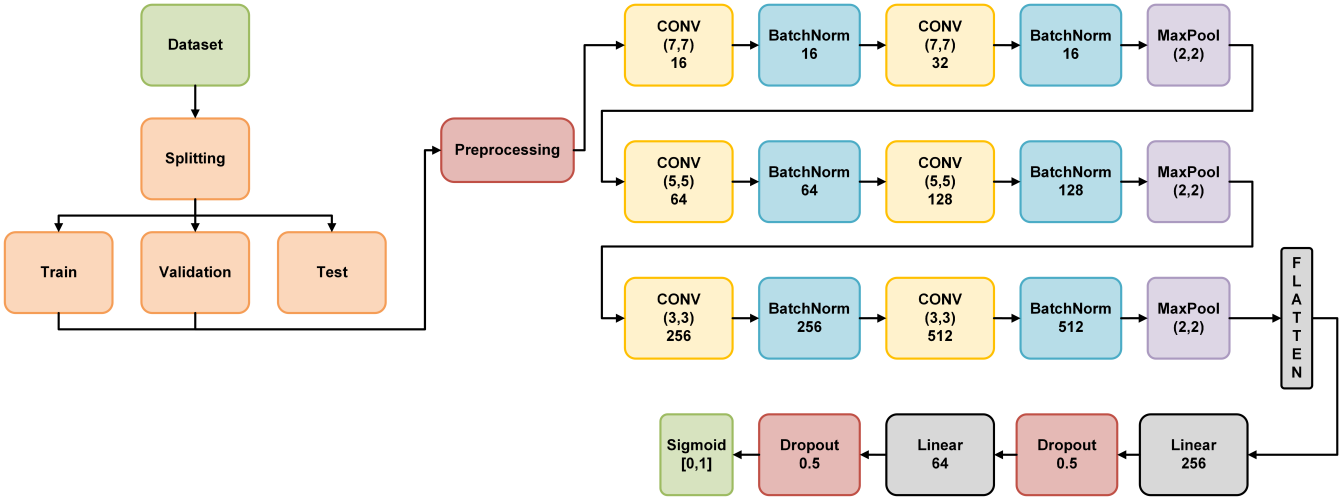

Figure 2: Architecture of the proposed network

neurons and the result is an ensemble of the learned results of these sparsely connected networks, thus improving the model's understanding on the whole, rather than relying on certain neurons which can overfit during testing/production.

After model construction, the architecture had 2.95M total trainable parameters, bearing a total size of 44MB, which is fairly lightweight and accessible on edge devices with less memory and storage capabilities.

We used this proposed architecture to train our data collection of DCT spectral samples of real and fake faces, as well as the original image samples to demonstrate how CNNs can easily capture frequency artifacts for efficient deepfake prediction, than being trained on the original facial images. Initial preprocessing to both the datasets were resizing to (256, 256) and pixel-wise standardization, and were converted to tensors before being fed for model training.

Results & Discussion

The proposed experiments were executed utilizing PyTorch, an open-source framework based on Python, and the Torch library. All experiments were carried out on Google Colab, which provided NVIDIA A100 GPU for training the deep learning network.

Our experiments include training the proposed network on two different datasets: collection of DCT spectral representations of real and fake face images, and a collection of the original images, in their RGB format.

The proposed network was trained after a performance analysis with its hyperparameters using grid search. The search space included (1) learning rate (LR), (2) weight decay (WD), and (3) gradient update optimizer algorithm as parameters of interest to optimize. The learning rate and weight decay parameters ranged from 1e-3 to 2e-5, and gradient optimizers, namely Adam (with weight decay), NAdam, and Adamax. Table 2 describes the best set of parameters after the search, that gave the efficient forthcoming results. After a number of experiments on a smaller subset of the spectral image dataset, Adamax optimizer with LR = 2e-5 and WD = 1e-5 gave the best results. The batch size was set as 32 in all experiments. This configuration was utilized for the model training experiments mentioned below.

#	LR	WD	Optimizer	Val accuracy
1	1e-3	1e-3	AdamW	0.4970
2	1e-4	1e-4	AdamW	0.5014
3	1e-3	1e-3	NAdam	0.4977
4	1e-4	1e-4	NAdam	0.9056
5	2e-5	1e-5	AdamW	0.9347
6	2e-5	1e-5	Adamax	**0.9567**

Table 2: Hyperparameter tests & analysis

The model training was conducted for 50 epochs each, on the aforementioned system configurations. The experimental results for both data sets are presented in Table 3. It was evident that convergence was efficient with the DCT spectral images, as these images are less complex, and carry visible artifacts for the model to learn. After 50 epochs and around 150 mins of training time, the model produced a testing accuracy of 0.9543. Fig. 5 displays the accuracy and loss variations over epochs for the spectral-trained model, and we can infer that the model has not overfitted the samples. Fig.6 shows the confusion matrix distribution for the same, and Fig.7 displays the ROC curve of the model on the testing set.

