CS-GY 6923 INET: Project Report

Topic: Det-IGEN: Detectable Watermarking for Image Generation

---

Shashank Pandey (NetID: sp8108)
Sirish Parupudi (NetID: vsp7230)

---

1. **Introduction**

The rapid proliferation of diffusion models in image generation, exemplified by popular tools like DALL-E, Gemini, and Grok2, has revolutionized the field of artificial intelligence. These models have demonstrated remarkable capabilities in creating highly realistic and diverse images from textual descriptions. However, this advancement brings forth a critical challenge: the need for a reliable method to distinguish AI-generated images from authentic ones. As the quality of AI-generated content continues to improve, traditional detection methods become increasingly inadequate, necessitating a more robust approach to ensure the integrity and provenance of visual media.

This is an urgent problem, as recent events have shown us the potential consequences when AI-generated images are perceived as real. During Hurricane Helene in 2024, AI-created images of a tearful child holding a puppy in a rescue boat went viral (Image 1), stirring strong emotional reactions and misleading millions of viewers. These images were used to spread partisan narratives and criticize disaster response efforts, undermining public trust in first responders and government institutions, including FEMA (the Federal Emergency Management Agency).

Beyond natural disasters, AI-generated images were exploited for political propaganda to great effect during the 2024 election, with examples including a manipulated image depicting Vice President Harris in communist attire (Image 2), and an image that appeared to show pop star Taylor Swift endorsing former President Trump (Image 3). In both cases, the images were shared with millions of Americans by famous personalities, including X (and Grok2) owner Elon Musk and former President Trump himself, without indicating that the images were AI-generated. These incidents show how AI-generated content can be used to spread misinformation, manipulate public opinion, and even facilitate scams targeting vulnerable individuals.

(Image 1)

(Image 2)

(Image 3)

These examples only strengthen our argument that AI-generated images should be easily detectable, and these generative models should explicitly encode methods that indicate that their outputs are synthetic, generated images (solving this problem at source).

In this report, we present our framework, Det-IGEN (**Det**ectable Watermarking for **I**mage **GEN**eration). Det-IGEN addresses this challenge by proposing a novel watermarking technique integrated directly into the diffusion model's generation process. This method embeds a unique, noise-based signature within the model, which generates images that contain frequency-based watermarks. The watermark is designed to be resilient against common image manipulations while maintaining the visual quality of the generated images.

2. **Background and Related Work**

The problem of designing a reliable method for AI-generated image detection has been an exploding question in the image generation field, with GANs (Generative Adversarial Networks) being one of the most popular - and best-in-class approaches used to generate images. In recent years, transformer-based structures have also been explored as an alternative, where images are converted to token lists using an image decoder. Diffusion models, where we slowly add random data as part of a process where we learn how to iteratively reverse the diffusion process, in a denoising process (to generate an image), have brought huge improvements in text-conditional image generation, and can synthesize high-resolution photo-realistic images for a wide variety of text prompts. Because of their iterative denoising algorithm, diffusion models can also be adapted for image editing in a zero-shot fashion by guiding the generative process. These models are capable of sampling new images at inference time by iteratively processing an initial noise map.

There are two broad approaches to generative image detection - passive detectors, and watermark-based detectors.

In passive detectors, we use features from the image to classify images into AI-generated and non-AI-generated classes. These features can include frequency domain information (as seen in FreqCLIP), or spatial information from the image (like in UnivCLIP). These methods have, with time, proven to give us sub-optimal results. We are following a watermarking-based approach to design our model. Broadly, in these approaches, noise (a watermark) is added at some point in the generative process. This noise is what is leveraged to make out that the image generated is AI-generated.

A watermarking method typically consists of three components: watermark, encoder, and decoder. The encoder embeds a watermark into content, while a decoder decodes a watermark from watermarked/unwatermarked content. When a content has a watermark, the decoded watermark is similar to our actual watermark. The encoder and watermark can also be embedded into the parameters of our model, such that its generated content is inherently watermarked.

3. **Dataset**

In our evaluation of our approach, we test our performance on two tasks: in one task, we see how our detection approach (which aims to find the presence of a watermark) works on a mixture of AI-generated images from both our augmented model and from a standard diffusion model. In the other task, we test our detector's performance in classifying images that are AI-generated (images from our model) against real-world images.

For the two tasks, therefore, we need two different datasets - a collection of prompts for the first task, and a set of real-world images for the second task.

3.1 DiffusionDB:

For the first task, we use the DiffusionDB dataset. The DiffusionDB dataset contains over 14 million detailed prompts, with well-defined instructions of high complexity. The dataset has been used in multiple SOTA image detection and image generation tasks.

