

SpiderMark: A Hybrid Watermarking Framework for Generative AI with Algorithmic Injection and Learning Verification

Chihcheng Hsieh, Jacinto C. Nascimento², Catarina Moreira¹

¹ Data Science Institute, University of Technology Sydney, Australia

² Instituto Superior Técnico, University of Lisbon, Portugal | INESC-ID

Motivation

Generative AI can now create images, video, and audio that are **virtually indistinguishable from reality**.

While this fuels innovation, it also triggers an **authenticity crisis**: *deepfakes* and synthetic media blur the boundary between the real and the artificial, spreading misinformation, manipulating opinion, and eroding democratic trust.

The inability to verify digital provenance threatens public confidence in what we see, hear, and believe.



Research Gap & Question

Current watermarking methods for text-to-image generators embed a hidden pattern in the **noise or latent space** of the diffusion process.

- **Poor Robustness**: The watermark can be easily corrupted by jailbreak attacks such as image rotation, cropping, or compression
- **Detection Difficulty**: After generation, models struggle to detect the hidden watermark, making verification unreliable

*How can we design a **watermarking** framework for text-to-image models that remains **invisible**, yet **verifiable** and **robust** under real-world transformations?*

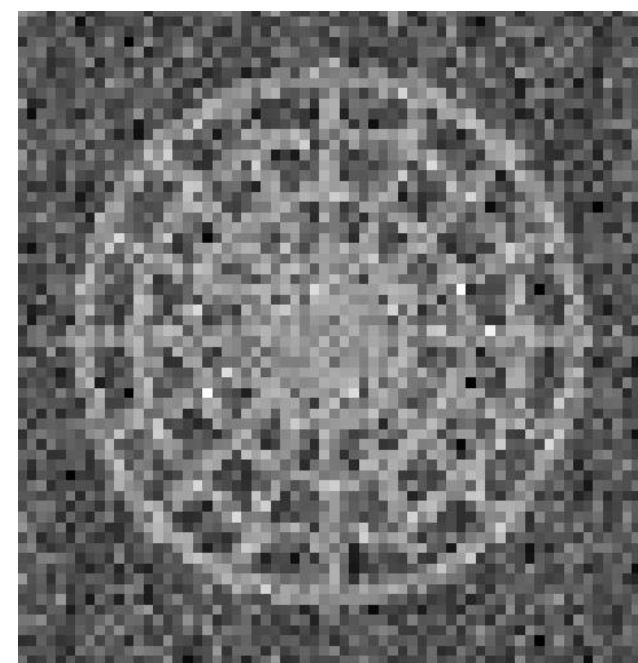
SpiderMark

1. Pattern Generation

We use log-polar coordinates to generate a spider pattern signature in Fourier Space

