Poster: Frequency-Domain Chaotic Watermarking for Image Authentication Against Deepfakes

Dabbrata Das
Uttara University
dabbrata@uttara.ac.bd

Kaniz Fatima
Rochester Institute of Technology
kf2366@rit.edu

Dhiman Goswami
George Mason University
dgoswam@gmu.edu

Sanchari Das
George Mason University
sdas35@gmu.edu

***Abstract*—Generative AI has made visual content increasingly difficult to authenticate, while detection-based defenses often fail under compression, editing, and unseen models. This work proposes a lightweight frequency-domain chaotic watermarking approach for reliable post-embedding authentication with high imperceptibility. Experiments on CelebA, Celeb-DF, and FaceForensics++ show strong visual quality (PSNR $>$ 51 dB, SSIM ≈ 0.998) and improved robustness compared to spatial-domain methods. The approach offers a practical solution for trustworthy media authentication and identity protection.**

I. Introduction

Modern generative models enable anyone to create highly realistic synthetic images, videos, and audio [1]. While these tools expand creativity, they also expose users to harms such as non-consensual content [2], fraud [3], and tampered evidence [4], [5], undermining safety, reputation, and trust.

AI-enabled fraud is rising rapidly, with some regions reporting annual increases of 410%–1720% [6] and over $410M losses in early 2025 [7]. More than 3000 deepfake tools are publicly available, with tens of thousands of model variants downloadable [8], making high-quality manipulations accessible to non-experts. GANs, autoencoders, and diffusion models produce convincing face swaps, voice clones, and fake identities that remain plausible after standard image processing.

Detection methods such as FaceForensics++, DFDC models, and transformer-based audiovisual forensics [9] fail under common scenarios like resizing, filtering, or unseen model outputs. Human judgment is similarly unreliable, highlighting the need for proactive protection at content creation. This motivates the need for proactive protection mechanisms that secure content at creation time, rather than relying solely on post-hoc detection.

However, current mitigation techniques remain fragile. Spatial-domain watermarks follow predictable patterns and are easily degraded by editing or generative attacks, learning-based watermarks often fail under compression or diffusion-based synthesis, and provenance systems require widespread platform adoption and trusted infrastructure. In contrast, frequency-domain components remain relatively stable under deepfake synthesis. Motivated by this observation, our proposed system embeds a secret, chaotic, key-dependent watermark into mid-frequency DCT (Discrete Cosine Transform) coefficients, enabling imperceptible yet verifiable protection that survives generative manipulation while avoiding the fragility of spatial or platform-dependent methods.

II. System Design

A. Overview

The proposed system presents an active defense framework that embeds a highly secure chaotic watermark into the frequency domain of images. The watermark is designed to be imperceptible to human observers while remaining reliably extractable for authentication purposes, even after deepfake generation and common image manipulations. This enables images to carry intrinsic proof of authenticity without requiring external forensic analysis.

As illustrated in Figure 1, the overall pipeline includes chaotic watermark generation using a secret key, DCT-based embedding during image creation or upload, and watermark extraction for authenticity confirmation. During authentication, the same key is used to regenerate the watermark and compare it with the extracted pattern, allowing the system to distinguish between authentic and manipulated content. This key-dependent mechanism strengthens security by preventing unauthorized parties from forging valid watermarks.

Overall, the work emphasizes an end-to-end authentication framework that enables practical deployment and continuous integrity checks, with a robust, imperceptible, and reliable design suitable for social media content protection, digital forensics, and secure media distribution.

B. Watermark Generation

The watermark is generated using a logistic chaotic map, producing a key-dependent pseudo-random sequence highly sensitive to the initial seed. The map is defined as $x_{n+1} = rx_n(1 - x_n)$ with $x_n \in (0, 1)$, where $r = 3.99$ ensures fully developed chaotic behavior. A secret key sets the initial value x_0, and iterative application of the map generates a sequence $\{x_n\}_{n=1}^{N}$ that forms the basis of the watermark. The generated sequence is reshaped into a two-dimensional chaotic pattern and used as the watermark. Due to the sensitivity of chaotic systems, small changes in the key produce different patterns, making watermark difficult to predict or reproduce while remaining reliably detectable after deepfake manipulation.

