2D-FACT: Dual-Domain Fake Image Detection Against Text-to-Image Generative Models

Eric Ji
University of Illinois Urbana-Champaign
Champaign, IL, USA
ericji3@illinois.edu

Boxiang Dong, Bharath Samanthula, Na Zhou
Montclair State University
Montclair, NJ, USA
{dongb, samanthulab, zhoun}@montclair.edu

***Abstract*—Recent developments in generative artificial intelligence are bringing great concerns for privacy, security and misinformation. Our work focuses on the detection of fake images generated by text-to-image models. We propose a dual-domain CNN-based classifier that utilizes image features in both the spatial and frequency domain. Through an extensive set of experiments, we demonstrate that the frequency domain features facilitate high accuracy, zero-transfer learning between different generative models, and faster convergence. To our best knowledge, this is the first effective detector against generative models that are finetuned for a specific subject.**

***Index Terms*—AI generative model, fake image detection, frequency domain**

I. Introduction

Recent advancements in generative artificial intelligence (GenAI) have transitioned our society into a new AI-facilitated era. Large language models (LLMs) like ChatGPT are being used in areas such as customer support, financial reporting, and code refactoring. From a multimedia perspective, other breakthroughs like Stable Diffusion [7] are capable of portraying photo-realistic images from text prompts with little computation costs. Additionally, with a few image samples of a subject, DreamBooth [8] can synthesize them into any scene, pose, or view.

Even though generative models significantly boost productivity across various industries, they also raise tremendous concerns for security and privacy. The FBI recently observed an uptick in the wrongful use of AI-generated content for harassment[1]. In July 2023, AI companies, including OpenAI, Google, and Meta made voluntary commitments to the White House to implement measures such as watermarking AI-generated content to help make the technology safer in the future[2]. However, research has shown that watermark-based detection is not enough [3]. These actions do not resolve the immediate need for a reliable tool to detect AI-generated content.

This paper focuses on the detection of fake images generated by text-to-image generative models, specifically diffusion models. There is very limited existing work in this field. [9] is the closest to ours and proposes a CNN-based classifier to attribute a fake image to a specific AIGen model with the help of a prompt. Therefore, it requires additional models for predicting captions. This increases both the computation power and space necessary. Also, no existing work targets detecting finetuned generative models like DreamBooth [8]. These types of GenAI models pose the highest security threats to the society, since it makes the synthesized image more trustworthy and disruptive by implanting a certain subject.

In this paper, we propose 2D-FACT, a dual-domain fake image detection model. We propose to build a CNN-based detection model that takes the spatial domain and frequency domain features of the images as input. To our best knowledge, this is the first effective fake image detector against finetuned generative models. In this paper, we make the following contributions.

- We build a dataset that includes fake images generated by diffusion models and finetuned models.
- We demonstrate the superiority of dual domain features in encapsulating patterns unique to the underlying AIGen models.
- We perform an extensive set of experiments to show the effectiveness of the model in detecting fake images.

The rest of the paper is organized as follows. Section II introduces preliminaries. Section III discusses 2D-FACT in details. Section IV presents the experiment results. Section V reviews related work. Section VI concludes the paper.

II. Preliminaries

A. Generative AI Models

In recent years, there has been a notable emergence of GenAI models. LLMs like GPT, BERT and Llama push the boundaries of natural language understanding and generation. They showcase the ability to generate coherent and contextually relevant text, leading to improvements in chatbots, virtual assistants and content creation. Even though effective, these LLMs are only limited to analyze and generate text content and thus have limited applications.

Another flavor of GenAI models focus on multimedia data and produces realistic artistic work. Stable Diffusion [7] is a text-to-image model based on a latent diffusion model. The latent diffusion model is a state-of-the-art probabilistic

[1]https://www.ic3.gov/Media/Y2023/PSA230605
[2]https://www.reuters.com/technology/openai-google-others-pledge/

979-8-3503-0965-2/23/$31.00 ©2023 IEEE

model capable of producing impressive results in multiple tasks such as image synthesis, super-resolution, inpainting, and colorization [7]. Unlike previous GAN-based image generation models [12], diffusion models do not require billions of parameters, making them relatively more accessible and less computationally intensive. Due to Stable Diffusion's strong capabilities and open-source nature, it has hatched many applications such as Draw Things[3] and DreamStudio[4].

DreamBooth [8] is a powerful extension of Stable Diffusion and brings content generation to a higher level. What DreamBooth brings to the table is this aspect of "personalization". By fine-tuning a pre-trained diffusion model with a few image samples of a subject along with a text prompt, it learns to implant the subject into the model's output domain. The applications of DreamBooth include recontextualization, art renditions, novel view synthesis, and property modifications [8]. The "personalization" aspect of DreamBooth is impressive but does pose large threats when used with the wrong intentions. As a result, our work highlights a model's robustness in detecting images generated by DreamBooth.

