IBIMA Publishing
Communications of International Proceedings
https://ibimapublishing.com/p-articles/46AI/2025/4626425/
Vol. 2025 (34), Article ID 4626425

Diffusion Models as the Future of Robust Watermarking: the presentation of ZoDiac and SuperMark Approaches*

Marta BISTROŃ and Zbigniew PIOTROWSKI

Military University of Technology, gen. Sylwestra Kaliskiego 2, 00-902 Warsaw, Poland

Correspondence should be addressed to: Marta BISTROŃ, marta.bistron@wat.edu.pl

* Presented at the 46th IBIMA International Conference, 26-27 November 2025, Ronda, Spain

Abstract

The rapid advancement of generative AI has created unprecedented challenges for the authenticity and security of digital content. Traditional watermarking methods—based on spatial, frequency, or deep learning—are proving increasingly vulnerable to contemporary attacks, including adversarial perturbations, generative model regeneration, and latent space manipulation. This paper examines the threats posed to watermarking and highlights diffusion models as a promising foundation for robust watermarking. Two representative approaches, ZoDiac and SuperMark, are presented and their characteristics and empirical robustness are compared, both being training-free approaches. Experimental evidence shows that diffusion-based watermarking achieves near-perfect robustness (>99%) against standard distortions and high robustness against adaptive AI-based attacks, while maintaining high fidelity. These findings suggest that diffusion models provide an inherently robust framework that can overcome the limitations of conventional watermarking strategies.

Keywords: diffusion models, robust watermarking, SuperMark, ZoDiac.

Introduction

The rapid development of generative artificial intelligence has transformed the way digital content is created. Modern diffusion-based image generators, such as Stable Diffusion and Imagen, enable the creation of photorealistic visualizations at scale, opening up new possibilities for creativity, design, and communication. At the same time, these technologies pose unprecedented challenges to the authenticity, ownership, and trust of content. Images and videos can be easily synthesized, manipulated, or reproduced, raising concerns about copyright protection, disinformation, and compliance with emerging regulations such as the EU Artificial Intelligence Act (AI Act) (European Commission, 2024) and the C2PA (C2PA, 2023) digital provenance standard.

The classic approach to protecting visual content is digital watermarking—embedding unnoticeable signals into an image or video, which can then be easily extracted to verify their origin. Early watermarking algorithms focused on spatial or frequency transformations, and later on deep learning methods such as convolutional neural networks or GANs. While effective in combating traditional distortions such as JPEG compression or simple filtering, these approaches are becoming increasingly vulnerable to modern attacks, including adversarial attacks, regeneration via generative models, and automated sanitization processes, which can remove or destroy watermarks with high efficiency without affecting the quality of the visual content.

This article examines the evolving nature of watermarking threats in the era of generative AI and discusses why diffusion models can provide a path to more robust watermarking without the need for robustness-enhancing mechanisms, such as a noise layer, during algorithm training.

Cite this Article as: Marta BISTROŃ and Zbigniew PIOTROWSKI, Vol. 2025 (34) "Diffusion Models as the Future of Robust Watermarking: the presentation of ZoDiac and SuperMark Approaches " Communications of International Proceedings, Vol. 2025 (34), Article ID 4626425, https://doi.org/10.5171/2025.4626425

The article is structured as follows: Section 2 discusses the evolution of watermarking attacks and highlights new challenges posed by AI-based manipulation. Section 3 discusses the inherent robustness of diffusion models and introduces new watermarking approaches. Section 4 concludes the discussion.

Attacks on Watermarking Algorithms: Past and Present

Over the past two decades, watermarking algorithms have been challenged by a wide range of attacks. Early threats often posed unintended side effects of multimedia processing, such as lossy compression or resampling. Later, attackers deliberately designed manipulations to alter the watermark structure, including geometric transformations, statistical analysis, or destructive compression.

With the development of deep learning, new classes of attacks have emerged. Generative models can regenerate or inpaint content, effectively removing watermarks without visible artifacts. Adversarial perturbations exploit weaknesses in watermark detectors, while encoder-decoder networks can be trained specifically for watermark removal. Recent latent space attacks utilize diffusion or variational autoencoder (VAE) models, which enable multimedia regeneration in a way that bypasses or completely removes embedded signals. **Table *1*** summarizes the main categories of watermarking attacks, their mechanisms, and their impact on watermarking.

Table 1. Overview of watermarking attacks: evolution and characteristics

(Bistroń, Żurada and Piotrowski, 2025).