C. Watermark Embedding

The proposed system embeds a key-dependent chaotic watermark in the frequency domain to ensure imperceptibility and robustness against manipulations and generative attacks. The input image is resized and split into channels to align

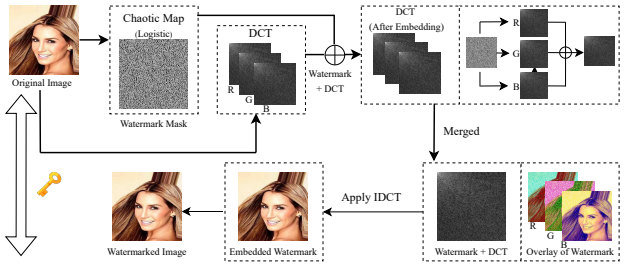

Fig. 1. Chaotic watermark embedding on persistent identity protection.

with the chaotic pattern, which is generated from a secret key. Each channel undergoes a 2D DCT, and the watermark is added to the mid-frequency coefficients, balancing visual fidelity and resilience. The modified coefficients are converted back via inverse DCT, and the channels are recombined to produce the watermarked image, which remains visually indistinguishable from the original while carrying a traceable, resilient watermark.

D. Authentication

Authentication is performed by extracting the watermark from an image and comparing it with a reference chaotic pattern generated from the same secret key. The overall procedure, including watermark embedding and authentication, is summarized in Algorithm 1.

Algorithm 1 Chaotic Watermark Embedding & Authentication

Require: Image f, secret key k, scaling factor α
Ensure: Watermarked image f_{wm} and authentication result
Generate watermark: M from key k using logistic map
Embed watermark:
for each channel c of f **do**
  $D^c \leftarrow \text{DCT}(f_c)$
  $D^c_{\text{wm}} \leftarrow D^c + \alpha \cdot M$
  $f^c_{\text{wm}} \leftarrow \text{IDCT}(D^c_{\text{wm}})$
end for
$f_{\text{wm}} \leftarrow$ recombine channels
Authenticate: regenerate M from k
for each channel c of test image f_{test} **do**
  $W^c \leftarrow (\text{DCT}(f^c_{\text{test}}) - \text{DCT}(f^c))/\alpha$
end for
Compare W with M to verify authenticity

III. Evaluation

Experiments on CelebA, Celeb-DF, and FaceForensics++ demonstrate that the key-dependent chaotic watermark can be efficiently embedded and extracted, enabling reliable authentication of images. The watermark survives GAN-based manipulations, allowing the system to distinguish genuine images from altered ones based on watermark comparison. This mechanism preserves high visual quality (PSNR > 51 dB, SSIM ≈ 0.998) while providing a fast and practical authentication method suitable for real-world deployment.

IV. Preliminary Work

Our currently under-submission research, explores frequency-domain chaotic watermarking for deepfake mitigation, including evaluation, authentication and verification, guidance on threshold selection to enable effective mitigation, and robustness analysis. In contrast, this poster presents a early-stage study focuses on efficient embedding and fast extraction of a key-dependent chaotic watermark using the same secret key. The poster highlights the feasibility and practical potential of watermark-based authentication, without incorporating the full evaluation and threshold-based verification explored in the submitted work.

V. Conclusion and Future Work

The proposed approach provides a lightweight and practical solution for tamper-resistant media authentication. By embedding a key-dependent chaotic watermark without relying on deep learning models, the system enables fast and efficient authentication across images of varying resolutions. The watermark persists through deepfake manipulations while maintaining high visual quality, demonstrating the feasibility of secure, real-time media verification. Currently, a fixed embedding strength and a single chaotic map are used, which perform well but may not be optimal for all images. Future work will explore adaptive embedding that adjusts the watermark strength based on image quality, compression, or platform-specific constraints to enhance robustness and versatility.

Acknowledgment

We would like to thank the Data Agency and Security (DAS) Lab at George Mason University, Uttara University, and the Rochester Institute of Technology, where the study was conducted. The opinions expressed in this work are solely those of the authors.

