Analyzing Performance: A Study on Factors Affecting Under-performance of Diffusion Generated Image Detection Models

Eunbeen Hong
Dept. of Computer Science and Engineering
College of Informatics
Korea University
2021320120

Hyejin Jo
Dept. of Computer Science and Engineering
College of Informatics
Korea University
2020320159

Jeemin Oh
Dept. of Computer Science and Engineering
College of Informatics
Korea University
2021320150

Jiyoung Kim
Dept. of Computer Science and Engineering
College of Informatics
Korea University
2021320123

***Abstract*—Diffusion models in the field of generative AI have been extremely successful in creating high-quality images for a variety of tasks. However, synthetic images that are nearly indistinguishable from real images can pose a great threat to society, particularly in the hands of those with malicious intent. Thus, it is important to find a reliable method to differentiate between machine-generated synthetic images and real images. To this end, we analyze different models and factors that contribute to successfully detecting synthetic diffusion-based images and compare the effectiveness of using PRNU pattern information to other proposed techniques. In this paper, we will discuss our model architecture, compare its performance to existing models, and provide an analysis of our results.**

I. Introduction

Diffusion models have been incredibly successful at synthesizing images for a variety of tasks. These synthesized images can be created to such a high quality that it is indistinguishable from "real" images to the human eye. While these breakthroughs have great implications for the future of many fields, it also creates security concerns. Malicious users can use these tools to propagate the dissemination of fake information, or fool victims with high-quality deepfakes [1]. Thus, with the development of image synthesis models, it is critical that we also secure a reliable method of differentiation and verification.

However, the task of detecting synthetic images is extremely difficult, especially due to the rapid advances in generative technologies. Specifically, because different generative techniques may leave behind vastly different fingerprints or residuals and more recent models have been able to successfully cut down on the significant residual left on synthesized images, it is particularly difficult to design a generalized model for detecting synthesized images.

There has been recent active research on the weaknesses of diffusion-generated images. Particularly, it has been found in [6] that due to the lack of explicit 3D modeling, there can be inconsistencies in shadows or reflection symmetries. Previous works [3] have also established that many generative models leave behind fingerprints, which can be extracted through a denoising process. Spatial domain fingerprints, as well as spectral peaks in frequency analysis, have been useful for detecting GAN-based synthetic images, but have yet to be used successfully in detecting diffusion-based synthetic images.

In our paper, we will analyze the effectiveness of various known architectures in detecting diffusion-generated synthesized images, then compare the use of noise fingerprints from PRNU signals to existing models and analyze the effectiveness of our technique. Lastly, we will conclude with our findings and suggest the most effective architectures for this task.

II. Background

We will briefly introduce the topics of diffusion-based image synthesis, synthetic image detection techniques, and frequency signal processing.

A. Diffusion-Based Image Generation

Diffusion models have been a recent breakthrough in the field of image synthesis. Furthermore, Diffusion Probabilistic Models [9] have also gained attention in the field. By iteratively approximating the true data distribution, diffusion probabilistic models offer a flexible framework for generating images with controllable attributes, such as style and content.

B. Generated Image Detection

There have been numerous recent studies to identify differentiating characteristics between real and computer synthesized images that have revealed key insights.

Corvi, in [3] found that similar to GAN-based images, images created through diffusion based means also leave behind fingerprint traces. These traces can be recovered using denoising techniques similar to the ones used for device identification using PRNU signals. Another central discovery made in this paper is the necessity of preserving information through the avoidance of heavy pre-processing of data, which may erase evidence and fingerprints needed for detection. Most of all, it is critical to avoid resizing images, especially in the first few layers, as this requires resampling and interpolation, which would erase much of the frequency information needed. To preserve these artifacts, Corvi suggested local patches from image crops and late fusion strategies to minimize information loss.

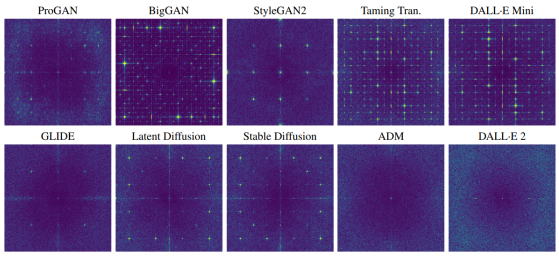

Fig. 1. Fourier transform (amplitude) of the artificial fingerprint estimated from 1000 image residuals, as shown in Corvi, 2022

Moreover, in [2], Chai finds that while global errors may vary from generative model to model, local errors may transfer more. Thus the focus should be on textures, and more detailed frequency information. Specifically, they find that on facial images, patch-based classifiers often perform better on out-of-domain synthetic images than full-image classifiers. We hypothesize that this may apply similarly to other classes of synthetic images as well.