(a) A real image

(b) A fake image

Fig. 1: Example of recontextualization for Leonardo DiCaprio with the prompt "Zwx man in a prison cell"

B. Risks of Generative AI Models

A significant privacy concern associated with these large pre-trained models is the leakage of private information. Often the training datasets consist of images scraped from the web, which leads to potential violations of privacy rights and copyright infringement [1]. In reality, images of real people have appeared in AI-generated images without their knowledge. In supervised deep learning, the model repeatedly scrutinizes the training samples to learn the relationship between features and labels. Due to this nature, the trained model inevitably memorizes information from the training set. Member inference attacks [10] have proved to be effective in determining if a sample is included in the training set of a machine learning model.

[3]https://drawthings.ai/

[4]https://beta.dreamstudio.ai/generate

Another large concern that has lately been plaguing the internet is misinformation. Having text-to-image models in the hands of malicious actors can lead to the most convincing fake news, hoaxes, and harassment [1]. In Figure 1, we show a pair of real and fake images of Leonardo DiCaprio. As displayed in Figure 1 (b), with DreamBooth, we can fake a realistic image of him in prison. This can be especially dangerous if AI-generated content influences individuals to perform dangerous actions or crimes. Even in academia and industry, text-to-image models present considerable risks. Aside from diminishing learning experiences, utilizing AI-generated images may lead to the reinforcement of harmful stereotypes, toxicity, and hate present in datasets [1]. The area of copywriting around AI-generated content is also still unclear and may put individuals in jeopardy. All these risks mentioned only scratch the surface and further support the need for the detection of AI-generated images.

III. Method

In this section, we first formally define our problem in Section III-A, then introduce the datasets that we build in Section III-B. Finally we present the design of 2D-FACT in Section III-C.

A. Problem Definition

In this paper, we aim to build a detection model to accurately identify images produced by text-to-image models. The resulting product would function as an effective fort to tackle the security threats posed by the rapid advancements in GenAI models. Besides being capable of differentiating fake and real images, the model should also satisfy the following requirements.

- **Lightweight and efficient.** The detection model should not demand massive computing resources such as GPU clusters and should have a fast response time.
- **Agnostic to various GenAI models.** Considering the fast development of LLMs, there are many derivatives of GenAI models. The detection model should be robust to the adversary's choice of generative models.

Some existing work not only aim to detect the fake images produced by GenAI models but also intend to accurately attribute them to the specific underlying model [9]. This is out of the scope of our paper.

B. Datasets

To train and test our models, we produced two separate datasets. All images gathered have a size of 768 x 768 pixels with RGB channels. The datasets will be made public upon the publication of this paper.

1) Diffusion Dataset (where fake images are generated directly from Stable Diffusion): We generate this dataset by following a similar approach in [9]. We take 12,500 real images from MSCOCO [4]. Despite being originally used for captioning tasks, the image descriptions provided by MSCOCO served as excellent prompts for text-to-image models. We implemented a pre-trained Stable Diffusion 2.1

Fig. 2: Real image (left) and fake image (right) derived from "A boat traveling through the water towards a rocky shore." in the *Diffusion* dataset

model from Diffusers [11] to generate corresponding fake images. This process allowed us to acquire an even split of fake and real images stemming from 12,500 unique prompts. This dataset consists of 25,000 images, with 20,000 images for training and 5,000 images for testing. A pair of real and fake images in this dataset is displayed in Figure 2

(a) A real human image (b) A fake human image

(c) A real object image (d) A fake object image

Fig. 3: Sample images in the *Finetuned Dataset*

2) Finetuned Dataset (where fake images are generated by finetuned model): We collect a dataset of 20 subjects, including 10 celebrities and 5 objects. For human subjects, we scrape real images of celebrities from the internet, and for objects, we use product images from IKEA. To generate fake images, we use DreamBooth [8] on Stable Diffusion 2.1 [7]. Specifically, we fine-tune the DreamBooth model with 5-7 real images and generate fake ones with a variety of prompts to cover diverse settings. In Figure 6, we provide a pair of real and fake images for one human and one object.

C. Detection Model

Discrete Fourier Transform (DFT) transforms the signal or an image from its spatial domain representation to the frequency domain. It has been widely used in image processing for various purposes, including noise reduction, sharpening, blurring, and compression. In our study, we find the frequency components of an image carry patterns, textures, and anomalies that are not immediately apparent in the spatial domain. Therefore, we build a dual-domain detector to identify fake images generated by AIGen models. In particular, to convert images into the frequency domain, we first convert all images into a single grayscale channel, then apply the DFT followed by a shift. Since the frequency response of an image consists of an imaginary component, we concatenate the magnitude response with the spatial domain features to create a dual-domain input.