References

[1] H. Lee, C. Lee, K. Farhat, L. Qiu, S. Geluso, A. Kim, and O. Etzioni, "The tug-of-war between deepfake generation and detection," *arXiv preprint arXiv:2407.06174*, 2024.
[2] J. Langa, "Deepfakes, real consequences: Crafting legislation to combat threats posed by deepfakes," *BUL Rev.*, vol. 101, p. 761, 2021.
[3] F. Muhly, E. Chizzonic, and P. Leo, "Ai-deepfake scams and the importance of a holistic communication security strategy," *International Cybersecurity Law Review*, pp. 1–9, 2025.
[4] Y. Apolo and K. Michael, "Beyond a reasonable doubt? audiovisual evidence, ai manipulation, deepfakes, and the law," *IEEE Transactions on Technology and Society*, vol. 5, pp. 156–168, 2024.
[5] N. Gupta, S. Das, K. Walsh, and R. Chatterjee, "A critical analysis of the prevalence of technology-facilitated abuse in us college students," in *Extended Abstracts of CHI*, 2024.
[6] U. Arshad, A. Tubaishat, S. Anwar, Z. Halim, A. Abualkishik, and A. Ullah, "Web3-based identity and kyc innovations for next-generation fintech," *ACM Transactions on the Web*, 2025.
[7] A. de Rancourt-Raymond and N. Smaili, "The unethical use of deepfakes," *Journal of Financial Crime*, vol. 30, pp. 1066–1077, 2023.
[8] W. Hawkins, B. Mittelstadt, and C. Russell, "Deepfakes on demand: The rise of accessible non-consensual deepfake image generators: The rise of accessible non-consensual deepfake image generators," in *Proceedings of FAccT*, 2025.
[9] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE CVF*, 2019, pp. 1–11.

Frequency-Domain Chaotic Watermarking for Image Authentication Against Deepfakes

Dabbrata Das, Kaniz Fatima, Dhiman Goswami, Sanchari Das
dabbrata@uttara.ac.bd, kf2366@rit.edu, dgoswam@gmu.edu, sdas35@gmu.edu

Motivation

- Need for a practical, imperceptible mechanism to protect media integrity without extra tools.
- Deepfake models primarily manipulate spatial features, while frequency components of the image remain largely preserved.
- Frequency-domain watermarks are difficult to detect or remove and provide a reliable way to verify authenticity.

Abstract

Generative AI increasingly undermines **image authenticity**, while many existing defenses fail under compression, editing, unseen models, and recapture. We introduce a **lightweight frequency-domain chaotic watermarking** approach for reliable authentication. A **secure key-based watermark** embedded in mid-frequency DCT coefficients enables clear separation between authentic and altered images without degrading visual quality, offering a practical and scalable solution for **trustworthy media provenance**.

Results

Table: Imperceptibility Evaluation (Original vs. Watermarked Images)

Dataset	PSNR ↑	FSIM ↑	SSIM ↑
CelebA	51.19	0.998	0.999
Celeb-DF	51.21	0.983	0.997
FF++	51.18	0.998	0.999

Table: Watermark Extraction Time in the Frequency Domain (s/sample) Across Datasets

Domain	Dataset	Extraction Time
Frequency	CelebA	0.0390
	Celeb-DF	0.0420
	FF++	0.0390

Watermark Authentication

Algorithm 1 Authentication via Watermark Extraction

Require: Watermarked image f_{wm}, original image f, key k, strength α
Ensure: Extracted watermark W and authentication result
1: Split f_{wm} and f into channels
2: Generate watermark M from key k
3: **for** each channel c **do**
4: $W^c \leftarrow (\text{DCT}(f_{\text{wm},c}) - \text{DCT}(f_c))/\alpha$
5: **end for**
6: W is compared with M to confirm authenticity

Contribution

- Embeds a key-protected chaotic watermark in DCT coefficients for secure image authentication without affecting visual quality.
- Enables secure authentication of images using the same key, ensuring authenticity without exposing the watermark.
- Modifies only a small portion of DCT coefficients and requires no deep learning, enabling lightweight and real-time deployment.

Design Overview

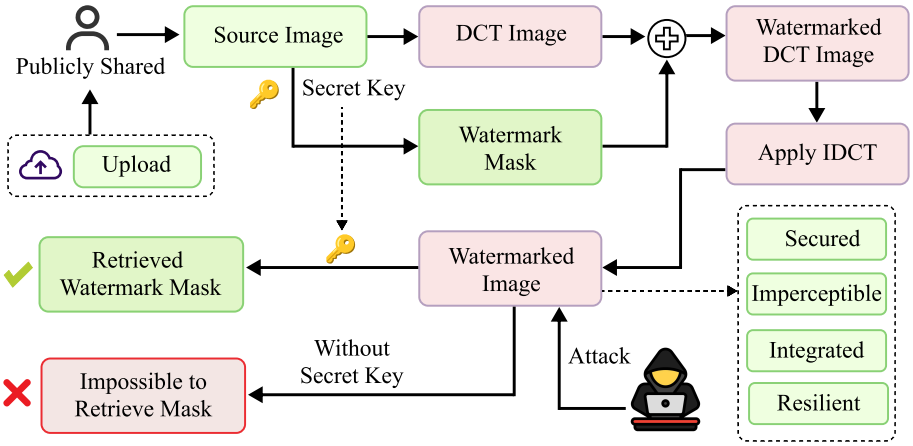

Figure: The system embeds a secret-key chaotic watermark in the DCT frequency domain to enable robust image authentication against deepfakes while preserving imperceptibility, image quality, and high-quality retrieval.

