Informatics and Control Problems 45 Issue 1 (2025)
journal homepage: www.icp.az

Autoencoder-based invisible watermarking: A lightweight deep learning approach for robust embedding and recovery

Abdurahman Vagifli

Azerbaijan Technical University, Baku, Azerbaijan

A R T I C L E   I N F O

Article history:
Received 21.04.2025
Received in revised form 30.04.2025
Accepted 06.05.2025
Available online 04.06.2025

Keywords:
Invisible watermarking
Autoencoder
Deep learning
Image authentication
Tensor flow

A B S T R A C T

This paper presents a lightweight yet robust invisible watermarking framework based on deep autoencoders implemented in TensorFlow. Unlike traditional watermarking methods relying on handcrafted transformations such as DCT, DWT, or SVD, our approach uses a neural network to learn the embedding and extraction processes in an end-to-end manner. We demonstrate that a simple autoencoder trained on a single host image can achieve high imperceptibility and strong robustness against common image processing attacks, including JPEG compression and Gaussian noise. Experimental results show an SSIM of 0.9790 and PSNR of 27.28 dB between the original and watermarked images, with watermark recovery correlation reaching 0.9726 even after JPEG compression and 1.0000 after Gaussian noise, validating the method's effectiveness for real-world applications.

1. Introduction

The protection of digital media through watermarking has become an essential strategy in digital rights management, copyright enforcement, and tamper detection [1]. Traditional invisible watermarking methods typically leverage spatial or frequency-domain transforms such as Least Significant Bit (LSB), Discrete Cosine Transform (DCT), Discrete Wavelet Transform (DWT), and Singular Value Decomposition (SVD). While effective, these approaches often rely on fixed embedding rules that can be brittle under image processing operations.

With the rise of deep learning, autoencoders – neural networks designed to learn data representations through compression and reconstruction – offer a new perspective on watermarking. In this work, we design an autoencoder-based system that learns to embed and later extract a watermark from an image while maintaining imperceptibility and robustness. Although trained on a single host image, the model achieves impressive performance under both clean and attacked conditions.

2. Problem statement

Conventional invisible watermarking techniques, such as DCT, DWT, and SVD, rely on predefined transformation rules to embed information into images. While effective in controlled environments, these methods often lack adaptability and show reduced robustness when faced with distortions like compression or noise. Additionally, they require domain-specific design and fail to

*Corresponding author.
E-mail address: avaqifli77@gmail.com (A.V. Vagifli)

www.icp.az/2025/1-06.pdf   https://doi.org/10.54381/icp.2025.1.06

generalize across different image characteristics or attack types [2].

This creates a need for a more flexible and learnable embedding strategy that can adapt to various conditions without manual tuning.

The core problem addressed in this study is:

Can a deep autoencoder, trained even on a single host image, learn to embed a watermark in a visually imperceptible manner and ensure reliable extraction after common image distortions?

Solving this problem involves assessing:

- The ability of an autoencoder to preserve imperceptibility in watermarked images,
- Its capacity to ensure robust extraction under attacks such as JPEG compression and Gaussian noise.

By focusing on this problem, the study aims to provide a foundation for data-driven watermarking models that learn embedding and extraction in an end-to-end manner, even under data-scarce conditions.

3. Methodology

3.1. Autoencoder Architecture

The system is composed of two major components [3]:

- Encoder: Accepts a host image and a watermark, resizes the watermark, and embeds it into the host image using convolutional layers.
- Decoder: Receives the watermarked image and attempts to reconstruct the original watermark.

The input watermark is a 32×32 RGB image containing the text "AI", while the host is a 128×128 RGB image. The network uses 3×3 convolutions and sigmoid activation to ensure pixel values remain between 0 and 1.

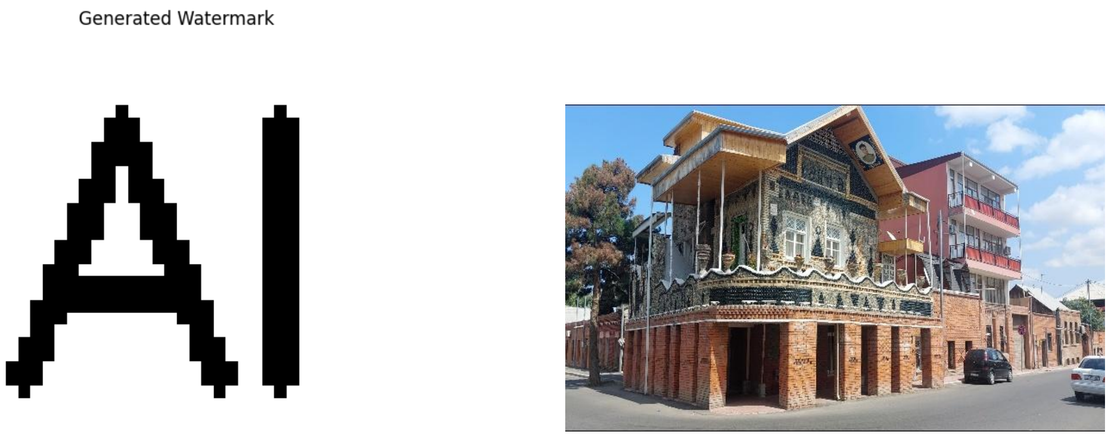

Fig. 1. Generated watermark and host image

3.2. Loss Functions

The model is trained using a combination of two mean squared error (MSE) losses:

- MSE between the original and watermarked image (for imperceptibility),
- MSE between the original and extracted watermark (for recoverability).

The total loss is a simple weighted sum of both terms.