Comparing these results with the CNN built on RGB images (please find the respective visualizations in the notebooks), we can infer that the model built on DCT spectrograms is efficient and takes a lesser inference time per image (as seen in Table 3). The faster training and inference time can be explained by the nature of the image input being sent. Spectrograms are 1-dimensional, and processing them is much easier and allows faster convergence with a decent hyperparameter configuration and epoch count. On the other hand, spectrograms have proved to leverage enough information to the model through feature maps, leading to im-

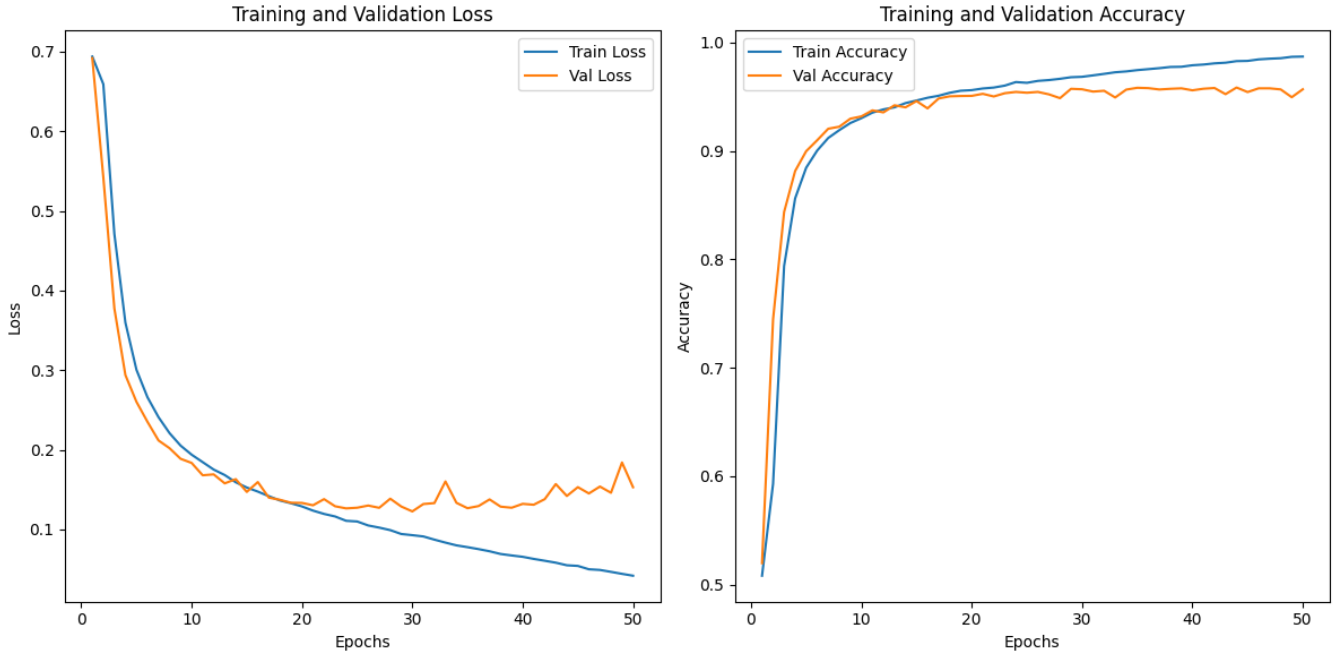

Figure 5: Proposed Network - Accuracy and loss variations over epochs

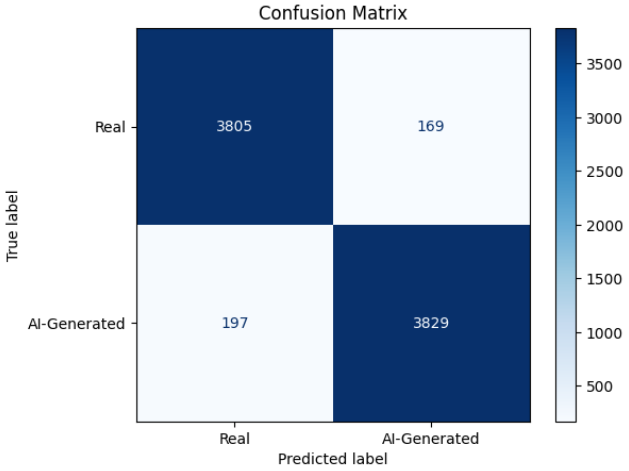

Figure 6: Proposed Network - Confusion matrix

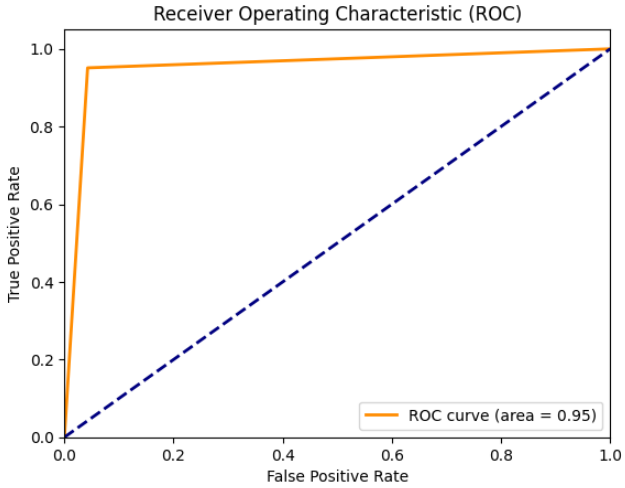

Figure 7: Proposed Network - ROC curve

proved generalization. The other performance metrics such as Precision, Recall, Area under the curve, and F1 scores are improved with the model trained on DCT images, and thus this model proves to be better with reducing Type I and Type II errors.

Our experiments reveal that spectral artifacts are a consistent fingerprint that can be systematically exploited. Through comprehensive evaluations on real and spectral-transformed datasets, our method achieved remarkable precision, recall, and F1 scores, outperforming state-of-the-art spatial domain approaches while maintaining a lightweight architecture suitable for resource-constrained environments.

Metric	RGB images	DCT spectral images
Precision	0.8026	0.9577
Recall	0.8371	0.9511
F1	0.8194	0.9544
AUC	0.8142	0.9543
Train Acc	0.8390	0.9869
Val Acc	0.8377	0.9567
Test Acc	0.8144	**0.9543**
Training time (min)	~ 250	~ 150
Inference time (ms)	4.562	**2.5**

Table 3: Performance metrics analysis of our method on the original image dataset vs spectral image dataset

Conclusion

In this work, we introduced a novel framework for detecting artificially generated faces by exploiting the spectral artifacts left behind in the frequency domain. By employing discrete cosine transform (DCT) preprocessing and designing a lightweight CNN tailored to frequency domain features, we demonstrated the efficacy of our approach in achieving superior detection accuracy with reduced computational overhead. Our findings underscore the limitations of traditional spatial-domain methods, which often fail to generalize across diverse GAN architectures, and highlight the resilience and efficiency of frequency-domain analysis.