Evaluation

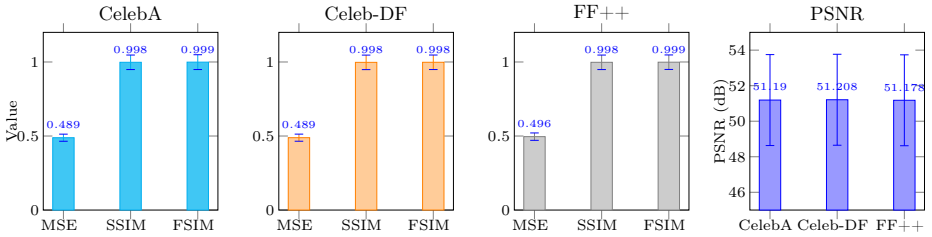

Figure: Comparison across CelebA, Celeb-DF, and FF++ datasets for MSE, SSIM, FSIM, and PSNR. Error bars indicate 5% significance level.

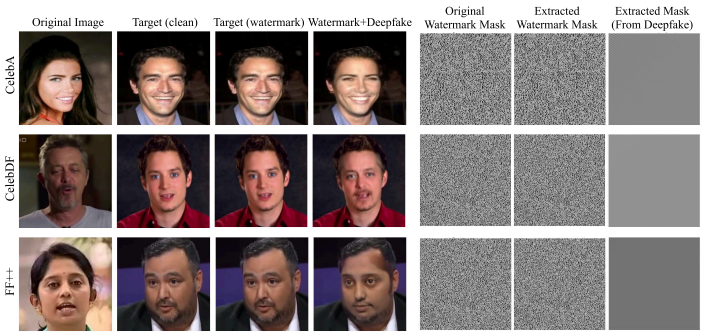

Figure: Watermark visualization and extraction, showing original, watermarked, and deepfake images alongside original and extracted masks.

Table: Comparison of Watermark-based authentication across methods. Symbols: ✓ = strong, ● = moderate, ✗ = weak.

Method	Core Idea	Deepfake Resistance	Imperceptibility	Efficiency
DiffMark [1]	Diffusion watermark	●	●	✗
CMUA-Watermark [2]	Universal perturbation	●	✗	✗
FaceShield [3]	DM attention + perturbation	●	●	●
LEAT [4]	Latent disruption	●	●	✗
CIN [5]	Dual-mode watermark	●	✓	●
LandMark [6]	Landmark watermark	●	●	●
Ours	**Chaotic DCT watermark**	✓	✓	✓

References

[1] Sun, C., Sun, H., Guo, Z., Diao, Y., Wang, L., Ma, D., Yang, G., & Li, K. (2026). DiffMark: Diffusion-based robust watermark against Deepfakes. *Information Fusion*, 127, 103801. https://doi.org/10.1016/j.inffus.2025.103801
[2] Huang, H., Wang, Y., Chen, Z., Zhang, Y., Li, Y., Tang, Z., et al. (2022). CMUA-Watermark: Cross-model universal adversarial watermark for combating deepfakes. AAAI 36, 989–997.
[3] Jeong, J., In, S., Kim, S., Shin, H., Jeong, J., Yoon, S. H., Chung, J., & Kim, S. (2025). FaceShield: Defending facial image against Deepfake threats. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10364–10374, October.
[4] Shim, J., & Yoon, H. (2025). LEAT: Towards robust deepfake disruption in real-world scenarios via Latent Ensemble Attack. *Expert Systems with Applications*, 279, 127417. https://doi.org/10.1016/j.eswa.2025.127417
[5] Ma, R., Guo, M., Hou, Y., Yang, F., Li, Y., Jia, H., & Xie, X., 2022. Towards blind watermarking: Combining invertible and non-invertible mechanisms. ACM Multimedia, 1532–1542.
[6] Wang, T., Huang, M., Cheng, H., Zhang, X., & Shen, Z., 2024. LampMark: Training-free landmark perceptual watermarks for deepfake detection. ACM Multimedia, 10515–10524.