Comparison of Classical and DL-based Watermarking Methods

Multimedia and Laboratory - Project Report

Daniele Materia
University of Catania
Master's Degree in Computer Science

Academic Year 2025/2026

Abstract—**This project presents a comparative benchmark between classical watermarking techniques and modern Deep Learning-based approaches. We evaluate two traditional methods, Discrete Cosine Transform (DCT) and Discrete Wavelet Transform combined with Singular Value Decomposition (DWT-SVD), against the state-of-the-art DL-based framework PIXEL SEAL. The study focuses particularly on the balance between watermark robustness against common signal processing attacks (JPEG compression, Gaussian blur, Salt-and-Pepper noise) and imperceptibility, providing quantitative evidence of the trade-offs between classical methods and learned neural approaches.**

I. Introduction

Digital image watermarking remains a critical task for copyright protection and provenance tracking in an increasingly digital media landscape. Traditionally, this field has relied on well-understood mathematical transforms to embed data in the frequency domain; however, these approaches often struggle to maintain robustness against complex, non-linear distortions. The emergence of Deep Learning has recently shifted the focus toward learned, end-to-end architectures capable of achieving superior resilience through data-driven optimization.

This project report presents a comparative analysis of these paradigms. By benchmarking two classical transform-domain methods against one state-of-the-art neural framework, we investigate the trade-offs between payload capacity, visual fidelity, and extraction accuracy. We evaluate the performance of the chosen watermarking methods under three different kinds of attacks to the watermark to quantify the practical advantages and limitations inherent in both hand-crafted and learned watermarking strategies. The following sections provide a detailed review of the evolution of these methods and the specific methodologies employed in our benchmark.

II. Related Methods

The scientific literature on digital image watermarking has recently undergone a significant paradigm shift, transitioning from manual feature engineering in transform domains to end-to-end learned representations, following the steady rise of the field of Deep Learning [1], [2]. Traditional watermarking approaches primarily rely on mathematical transforms to identify perceptually insignificant regions in the target image for embedding the watermark. Among these, the Discrete Cosine Transform (DCT) is widely adopted as it operates in the same transform domain as the JPEG compression standard. By embedding the watermark directly into the DCT coefficients, these methods can better anticipate and mitigate the distortions introduced during the quantization stage. Further advancements introduced multi-resolution analysis through the Discrete Wavelet Transform (DWT), frequently coupled with Singular Value Decomposition (SVD). This hybrid approach exploits the inherent stability of singular values, which remain relatively constant even when the image undergoes minor geometric or spatial perturbations. By embedding the watermark within these stable components, the DWT-SVD scheme achieves greater resilience against distortions that typically compromise purely block-based methods.

However, classical methods face a fundamental limitation: their robustness is tightly coupled to the specific mathematical properties of the chosen transform. Each distortion type requires careful hand-tuning of embedding parameters, and they struggle with complex, non-linear attacks or combinations of several distortions, which cannot be always explicitly considered during design. DL-based approaches address these limitations by learning robust embedding and extraction functions directly from data.

The seminal work *HiDDeN* [3] pioneered the field of DL-based watermarking approaches by introducing a neural-network based model comprising three key components: an encoder network that embeds the watermark into the target image, a differentiable noise layer that simulates signal processing attacks (e.g., JPEG compression, Gaussian noise, cropping), and a decoder network that extracts the watermark from the distorted image. Through adversarial training, the encoder learns to embed imperceptible watermarks while the decoder simultaneously learns robust extraction strategies, enabling the system to automatically discover embedding patterns that survive realistic distortions without manual parameter tuning. However, HiDDeN and similar encoder-decoder architectures can exhibit visible artifacts in the generated watermarked images, particularly due to the reliance on pixel-wise reconstruction losses during training.