Sample prompt from DiffusionDB: "*a renaissance portrait of dwayne johnson, art in the style of rembrandt!! intricate. ultra detailed, oil on canvas, wet - on - wet technique, pay attention to facial details, highly realistic, cinematic lightning, intricate textures, illusionistic detail*"

3.2 COCO:

For the second task, we used the COCO (**C**ommon **O**bjects in **Co**ntext) dataset to source our real-world images. COCO is a widely used dataset that contains real-world images belonging to 80 categories.

Preprocessing

1. **Resize**: Scales the image to the target_size (default is 512 pixels on both dimensions).
2. **Center Crop**: Crops the center of the image to ensure it is square with dimensions target_size x target_size.
3. **Convert to Tensor**: Converts the processed image into a PyTorch tensor.
4. **Normalize**: Adjusts the pixel values to the range [-1.0, 1.0] using the formula 2.0 * image - 1.0.

4. **Solution Approach**

4.1 Masking:

As we have already mentioned, we decided to follow a watermarking-based approach in our design, i.e., adding a pattern to some component in the generative process, rather than adding the pattern directly to the generated image.

We decided to experiment with different types of watermarking to judge performances. They are as follows:

(i) Zero masking
(ii) Random masking
(iii) Ring masking

In zero-masking, we choose the mask to be a circular region to preserve invariance to rotations in image space. The key is chosen to be an array of zeros, which creates invariance to shifts, crops, and dilations. This key is invariant to manipulations but at the cost of departing severely from the Gaussian distribution. It also prevents multiple keys from being used to distinguish between models.

In random masking, we draw the fixed key k from a Gaussian distribution. The key has the same independent and identically distributed Gaussian nature as the original Fourier modes of the noise array, and so this strategy is expected to have the least impact on generation quality. This method also offers the flexibility for the model owner to possess multiple keys. However, it is not invariant to make image manipulations.

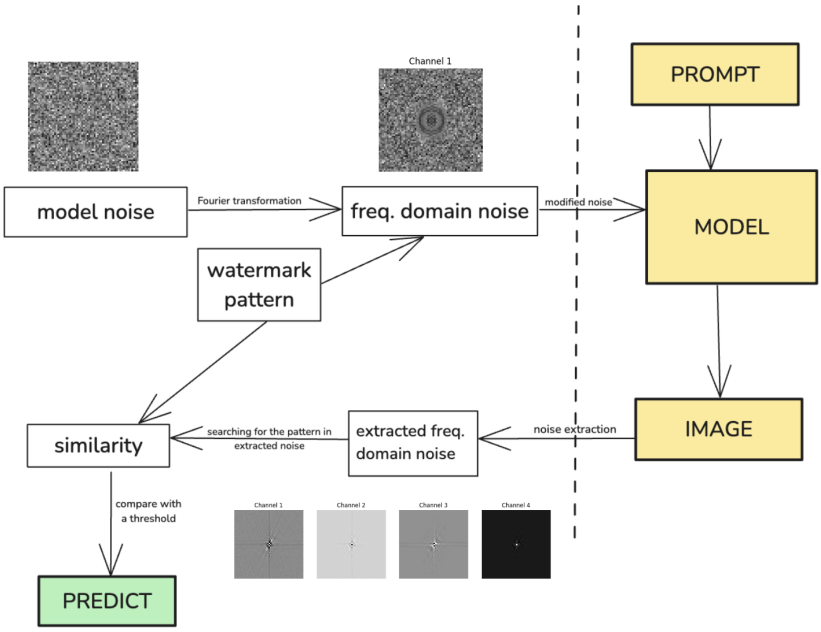
By adding patterns in the frequency domain of our image, we can exploit several classical properties of the Fourier transform for periodic signals:

- A rotation in pixel space corresponds to a rotation in Fourier space.
- A translation in pixel space multiplies all Fourier coefficients by a constant complex number.
- A dilation/compression in pixel space corresponds to a compression/dilation in Fourier space.
- Color jitter in pixel space (adding a constant to all pixels in a channel) corresponds to changing the magnitude of the zero-frequency Fourier mode.

This means that our method of augmenting our model is resistant to common attacks like distorting the image, cropping, and color-changing.

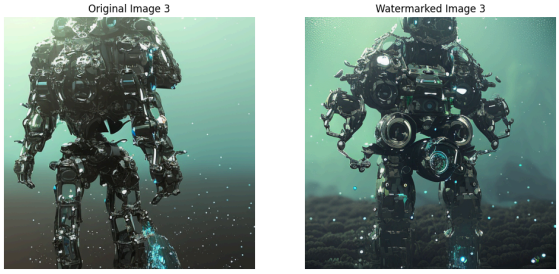
4.2 Solution Pipeline:

We can see our overall pipeline below:



An overview of the complete DetIGEN pipeline
(for both tasks of image generation and image detection)

As we see above, we modify the noise that is part of the diffusion model, by adding our pattern in the frequency domain of the noise. This augmented/modified model is used to generate images from prompts. Since there are no modifications done to the generated image, we do not see any loss in quality.