This work bridges the gap between theoretical insights into frequency-domain representations and their practical applications in digital forensics. The lightweight nature of our approach opens avenues for deployment in real-time systems, such as social media platforms and content verification pipelines, where rapid and reliable detection of deepfakes is critical. Furthermore, our method's robustness against diverse GAN models provides a strong foundation for future extensions, including adaptive detection mechanisms and cross-modal analysis of synthetic media. Our experiments show the spectral CNN model's performance exceeded that of the CNN built upon their original RGB representations. Also, the reduced dimensionality of DCT images has improved the training and inference time for the model.

We envision this research as a stepping stone toward a new era of secure and trustworthy AI, where frequency domain analysis plays a pivotal role in safeguarding digital ecosystems. Future work will explore integrating learnable frequency-domain transformations, expanding the approach to video deepfakes, and enhancing interpretability to ensure transparency and trust in detection systems. Together, these efforts will solidify the defense against the evolving threat of synthetic media, ensuring the integrity of digital content in an era of rapid technological advancement.

References

Amoroso, R.; Morelli, D.; Cornia, M.; Baraldi, L.; Bimbo, A. D.; and Cucchiara, R. 2024. Parents and Children: Distinguishing Multimodal DeepFakes from Natural Images. arXiv:2304.00500.

Corvi, R.; Cozzolino, D.; Zingarini, G.; Poggi, G.; Nagano, K.; and Verdoliva, L. 2022. On the Detection of Synthetic Images Generated by Diffusion Models. *arXiv preprint arXiv:2211.00680*.

Durall, R.; Keuper, M.; and Keuper, J. 2020. Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks Are Failing to Reproduce Spectral Distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7890–7899.

Hao, K. 2019. The biggest threat of deepfakes isn't the deepfakes themselves. Accessed: 2024-12-18.

Karras, T.; Laine, S.; and Aila, T. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4396–4405.

Martin-Rodriguez, F.; Garcia-Mojon, R.; and Fernandez-Barciela, M. 2023. Detection of AI-Created Images Using Pixel-Wise Feature Extraction and Convolutional Neural Networks. *Sensors*, 23(22): 9037.

NVIDIA Corporation. 2019. Flickr-Faces-HQ (FFHQ) Dataset. Accessed: 2024-12-18.

NVIDIA Corporation. 2021. Alias-Free Generative Adversarial Networks (StyleGAN3). Accessed: 2024-12-18.

Appendix

- Project Notebook - Model on DCT images
- Project Notebook - Model on RGB images
- Proposed model - DCT images
- Proposed model - RGB images
- GitHub Repo link