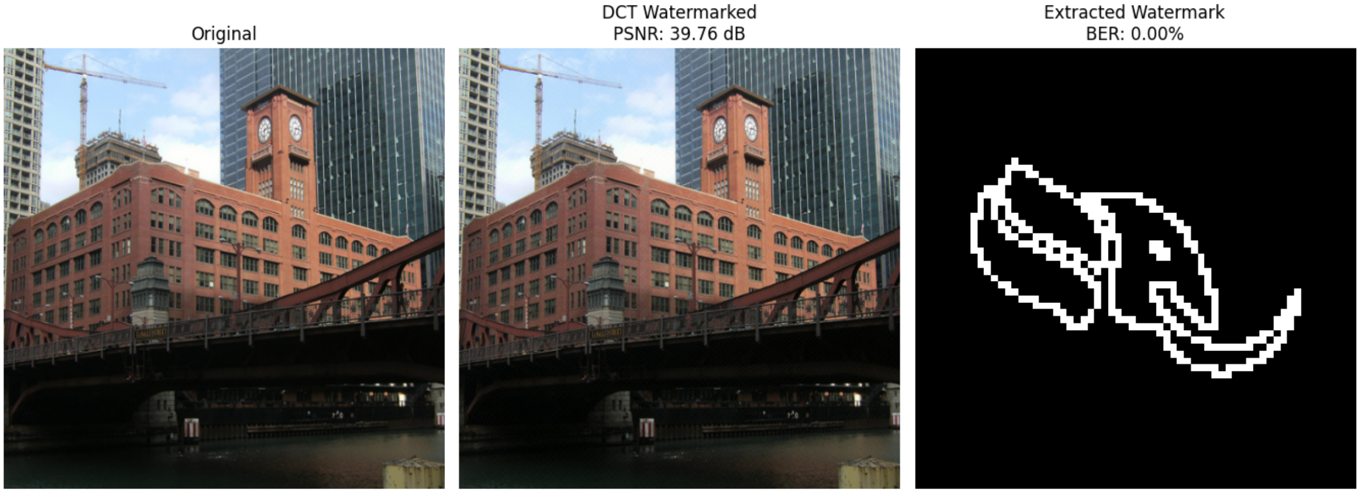

Figure 1. Qualitative example of a 64×64 binary watermark embedded with DCT in a 512×512 image and its extraction in ideal conditions (no attacks).

A different architectural paradigm leverages Invertible Neural Networks (INNs) to model watermark embedding as a structurally reversible transformation. Unlike conventional encoder-decoder architectures, INN-based approaches rely on bijective mappings that preserve dimensionality, enabling the joint transformation of the cover image and watermark into a latent representation from which both signals can be reconstructed through an explicit inverse pass. Under ideal, attack-free conditions, this formulation can approximate lossless reconstruction. By avoiding information bottlenecks typical of encoder-decoder models, INNs can reduce reconstruction error and limit visual distortion in the watermarked image. However, robustness to common image degradations is not inherently guaranteed by invertibility itself. Achieving competitive robustness typically requires explicit attack simulation and carefully designed training objectives. While INN-based methods represent an interesting direction in neural watermarking, this study focuses on adversarial-trained architectures, which have demonstrated stronger empirical robustness.

The state-of-the-art DL-based watermarking model *PIXEL SEAL* [4] addresses this limitation through adversarial-only training, where a discriminator network directly optimizes for visual imperceptibility rather than relying on pixel-wise proxy losses. This achieves superior transparency while maintaining the learned robustness advantages of neural approaches.

III. Methodology

The experimental framework is designed to evaluate the trade-off between watermark invisibility and robustness across the two distinct paradigms. The benchmark is implemented using a Python-based pipeline, where classical baselines leverage established signal processing libraries (OpenCV, NumPy) for transform-domain manipulations, and the deep learning model utilizes PyTorch with hardware-accelerated inference on Apple Silicon (Metal Performance Shaders backend).