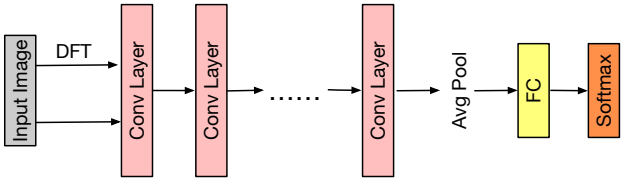


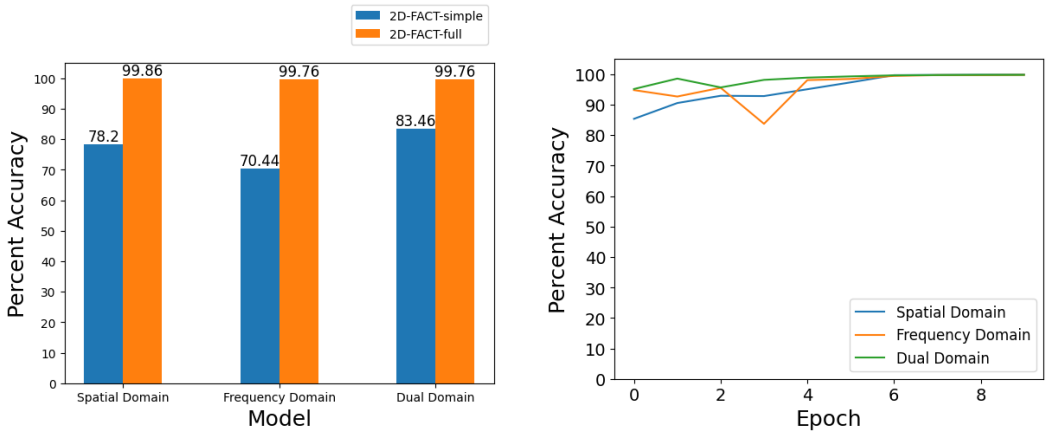
Fig. 4: The ResNet-18 architecture used for the dual-domain detector

For the detector, we trained ResNet-18 [2] from scratch. ResNet-18 is an 18-layer Convolutional Neural Network (CNN) widely known for its impressive performances in computer vision tasks like image classification and object detection. The learned detection model is named 2D-FACT-full. The architecture of ResNet-18 is displayed in Figure 4. We also train a simpler detector named 2D-FACT-simple by building a 7-layer CNN ($conv \rightarrow maxpool \rightarrow conv \rightarrow maxpool \rightarrow linear \rightarrow linear \rightarrow linear$).

IV. Experiment

A. Setup

In our experiments, we train the dual-domain detection model on the *Diffusion* dataset only. The *Finetuned* dataset is only used to evaluate model performance. This is to simulate the real-world setting where it is very difficult to collect a large set of fake images generated by finetuned models. This is because to generate such a fake image of a subject, we will need to first collect a few subject sample images taken at different angles and then finetune an existing text-to-image model. We train the detection model with stochastic gradient descent as our optimizer for 10 epochs with a batch size of 16. The training process is run on a computer with a 2.20 GHz Intel Xeon CPU and an NVIDIA T100 GPU. The 2D-FACT-full detector includes 11 million trainable parameters and takes about 150MB of memory. It is very affordable to store and run the model.



(a) Detection accuracy

(b) Testing accuracy of 2D-FACT-full at each epoch

Fig. 5: Detection accuracy of the fake images generated by diffusion model

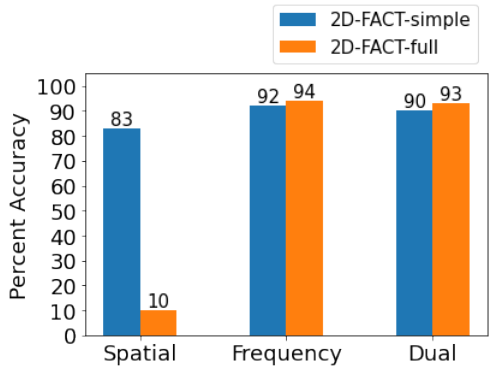


Fig. 6: Detection accuracy of the fake images generated by finetuned model

B. Detection of Diffusion Models

In Figure 5 (a), we show the detection accuracy on the *diffusion* dataset. We can see that 2D-FACT-full adequately identifies the fake images generated by Stable Diffusion, regardless of which domain the features are IN. On the other hand, 2D-FACT-simple also yields moderate detection accuracy (around 80%). This result demonstrates that 2D-FACT is fully capable of learning the patterns of fake image produced by Stable Diffusion. From Figure 5 (b), we can tell that the dual-domain features do help 2D-FACT-full to gain better accuracy in the first few epochs. However, even regardless of the frequency domain features, the detection models can converge to similar testing accuracy. We have similar observations from 2D-FACT-simple. We do not show it in the paper due to limited space. When supplementing the spatial domain features with frequency domain features, we do not yield significant boosts in overall performance. However, it does help with faster convergence, resulting in lower training costs.