Lastly, DIRE [14] proposes a novel image representation called "difference in residual error", which measures the difference between an input image and its reconstructed counterpart after deconstructing and subsequently reconstructing the image through the diffusion process. This method seems to have been the most successful thus far, showing nearly perfect accuracy and average precision on ADM, PNDM, iDDPM, Stable Diffusion, and LD models using the LSUN-Bedroom dataset.

However, in the works seen so far, results are presented only in ideal conditions and lack consideration for robustness analysis. Moreover, the dataset is rather limited and may even have pre-existing differences in data distribution, which may produce falsely over-encouraging results.

C. Frequency Signal Processing for Generated Image Detection

In our paper, we attempt to approach the task of detection with signal processing and denoising techniques, mainly focusing on the application of PRNU signal extraction and frequency analysis. PRNU (Photo Response Non-Uniformity) fingerprint extraction is a valuable technique used to detect computer synthesized images. PRNU fingerprints are unique noise patterns that arise from the imperfections of camera sensors and are introduced during the manufacturing process. When an image is captured by a camera, these fingerprints manifest as subtle pixel variations. However, computer synthesized images lack these inherent fingerprints, as they are typically created digitally without passing through a camera sensor. By extracting the PRNU fingerprints from an image and comparing them to a reference database of known camera fingerprints, it becomes possible to distinguish between genuine photographs and computer-generated images. The use of PRNU fingerprint extraction provides a powerful tool in combating the increasing sophistication of computer-generated visual content.

III. Approach

A. Motivation and overview

Previous paper [3] suggested that using a denoising filter such as PRNU extractor can be helpful to distinguish real and fake images, which has been proven to be successful for camera fingerprint extraction. Based on the findings obtained from the preliminary research and base paper [3], it was shown that the ViT and Swin models effectively detect images generated by the diffusion model. Each model demonstrated performance levels of 0.84 and 0.4 accuracy, respectively. In order to further enhance the performance of these two models, we decided to utilize them as backbones while incorporating additional feature extraction from the PRNU denoising filter. In the process of designing our model, we aimed to include three distinctive elements. Firstly, we proposed the removal of downsampling in the backbone stem layer. Secondly, while utilizing ViT as the backbone, we extracted feature information not from the class token but from the remaining output patches. Lastly, we introduced the process of late fusion, which involves integrating the feature information from the backbone model with the information obtained from the PRNU denoising filter.

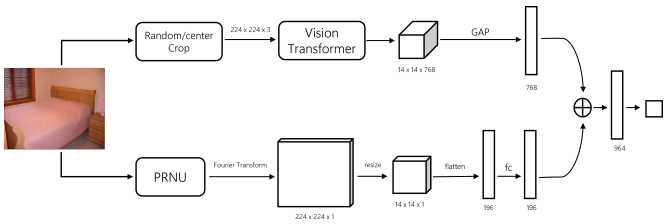

Fig. 2. Proposed model architecture with PRNU and FFT module

B. Removing downsampling method

The process of image resizing can have a detrimental impact on detection performance as also suggested in related GAN detection paper [7]. Particularly, our PRNU denoising filter is highly responsive to resolution, which makes it crucial to address spatial aliasing and checkerboard artifacts. In order

to solve these issues and avoid downsampling, we repeatedly apply crop transforms. In order to minimize the amount of quality loss, we used a random crop for the training and the center crop for the testing.

C. Extra feature token

In our approach, we assumed that the remaining output patches in the VIT model, excluding the class token, contain valuable additional evidence. Each patch in the ViT output represents a local region of the input image. These patches contain visual information specific to the corresponding region, capturing details and localized features that contribute to the overall understanding of the image. Also, the relative positions and spatial relationships between the patches carry important information about the structure and layout of objects in the image. The arrangement of patches can provide cues about the spatial organization of the visual elements. Also, the Fourier transform, which we use when extracting features from PRNU patterns, discards spatial information such as exact pixel positions and local relationships between image elements. The transformed representation focuses solely on the frequency content and does not retain explicit spatial localization, which highlights the importance of the extra features by leveraging the local or spatial details the images contain. Due to the reasons mentioned above, our model takes advantage of all the output patch tokens rather than utilizing the class token.