A. Classical Watermarking Approaches

The first classical approach considered in our study leverages a block-based Discrete Cosine Transform (DCT), where the cover image is partitioned into 8×8 non-overlapping blocks. A binary payload of 4096 bits is embedded by modulating the mid-frequency coefficients at position $(3, 3)$ of each DCT block with an intensity factor $\alpha = 0.08$. This technique is chosen to balance resistance to JPEG compression with visual transparency. The second classical method employs a hybrid DWT-SVD scheme. The image is decomposed into four sub-bands via a Haar wavelet transform, and Singular Value Decomposition is subsequently applied to localized 4×4 blocks within the LL (Low-Low) sub-band. The watermark bits are embedded by modifying the first singular value of each block with $\alpha = 0.08$, leveraging the mathematical stability of singular values against small pixel perturbations and geometric distortions. Both methods employ informed detection, where the original image is available at extraction time. Bits are recovered by comparing altered values (DCT coefficients or singular values) between watermarked and original images.

B. Deep Learning Approach: PIXEL SEAL

The neural watermarking system evaluated in this study is PIXEL SEAL [4], a state-of-the-art framework consisting of a joint encoder-decoder architecture trained with an adversarial-only paradigm. In our setup, the model operates at its maximum payload capacity of 256 bits, which corresponds to a 16×16 binary watermark image. Such constrained payload lengths constitute a significant bottleneck for neural watermarking frameworks, particularly when compared to the higher bit-densities achievable through classical transform-domain methods.

PIXEL SEAL is specifically designed to achieve superior imperceptibility by removing traditional pixel-wise proxy losses, instead letting a discriminator network define the boundaries of visual transparency during training.

A notable architectural detail discovered during implementation is that the extractor network outputs 257 prediction values rather than the expected 256 bits. Analysis revealed that the first element serves as a synchronization token or confidence score, with the actual message payload residing in indices 1 through 256.

C. Evaluation Protocol

The evaluation is performed on a test set of 200 images, randomly sampled from the COCO [5] 2017 Validation Set, with all images normalized to a resolution of 512 × 512 pixels. This dataset ensures a statistically significant variety of textures, luminance levels, and semantic content. To quantify the visual impact of the watermark, we employ the Peak Signal-to-Noise Ratio (PSNR), defined as:

$$\text{PSNR} = 20\log_{10}\left(\frac{255}{\sqrt{\text{MSE}}}\right) \tag{1}$$

where MSE is the mean squared error between the original and watermarked images. Robustness and extraction accuracy are measured through the Bit Error Rate (BER), defined as the ratio of incorrectly recovered bits to the total number of embedded bits N:

$$\text{BER} = \frac{\sum_{i=1}^{N} |\text{bit}_i^{\text{original}} - \text{bit}_i^{\text{extracted}}|}{N} \tag{2}$$

Three different attack scenarios are simulated: (1) JPEG compression with quality factor $Q = 30$, (2) Gaussian blur with kernel size 3×3 and $\sigma = 1$, and (3) Salt-and-Pepper noise with 1% pixel corruption.

IV. Results

To quantify the trade-offs between hand-crafted transforms and neural network-based models, we evaluate the two watermarking paradigms by analyzing payload capacity, watermark imperceptibility, and its robustness to signal distortions.

A. Classical Approaches: DCT and DWT-SVD

1) Baseline Performance (No Attacks): The block-based Discrete Cosine Transform (DCT) implementation (see Fig. 1 for a qualitative example) yielded a mean Peak Signal-to-Noise Ratio (PSNR) of 39.82 ± 0.15 dB across our test set. Visually, the watermarked images retain high fidelity, as modifications are localized within frequency bands to which the human visual system is less sensitive. In ideal conditions, the Bit Error Rate (BER) remained at 0%, confirming the reliability of the informed detection scheme.

The hybrid DWT-SVD scheme achieved a mean PSNR of 38.63 ± 0.18 dB. While achieving a PSNR that is slightly lower than DCT's (39.82 dB), this method improves robustness under attacks, as discussed below. By modifying singular values within the wavelet domain, DWT-SVD avoided subtle checkerboard artifacts that can emerge in DCT-based methods. The baseline BER was also 0%, as expected from the deterministic nature of these classical embedding functions.

Table I summarizes baseline performance metrics for both classical methods.