Key properties

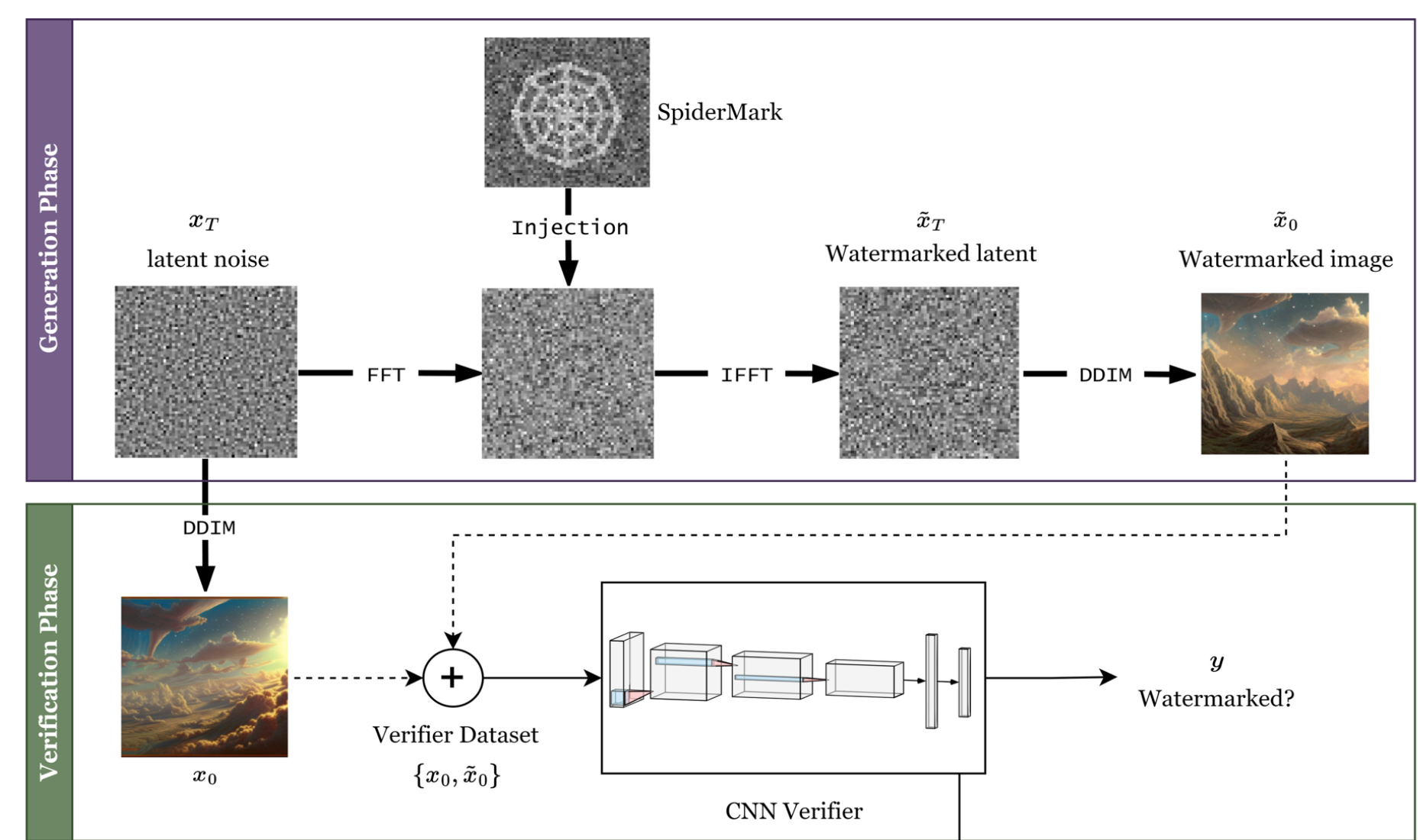
- **Dimensionality**. 2D structure (radial + angular)
- **Information**. High key capacity
- **Robustness**. Rotation and translation invariant
- **Easy CNN learning**. CNNs excel at detecting edges making them easier for verification



2.1. Injection Phase

The watermark is **injected directly into the diffusion process**. A lightweight algorithm embeds a **spider-web pattern** in the **Fourier domain of the latent noise** before image generation. Because the pattern resides in frequency space, it remains **invisible in the spatial domain** and **robust to geometric transformations** such as rotation or scaling.

2. Watermark Injection & Watermark Detection



2.2. Detection Phase

To verify authenticity, the image is **inverted back to its latent representation** using DDIM inversion.

A **CNN verifier** then analyses the Fourier-transformed latent and compares it against the **known SpiderMark pattern** and mask. This learned verification approach enables **reliable detection** even when the image has been compressed, cropped, or perturbed.