D. Late fusion strategy

Late fusion strategy refers to the combination or integration of multiple sources or modalities of information at a later stage in a data processing pipeline. It allows for the integration of different types of information which may have distinct representations or structures. The Fourier transform provides shift invariance, which means that a translation or shift of the image results in the same frequency representation with phase changes, which can be useful for certain tasks including detection. By combining frequency features and extra feature tokens at a later stage, the fusion process can handle the above heterogeneity more effectively. Furthermore, as mentioned earlier, frequency information inherently results in the loss of spatial details, making it unreliable. In cases where one or more modalities may be missing or contain inconsistent information, late fusion can provide robustness. The fusion process can adaptively weigh the available modalities, mitigating the impact of missing or noisy data on the overall performance.

IV. Experiments

A. Dataset

We first used LSUN-bedroom dataset as the real images and model-generated images trained to the LSUN-bedroom manifold as the synthetic images. This is because when we used a general dataset, it was ambiguous whether the model truly recognized fake images or only distinguished the difference in distribution of the dataset. We use DDPM [9], DDIM [12], guided-diffusion [4], latent-diffusion [11] as the model by which we generated the fake dataset. Then we extended it to general dataset.

To make our model robust, we apply data augmentation to the model, where we use Gaussian noise and Image Compression. However, as in table 2, the performance of the models with augmentation is subpar. This phenomenon was also suggested in [3] so we kept data augmentation to a minimum.

B. Setup and Training Details

We used Google Colaboratory to test all our models. We used a learning rate of 1e-4 in A100 GPU with Early Stopping with patience 10. We normalized the input image with the mean and standard deviation of the very model's training dataset. We padded the input image with zeros if the input image size is small for the model (e.g. ViT). We randomly cropped the input image for training and used the center crop for testing. Cross entropy function was used for loss, and Adam for the optimizer.

C. Results

We found the that the best model was EVA, with an accuracy of over 0.9 for detecting diffusion-generated images.

V. Analysis

A. Ablation Studies

We did several ablation studies using our own model (in Figure 1). The results are in Table 1. First, We changed the pretrained ViT backbone with extra feature tokens and global average pooling (GAP) to the cls tokens in our own model (2 track). The performance of the model used backbone with extra feature tokens is better (0.77) than cls tokens(0.58). However, when comparing vanilla ViT's performance (1 track), the cls token distinguished diffusion-generated images better (0.85) than extra feature tokens(0.78). The reason for this phenomenon is that while the 2-track model with late fusion requires additional feature information so it can be concatenated with PRNU signals, the single-track model only requires information about the classification. Thus it is better to use cls tokens that leverage this task. Compared to the result above, the performance of PRNU with extra feature tokens from ViT (0.5) did not show significant differences to using the cls tokens (0.5), both of which performed poorly. Because of this, we questioned the benefits of the PRNU patterns. So to test whether the PRNU patterns provide helpful information, we changed the model architecture, from 2 tracks to a single track, to take input which is original images or PRNU patterns with Fast Fourier Transform (FFT) and compared the performance. The standard ViT of the original images input detected diffusion-generated images better (0.85) than the PRNU signal input (0.5). We also used the signal of the original images computed with the FFT without the PRNU extraction, but the results are as bad as the PRNU signal.

Why don't PRNU signals work properly? This can be explained as a discrepancy in feature distribution. First, features from the regular backbone model and PRNU patterns seem to

Model	Accuracy
vanilla ViT-base with cls token	0.85
vanilla ViT-base with feature tokens	0.78
late fusion of PRNU and ViT-base feature tokens	0.77
late fusion of PRNU and ViT-base cls token	0.58
ResNet50 with PRNU and FFT	0.53
ResNet50 with PRNU	0.52
vanilla ResNet50	0.5
ViT-base with PRNU and FFT	0.5
ViT-base with PRNU	0.5
ViT-base with FFT	0.49