Group of attacks	Type of attacks	Description	Examples of operations	Affecting the watermark
Untargeted attacks	-	Attacks resulting from routine processing of the media, with no intention of removing the watermark, but which may affect its integrity.	Lossy compression (JPEG, MPEG), scaling, filtering.	Partial loss or distortion of the watermark.
Targeted attacks	General	Intentional manipulation of the media to remove, distort, or weaken the watermark.	Rotate, scale, change resolution.	Total or partial loss of watermark.
Targeted attacks	Statistical attacks	Modifications using statistical analysis of the media to identify and remove the watermark.	Histogram attack, frequency distribution analysis, autocorrelation attack.	Removal of the watermark without significant changes in the perception of the media.
Targeted attacks	Sensitivity attacks	Minimal modifications to the media that do not affect the visual quality, but destroy the watermark.	Bit depth reduction, delicate pixel changes.	Distortion or complete loss of the watermark without visible changes in the media.
Targeted attacks	Destructive compression	Aggressive compression to remove the watermark by extremely reducing the media data.	High-loss JPEG compression.	Significant loss of data, complete destruction of the watermark, degradation of media quality.
Targeted attacks	Geometry attacks	Manipulations in the spatial structure of the media that distort the position of the watermark.	Rotation, translation, change of proportions.	Disturbance of watermark position, loss of synchronization.
Deep learning-based attacks	Generative attacks	Use of generative models to regenerate media content and remove embedded watermark.	Image inpainting, deepfake generation, AI-based restoration.	Complete removal of the watermark without perceptual changes.
Deep learning-	Adversarial attacks	Modifications generated by neural networks to fool detection systems and	Adversarial noise, gradient-based attacks (FGSM, PGD).	Degradation or undetectability of the watermark.

based attacks		weaken watermark extraction.		
Deep learning-based attacks	Neural network removal	Use of DL models trained to detect and remove watermarks.	CNN-based watermark removal, encoder-decoder architectures.	High probability of watermark elimination with minimal distortion.
Deep learning-based attacks	Content replacement	Media content is regenerated using deep learning models to overwrite or bypass the watermark layer.	GAN-based texture replacement, style transfer techniques.	Loss or severe weakening of the watermark.
Deep learning-based attacks	Latent-space attacks	Attacks that exploit generative models (diffusion or VAE-based) to regenerate content in the latent space, effectively removing embedded watermarks.	Stable Diffusion regeneration, DERO, VAE sampling attacks.	Total removal of the watermark, especially in latent-domain watermarking schemes.

There is a clear shift in the nature of threats. While early watermarking systems primarily encountered unintended distortions or simple geometric transformations, today's algorithms must be resistant to sophisticated AI-based attacks that can regenerate or reconstruct entire multimedia files. Such attacks are often imperceptible to human observers, yet they destroy the integrity of the watermark. This evolution highlights the urgent need for innovative approaches. In this context, diffusion models emerge as a promising direction, as their intrinsic design—based on iterative denoising and latent space processing - can provide watermarking with a level of reliability that conventional methods cannot achieve.

Diffusion Models as Intrinsically Robust Watermarking

Diffusion models were developed as a method for generating high-quality data, capable of overcoming the limitations of earlier models such as GANs and VAEs. In contrast, diffusion models learn to represent the input data distribution by iteratively perturbing and reconstructing the signal, resulting in significantly improved stability and controllability of the generation process. They operate in a two-phase process:

- a forward diffusion step, in which the data is gradually transformed into noise, and
- a backward diffusion step, in which the model iteratively reconstructs clean data from noisy representations.

This iterative reconstruction, performed in the latent space, means that diffusion models are naturally trained to cope with distortions such as noise, resampling, and compression. For watermarking, this offers a key advantage: watermarks embedded via diffusion are more resistant to manipulations that would otherwise destroy signals embedded in the pixel or frequency domains. This allows the watermark to be permanently and invisibly encoded, while also being difficult to remove or distort.

Initially, diffusion models in watermarking focused primarily on generating synthetic content with a unique identifier. The goal of such solutions was to unambiguously determine that a given content was generated by a specific architecture and to attribute it to its source. However, as technology developed and the capabilities of diffusion models grew, they also began to be used to embed watermarks in real-world images. Several publications addressing this topic have appeared in the literature.

Two representative approaches illustrate how diffusion models can be used for watermarking:

- ZoDiac (Zhang et al., 2024)—a zero-bit watermarking method that embeds a Fourier-domain pattern into the latent noise representation of stable diffusion images.
- SuperMark (Hu et al., 2024) is a multi-bit method that uses a Stable Diffusion Upscaler with Gaussian shading to embed binary signatures while balancing transparency and robustness through residual signals.