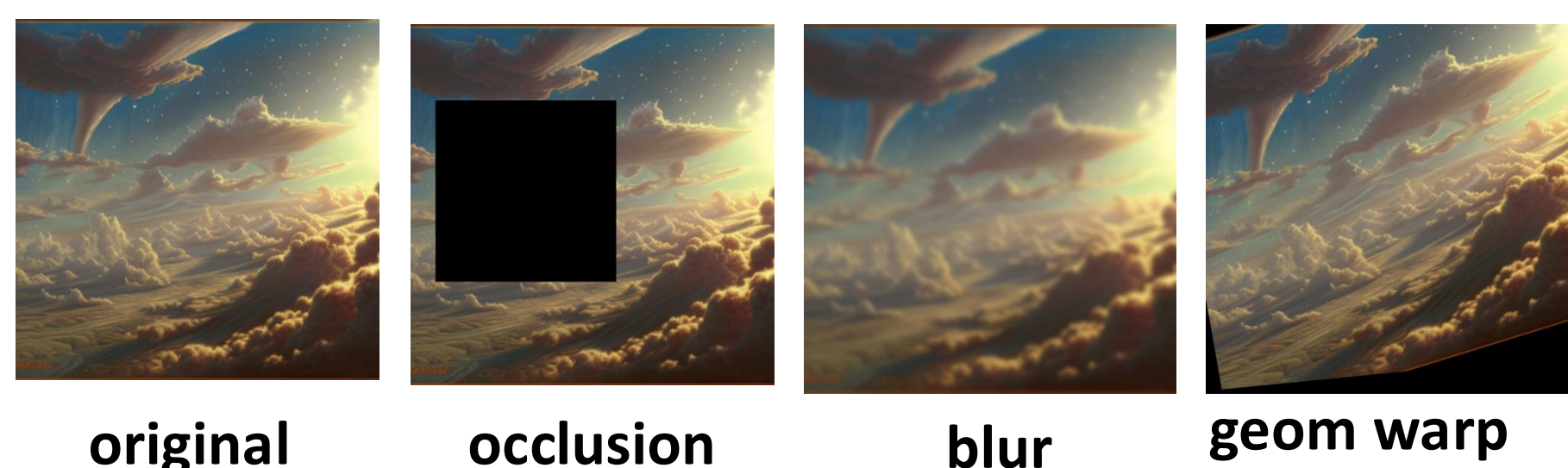
Results

SpiderMark was evaluated on **Stable Diffusion v2.1** using **5 K MS-COCO 2017** images.

Each image was generated in **clean** and **watermarked** forms. A **ResNet-18 verifier** was trained to classify whether an image contained the watermark under diverse real-world attacks.

Attack Scenarios

We tested our model in seven different attacks: jpeg compression, down/up scale, blur, cropping, occlusion, geometric wrapping and training augmentation (a mix of rotation, scaling, jittering, etc)



Attack	Baseline (PSNR)		SpiderMark	
	Accuracy	AUROC	Accuracy	AUROC
clean (no attack)	0.9533	0.9557	0.9333 (-0.02)	0.9918 (+0.04)
jpeg strong	0.58	0.6013	0.7467 (+0.17)	0.8297 (+0.23)
down/up rescale	0.5533	0.5759	0.8133 (+0.26)	0.8645 (+0.29)
blur	0.4733	0.5	0.6213 (+0.15)	0.7027 (+0.20)
random crop	0.56	0.5823	0.808 (+0.25)	0.8912 (+0.31)
occlusion	0.92	0.9241	0.9227 (+0.00)	0.9774 (+0.05)
geom warp	0.5333	0.557	0.7973 (+0.26)	0.8822 (+0.33)
train aug mix	0.5533	0.5759	0.728 (+0.17)	0.7926 (+0.22)

Results show that SpiderMark is **invisible**, and can be embedded directly in the diffusion process, **achieving near-perfect detection accuracy** while preserving image quality and robustness to real-world distortions.

Conclusions

SpiderMark combines algorithmic injection and learned verification for **robust, invisible watermarking** in diffusion models.

It improves robustness by **+18%** under real-world attacks while preserving image quality.