Table I
BASELINE PERFORMANCE OF CLASSICAL METHODS (NO ATTACKS). PSNR VALUES REPORTED AS MEAN ± STD.

Method	PSNR (dB) ↑	BER (%) ↓
Block-DCT	39.82 ± 0.15	0.0
DWT-SVD	38.63 ± 0.18	0.0

2) Robustness Under Attacks: To quantify robustness, we measured BER after applying three attack types to the watermarked images.

JPEG Compression. Under JPEG compression with $Q = 30$, DCT exhibited substantial degradation with a BER of $31.7 \pm 8.2\%$. This is attributable to the informed detection scheme: compression alters DCT coefficients in the marked image while the original remains unchanged, and when quantization distortions exceed the embedding strength ($\alpha = 0.08$), comparison-based extraction fails. In contrast, DWT-SVD's singular values demonstrate greater stability under JPEG quantization, achieving a BER of only $0.9 \pm 0.7\%$.

Gaussian Blur. Both classical methods face challenges under Gaussian blurring due to frequency attenuation. DCT achieved a BER of $11.7 \pm 7.3\%$, while DWT-SVD demonstrated superior resilience at $1.0{\pm}0.9\%$. The LL sub-band used in DWT-SVD inherently preserves image structure, implying less degradation.

Salt-and-Pepper Noise. Under impulse noise, DCT achieved a BER of $5.3 \pm 0.9\%$, slightly outperforming DWT-SVD ($7.1 \pm 4.5\%$). The localized nature of impulse noise primarily affects spatial-domain statistics while leaving frequency-domain coefficients relatively stable, explaining DCT's better performance. The impulse noise appears to disrupt singular value stability within the LL sub-band more significantly than the selected DCT coefficient modulation.

B. Deep Learning Approach: PIXEL SEAL

1) Baseline Performance: The PIXEL SEAL framework, operating with a 256 bit payload (16×16 binary watermark), achieved a mean PSNR of 46.22 ± 1.69 dB, significantly outperforming both classical approaches. This superior imperceptibility stems from two factors: (1) the adversarial discriminator's explicit optimization for visual transparency, and (2) the substantially reduced payload capacity (256 vs 4096 bits). By embedding fewer bits, the watermark signal is distributed more sparsely, reducing injected noise energy. Under ideal conditions, PIXEL SEAL maintained perfect extraction reliability with a mean BER of 0.0%.

2) Robustness Under Attacks: The PIXEL SEAL model showed distinct robustness characteristics compared to classical methods.

JPEG Compression. PIXEL SEAL achieved a BER of $9.4 \pm 5.6\%$, positioning it between DWT-SVD (0.9%) and DCT (31.7%). While not matching DWT-SVD's remarkable robustness, PIXEL SEAL significantly outperforms DCT. This reflects the benefits of adversarial training with explicit compression-based augmentations.

Table II
COMPREHENSIVE PERFORMANCE COMPARISON BETWEEN CLASSICAL AND DL-BASED APPROACHES, INCLUDING ATTACKS TO WATERMARK (MEAN $\pm$ STD)

Method	**Payload (bits)**	**No Attack**		**BER (%) Under Attack** $\downarrow$		
		PSNR (dB) $\uparrow$	**BER (%)** $\downarrow$	**JPEG (Q=30)**	**Gaussian Blur**	**Salt-Pepper**
Block-DCT	4096	39.82 ± 0.15	0.0	31.7 ± 8.2	11.7 ± 7.3	5.3 ± 0.9
DWT-SVD	4096	38.63 ± 0.18	0.0	0.9 ± 0.7	1.0 ± 0.9	7.1 ± 4.5
PIXEL SEAL	256	46.22 ± 1.69	0.0	9.4 ± 5.6	0.1 ± 0.2	12.8 ± 0.6

Gaussian Blur. PIXEL SEAL exhibited exceptional resistance, maintaining a BER of merely $0.1 \pm 0.2\%$ – a strong improvement over both DWT-SVD (1.0%) and DCT (11.7%). This remarkable robustness suggests that the model may have learned to encode watermark information in low-frequency components inherently resistant to blur-induced attenuation, demonstrating a key advantage of learned representations over hand-crafted embeddings.