TABLE I
ABLATION STUDY

Dataset	Method	Accuracy	Data Aug
LSUN-B	ResNet50	0.49	O
	Swin-base	0.82	
	ViT-base	0.78	
	ViT-large	0.84	
	ViT-base	0.85	X
	ViT-large	0.84	
	Swin-base	0.84	
	MLPMixer	0.76	
	CNNDet [13]	0.87	
	Patchfor [2]	0.91	
	F3Net [10]	0.92	
	Grag2021 [7]	0.76	
	Eva-02 base	0.87	
	Eva-02-large	0.87	
	Eva giant (ours)	0.92	
General Dataset	ViT-base	0.85	
	Eva giant	0.87	

TABLE II
EXPLORING MODEL ARCHITECTURES

be incompatible. As described above, We tried several ways to use PRNU patterns or even only FFT trace with backbone features, but all of these trials ended up either overfitting or did not converge. When PRNU patterns are used, this hinders the training phase of the model as it has different distribution from features. As can be seen in Table 1, when PRNU or FFT is used, the result is always worse than vanilla models.

Second, there are fine-tuning challenges to the "already-knowing" model. We used the backbone such as ViT or Resnet only pretrained by ImageNet, not the signal. With this, it is hard to fine-tune the model with comparatively small (around 3k) PRNU data. PRNU has a significantly different distribution from real images, so if we want to successfully make our own detector, we may need to pre-train over 14M PRNU data (same as ImageNet) or fine-tune over 20k data, or it would have a hard time learning to interpret PRNU patterns.

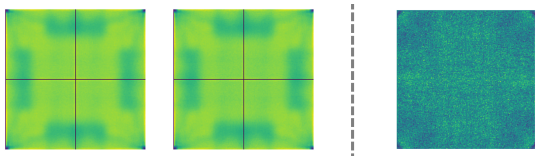

Fig. 3. Left are average PRNU patterns from real and fake respectively, and the right is the difference between the two

Third, PRNU patterns are indistinguishable. A critical drawback of PRNU is that PRNU patterns of real and fake images are indistinguishable. An average PRNU of each fake and real image is computed and visualized as on the left side of Figure 2. The two maps have little difference, and cannot be distinguished from the other. Further, as shown on the right side of 2, which is the subtraction of two patterns, any meaningful pattern has not been detected. As PRNU does not contain meaningful information, it would only confuse the classification. This could be due to an inept PRNU module. [3] has mentioned that PRNU has a meaningful pattern, but we suspect that this pattern (i.e. fingerprint) is different for the dataset is used or it has nothing to do with differentiating with the real image.

Additionally, as mentioned above, the cls token is better than extra feature tokens for classification tasks. We first assumed that if we used extra feature tokens, the model would be able to fuse the spatial information of semantic inconsistency of the image and some evidence from the PRNU pattern. However, as PRNU hardly has meaningful information, extra feature token is different from the output of traditional CNN and performs worse than the cls token.

Through this ablation study, we chose the architecture which is 1 track and takes original images as the input.

B. Exploring model architectures

Next, we explored various pretrained backbone and compared the performance. Through numerous attempts, we finally found the currently best performing model to be the EVA-giant (0.92). We predict the reason for this to be from the properties of EVA [5]. EVA is based on Vision Transformer, which is scaled up to 1B parameters and learned via self-supervised learning method using CLIP representation. Due to the large size of the model, it can extract overall levels of features, thus it can distinguish the synthetic images by the various aspect.

VI. CONCLUSION

We overviewed various models to classify diffusion-generated images from real images. We tried to use the PRNU pattern from the dataset but found that PRNU is hard to use with the feature space of the traditional model. To overcome this, we suggest trying to embed PRNU space with a bigger dataset. This is to ensure that model can learn to exploit and distinguish features from PRNU images. We also found a better model with over 90 percent accuracy on our dataset, which was the EVA model.

REFERENCES

[1] Amin Azmoodeh and Ali Dehghantanha. Deep fake detection, deterrence and response: Challenges and opportunities, 2022.

[2] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize, 2020.

[3] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models, 2022.
[4] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021.
[5] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale, 2022.
[6] Hany Farid. Lighting (in)consistency of paint by text, 2022.
[7] Diego Gragnaniello, Davide Cozzolino, Francesco Marra, Giovanni Poggi, and Luisa Verdoliva. Are gan generated images easy to detect? a critical analysis of the state-of-the-art. 2021.
[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.
[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.
[10] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues, 2020.
[11] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.
[12] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022.
[13] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. Cnn-generated images are surprisingly easy to spot... for now, 2020.
[14] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection, 2023.
[15] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.