(The left image is generated by the original model, while the right image is generated using our augmented model. As we can see, there is no difference in quality between the images)

During the detection process, we first effectively invert the generation process, using an extraction algorithm to obtain the random noise of the model.

We then compute the distance between the Fourier transform of this noise and our watermark in the frequency domain, and judge it to be watermarked if the distance is below a threshold. This comparative approach allows us to fine-tune our threshold such that we avoid false positives (or) true negatives.

We experimented with different patterns of the watermark (with the zero pattern, random pattern, and ring pattern), and with different sizes of the watermark (by changing the radius of the watermark).

5. Results

	Task 1: classifying watermarked images				Task 2: AI-generated image detection			
	accuracy	precision	recall	f1-score	accuracy	precision	recall	f1-score
radius = 10	0.99	0.99	1.0	0.99	0.99	0.99	1.0	0.99
radius = 5	0.99	0.99	1.0	0.99	1.0	1.0	1.0	1.00
ring-shaped watermark	0.96	0.93	1.0	0.97	0.98	0.96	1.0	0.98

We ran our pipeline across various parameters of radius, shape, and pattern, with three distinct pattern types:

- Tree-Ring-Zeros (a circular mask of zeroes)
- Tree-Ring-Rand (where we use randomness in our mask pattern)
- Tree-Ring-Rings (we induce a pattern of concentric rings)

Tree-Ring-Zeros: This method uses a circular mask to maintain invariance to rotations in image space. The key is an array filled with zeros, which ensures robustness against shifts, cropping, and scaling. However, this approach significantly deviates from the Gaussian distribution and limits the ability to use multiple distinct keys to differentiate models.

Tree-Ring-Rand: In this approach, a fixed key k is sampled from a Gaussian distribution, maintaining the independent and identically distributed (iid) Gaussian characteristics of the original Fourier noise. This method is expected to have minimal impact on generation quality due to its similarity to the original noise. Additionally, it allows for the use of multiple keys by the model owner. However, it lacks robustness to image manipulations.

Tree-Ring-Rings: This strategy introduces a pattern of concentric rings, with a constant value along each ring, ensuring rotational invariance. These constant values are sampled from a Gaussian distribution, offering some resistance to various image transformations while keeping the distribution close to an isotropic Gaussian.

While going through the results, we observed some interesting insights:

Across all models (and all tasks), not once did our detection algorithm incorrectly classify a watermarked AI-generated image as a real image (as we see that recall = 1.00 in all parameters). This is hugely encouraging, as it means that Det-IGEN does not miss a single AI-generated image. **This is effectively an 'inescapable' image detection paradigm.** This result intuitively makes sense, as we know that Fourier transforms are invariant and remain even after image modifications; and therefore has very positive real-world implications.

We are also encouraged by the sparsity of false positives. With precision scores > 0.9 in all runs, there are very few images that we are incorrectly classifying as AI-generated.

Lastly, as shown in the outputs, the fact that the visible nature/qualities of the images remain untouched means that we can preserve the realism found in best-in-class image generation models, while still being able to watermark the process and detect it.

6. Conclusions

We have successfully developed and implemented a robust pipeline capable of detecting whether an image was generated by AI. This system can be a valuable tool for AI wrapper companies and developers of deployable models, enabling them to integrate such detection methods to curb the spread of misinformation and false propaganda. By identifying AI-generated images, organizations can help foster transparency and trust in digital media, mitigating the risk of misuse and deception.

This problem can also be extended beyond simple detection to include attribution, where the origin of the AI-generated image can be traced back to the specific user or entity responsible for its creation. Such capabilities would be particularly useful for accountability and enforcement in scenarios where malicious actors attempt to exploit AI tools for harmful purposes.

Additionally, this framework can be enhanced further by incorporating advanced image encoding techniques. These techniques could embed imperceptible but unique messages or identifiers within images, acting as watermarks or signatures. These encoded messages would serve as an additional layer of detection and attribution, enabling the tracing of both the model and the user behind the generation of the image. Such innovations would pave the way for a more sophisticated and comprehensive system for managing AI-generated content, ensuring its ethical and responsible use across diverse applications.

7. References

1. Fernandez, Pierre, et al. "The stable signature: Rooting watermarks in latent diffusion models." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.
2. An, Bang, et al. "Benchmarking the robustness of image watermarks." arXiv preprint arXiv:2401.08573 (2024).
3. Wen, Yuxin, et al. "Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust." arXiv preprint arXiv:2305.20030 (2023).