Salt-and-Pepper Noise. Under salt-and-pepper noise, PIXEL SEAL showed moderate vulnerability with a BER of $12.8 \pm 0.6\%$, worse than both the classical approaches.

C. Cross-Paradigm Comparison

Table II presents a comprehensive summary comparing all evaluated methods under both ideal and attacked conditions.

1) Analysis and Trade-offs: The results reveal clear trade-offs between the two paradigms. It is critical to note that BER comparisons are influenced by payload size: classical methods embed 4096 bits, making them statistically more sensitive to bit corruption compared to PIXEL SEAL's 256 bits payload. This difference in information density partially explains the observed robustness patterns.

Classical methods offer higher payload capacity and deterministic extraction with minimal computational overhead. Among them, DWT-SVD emerges as the superior classical approach, demonstrating very strong robustness across all attacks. This comes at the cost of slightly lower baseline PSNR (38.63 dB vs DCT's 39.82 dB), reflecting the trade-off between imperceptibility and robustness controlled by the embedding strength parameter α.

PIXEL SEAL presents a different compromise: reduced payload, but superior PSNR (46.22 dB) and exceptional Gaussian blur resistance (BER = 0.1%). This comes at the cost of significantly higher inference latency compared to classical methods, which is inherent to deep neural network architectures, and may impact large-scale or real-time deployments. The DL-based models seem to prioritize compression and blur robustness, reflecting common real-world distortions and showing moderate vulnerability to less frequent attacks like impulse noise.

The paradigm choice ultimately depends on application requirements: classical methods suit high-capacity controlled environments, while PIXEL SEAL excels when imperceptibility and robustness to unpredictable distortions are main priorities, even at the cost of constrained payload.

V. CONCLUSION

This comparative study evaluated classical frequency-domain watermarking (DCT, DWT-SVD) against the deep learning model PIXEL SEAL. The experimental results quantify distinct trade-offs across payload capacity, visual fidelity, computational efficiency, and watermark robustness.

Classical methods allow for higher payload capacity and are computationally more efficient, with deterministic watermark extraction. However, robustness depends critically on embedding strategy: our DCT implementation showed vulnerability to JPEG compression (BER = 31.7%), highlighting the critical importance of embedding strategy and detection scheme design beyond mere transform-domain compatibility. DWT-SVD demonstrated superior resilience across most attacks, emerging as the more balanced classical approach.

PIXEL SEAL offers a fundamentally different compromise: a lower payload size but exceptional imperceptibility (PSNR = 46.22 dB) and remarkable Gaussian blur robustness (BER = 0.1%). However, this comes at the cost of significantly higher inference latency.

Consequently, the paradigm choice is application-dependent: classical methods excel in controlled environments and allow for high capacity payload, while PIXEL SEAL is preferable when resilience to unpredictable distortions is a priority, even at reduced payload capacity.

REFERENCES

[1] K. M. Hosny, A. Magdi, O. ElKomy *et al.*, "Digital image watermarking using deep learning: A survey," *Computer Science Review*, vol. 53, p. 100662, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574013724000467

[2] Y. Luo, X. Tan, and Z. Cai, "Robust deep image watermarking: A survey," *Computers, Materials and Continua*, vol. 81, no. 1, pp. 133–160, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1546221824007458

[3] J. Zhu, R. Kaplan, J. Johnson, and L. Fei-Fei, "Hidden: Hiding data with deep networks," 2018. [Online]. Available: https://arxiv.org/abs/1807.09937

[4] T. Souček, P. Fernandez, H. Elsahar, S.-A. Rebuffi, V. Lacatusu, T. Tran, T. Sander, and A. Mourachko, "Pixel seal: Adversarial-only training for invisible image and video watermarking," *arXiv preprint arXiv:2512.16874*, 2025.

[5] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft coco: Common objects in context," 2015. [Online]. Available: https://arxiv.org/abs/1405.0312