A comparison taking into account the most important properties of both methods and their resistance to attacks achieved without additional training is presented in **Table *2***.

Table 2. Diffusion-based watermarking methods: features and robustness

Feature	ZoDiac	SuperMark
Watermark type	Zero-bit (presence detection)	Multi-bit (signature extraction)
Diffusion model	Stable Diffusion (frozen)	Stable Diffusion Upscaler (frozen)
Embedding technique	DDIM inversion + Fourier-domain pattern	Gaussian Shading + residual signal
Extraction	Statistical test on latent spectrum	DDIM inversion + signature decoding
Training	Not required	Not required
Key strengths	Simple, per-image embedding; high robustness to common distortions	Balances transparency and robustness; supports multi-bit extraction
Robustness profile	✓ JPEG compression (>99% accuracy) ✓ Gaussian noise & blur (>99%) ✓ Brightness/contrast (~100%) ✓ VAE-based compression (Bmshj18, Cheng20, >98%) ✓ Stable Diffusion regeneration (Zhao23, ~99%) ✗ Rotation (large drop in detection)	✓ JPEG compression, Gaussian blur, Gaussian noise, brightness, cropping (avg. 99.46%) ✓ Adaptive attacks: VAE-based + diffusion-based regeneration (Zhao23, InsP2P) — avg. 89.29% ✓ Strong transferability across datasets & SR models

Empirical results confirm that ZoDiac achieves near-perfect robustness (>99%) against a wide range of standard distortions, including JPEG compression, Gaussian noise/blur, brightness variations, and even advanced regenerative attacks such as Stable Diffusion regeneration attack (Zhao23). Its main weakness is geometric rotations, which significantly reduce detection accuracy.

SuperMark, on the other hand, demonstrates greater robustness, maintaining watermark extraction accuracy of ~99.46% under normal distortions (JPEG, blur, noise, brightness, cropping) and ~89.29% under adaptive attacks, including VAE-based compression and diffusion-based regeneration (Zhao23, InsP2P). Furthermore, SuperMark demonstrates strong transferability across different datasets, high-resolution models, and watermark embedding techniques, suggesting its practical potential beyond controlled benchmarks.

Together, these findings indicate that diffusion models can handle both zero-bit and multi-bit watermarking in a training-free manner while providing inherent robustness unattainable with traditional deep watermarking frameworks.

Summary

The evolution of watermarking attacks, from simple distortions to sophisticated AI-driven manipulations, underscores the urgent need for new approaches to content protection. Diffusion models, by design, combine iterative denoising and latent space processing, providing inherent robustness to interference. Recent methods such as ZoDiac and SuperMark demonstrate that diffusion-based watermarking can achieve exceptional robustness without additional training, resisting both typical distortions and adaptive regenerative attacks. Both approaches confirm diffusion's potential as a paradigm shift in watermarking. Future work should explore not only the trade-off between invisibility and robustness, but also rotation-invariant watermarking and efficiency aspects for large-scale deployment.

Acknowledgment

This research work was funded by the project of the Military University of Technology titled: "*New Neural Network Architectures for Signal and Data Processing in Radiocommunications and Multimedia*." Project No. UGB 22-054/2025.

References

- Bistroń, M., Żurada, J.M., Piotrowski, Z. (2025), 'Deep Learning for Image Watermarking: A Comprehensive Review and Analysis of Techniques, Challenges, and Applications,' *SSRN preprint*, 2025, [Online], [Retrieved September 28, 2025], https://ssrn.com/abstract=5332823
- Coalition for Content Provenance and Authenticity (C2PA) (2023), 'C2PA Technical Specification v1.3,' [Online], [Retrieved September 28, 2025], https://spec.c2pa.org/specifications/specifications/2.2/specs/C2PA_Specification.html.
- European Commission (2024), 'Artificial Intelligence Act,' Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 on Artificial Intelligence, Official Journal of the European Union, L 168/1, 12 July 2024.
- Hu, R., Zhang, J., Li, Y., Li, J., Guo, Q., Qiu, H., Zhang, T. (2024), 'SuperMark: Robust and Training-free Image Watermarking via Diffusion-based Super-Resolution,' *arXiv preprint* arXiv:2412.10049, [Online], [Retrieved September 28, 2025], https://arxiv.org/abs/2412.10049
- Zhang, L., Liu, X., Viros Martin, A., Xiong Bearfield, C., Brun, Y., Guan, H. (2024), 'Attack-Resilient Image Watermarking Using Stable Diffusion,' *arXiv preprint* arXiv:2401.04247v2, [Online], [Retrieved September 28, 2025], https://arxiv.org/abs/2401.04247