C. Detection on Finetuned Models

In this section, we report our findings on using a model trained on Stable Diffusion to detect images generated by DreamBooth.

In Figure 6, we can see that 2D-FACT-full exhibits contrasting accuracy on the spatial domain and frequency/dual domain (i.e., 10% v.s. 93%), while the difference for 2D-FACT-simple is significantly smaller (i.e., 83% v.s. 90%). From this comparison, we can tell that: (1) 2D-FACT-full overfits the training set. However, the spatial domain features do not carry sufficient patterns of finetuned model, which leads to the catastrophic performance on the *Finetuned* dataset. Even a random guess would be able to outperform this model. (2) Frequency domain features encapsulate patterns unique to diffusion models and finetuned models, hence even a simple detection model (i.e., 2D-FACT-simple) can achieve 90% detection accuracy. (3) The dual domain features excel in fake image detection. They facilitate higher detection accuracy, enable zero-transfer learning between different AIGen models, and allow more efficient training.

V. Related Work

In this section, we discuss related work on existing fake image detection algorithms and image forensics.

A. Existing Fake Image Detection

Images in the spatial domain can exhibit a variety of patterns that are common among images generated by diffusion models. An example of such is distorted writing and meaningless text [8]. This is often a key giveaway that an image is not real. Artifacts in the form of deterioration are also commonly found near the edges of an image, providing another feature for classifiers to learn [5]. Maybe's AI Art Detector attempted to learn these features by training a Vision Transformer model on thousands of images scraped from Reddit. To an extent, the model was successful with high test accuracy. Unfortunately, in real-world applications, realistic fake images were often left undetected.

Our work recognizes this issue and addressed it with a larger and more comprehensive dataset. Additionally, our dual-domain model proves this is no longer an issue with high accuracy for detecting the hyper-realistic images that resulted from fine-tuning with DreamBooth.

B. Image Forensics

Previous work shows the frequency response of images from diffusion models exhibits a thin grid-like structure due to the use of an adversarially trained auto-encoder [6]. As GenAI improves, many of the artifacts visible in the frequency domain are slowly fading, making training in the frequency domain potentially more difficult.

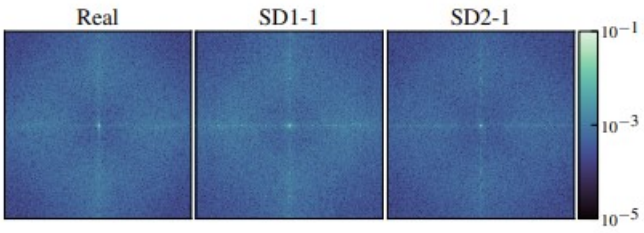


Fig. 7: Frequency spectrum of a real image and images from Stable Diffusion 1-1 and 2-1 as illustrated in [8, Fig. 4]

With a dual-domain approach, even if artifacts in the frequency domain become less prominent, the accuracy will not suffer as much as models trained only in the frequency domain. Dual-domain models will simply assign greater importance to the spatial features. This further supports the resilience of a dual-domain approach.

VI. Conclusion

In this paper, we propose a dual-domain CNN-based model to detect fake images generated by text-to-image models. The introduction of frequency domain features facilitates higher detection, accuracy, and faster convergence. To our knowledge, this is the first detection model specifically tested against models which produce ”personalized” images.

In the future, we plan to build an exhaustive dataset of images from DreamBooth and incorporate it into the training process. Furthermore, we also plan to analyze our model’s resistance to image processing techniques like Gaussian blur and anti-aliasing.

VII. Acknowledgement

This research is based upon work supported by the National Science Foundation under the REU Site Grant CNS-2050548.

References

[1] Chen Chen, Jie Fu, and Lingjuan Lyu. A pathway towards responsible ai generated content, 2023.
[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.
[3] Zhengyuan Jiang, Jinghuai Zhang, and Neil Zhenqiang Gong. Evading watermark based detection of ai-generated content, 2023.
[4] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.
[5] Matthew Maybe. Can an ai learn to identify “ai art”?, May 2023.
[6] Jonas Ricker, Simon Damm, Thorsten Holz, and Asja Fischer. Towards the detection of diffusion model deepfakes, 2023.
[7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.
[8] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, 2023.
[9] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. De-fake: Detection and attribution of fake images generated by text-to-image generation models, 2023.
[10] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.
[11] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022.
[12] Fangneng Zhan, Hongyuan Zhu, and Shijian Lu. Spatial fusion gan for image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.