3.3. Training Protocol

Since this is a prototype, the model is trained on a single host-watermark pair repeated 64 times to simulate a dataset. The model was trained using the Adam optimizer with a learning rate of 0.001 over 100 epochs.

4. Experiments and Results

4.1. Imperceptibility Evaluation

The evaluation process in watermarking refers to the set of techniques and metrics used to assess the effectiveness of a watermarking algorithm. The goal is to ensure that the watermark is successfully embedded in the image without significantly degrading its quality, and that it remains detectable and extractable even after the image undergoes transformations or attacks (like compression or noise) [4]. The evaluation process typically involves:

1. Assessing Image Quality: Ensuring that the watermarking process does not noticeably degrade the visual quality of the image.
2. Evaluating Watermark Detection: Checking whether the watermark can still be extracted correctly after the image is modified.
3. Testing Robustness: Evaluating how well the watermark survives attacks like compression, noise addition, or cropping.

Structural Similarity Index (SSIM)

Purpose: SSIM measures the perceptual quality of the watermarked image by comparing the structural information between the original and the watermarked image. It assesses luminance, contrast, and structure.

Formula: SSIM compares the local patterns of pixel intensities between the original and watermarked images. The formula for SSIM is:

$$\text{SSIM}=\frac{(2\mu_x\mu_y+c_1)(2\sigma_{xy}+c_2)}{(\mu_x^2+\mu_y^2+c_1)(\sigma_x^2+\sigma_y^2+c_2)}$$

Where:

- μ_x *and* μ_y are the mean pixel values of the original (x) and watermarked (y) images, respectively.
- σ_x^2 and σ_y^2 are the variances of the original and watermarked images.
- σ_{xy} is the covariance between the original and watermarked images.
- C_1 and C_2 are constants to stabilize the division with weak denominators (usually small values like $C1=0.01^2$and $C2=0.03^2$)

Interpretation: SSIM returns a value between -1 and 1, where 1 indicates that the two images are identical in terms of structure, and 0 indicates no similarity. A high SSIM value means that the watermarked image is perceptually very similar to the original image.

Mean Squared Error (MSE)

Purpose: MSE measures the average squared difference between corresponding pixel values of the original and watermarked images. It is a simple method to quantify the error or distortion introduced by the watermarking process.

$$\text{MSE}=\frac{1}{N}\sum_{i=1}^{N}(I_{original}(i)-I_{watermarked}(i))^2$$

Where:

- $I_{original}(i)$ and $I_{watermarked}(i)$ are the pixel values of the original and watermarked images, respectively.
- N is the total number of pixels in the image.

Interpretation: A lower MSE indicates better image quality (less distortion), as it means the difference between the original and watermarked images is minimal.

Peak Signal-to-Noise Ratio (PSNR)

Purpose: PSNR is a metric used to measure the quality of the watermarked image in terms of pixel-level differences. It compares the original and watermarked images based on the maximum possible pixel value (usually 255 for an 8-bit image) and the MSE (Mean Squared Error).

$$\text{PSNR}= 10\text{x}\log_{10}(\frac{R^2}{MSE})$$

Where:

- R is the maximum pixel value of the image (typically 255 for an 8-bit image).
- MSE is the Mean Squared Error between the original and watermarked images.

Interpretation: A higher PSNR indicates better quality (less distortion), meaning the watermarked image is more similar to the original. Typical PSNR values are in the range of 30-50 dB, where higher values (closer to 50) generally indicate better quality.

Fig. 2. Original host image and watermarked image

Table 1
The visual quality of the watermarked image

Metric	Value
SSIM	0.9790
PSNR	27.28 dB
MSE	0.0019

These results indicate minimal distortion, confirming that the watermarking process does not degrade visual quality.

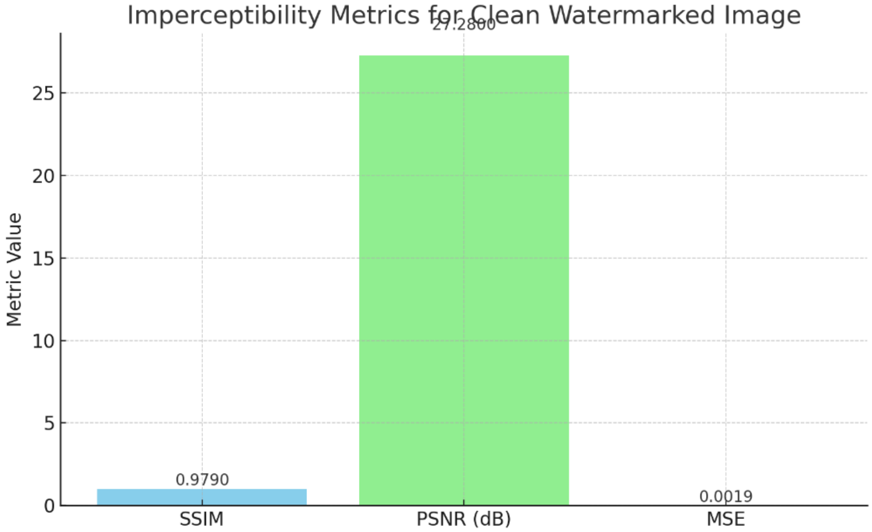

Fig. 3. Imperceptibility metrics (SSIM, PSNR, MSE)

4.2. Robustness Evaluation

There are three main categories of attacks applied to watermarked images. **Natural Distortions** occur during regular image processing or transmission, such as compression, noise, or blurring, which can degrade image quality without intent to remove the watermark. **Intentional Attacks** are specifically designed to remove or degrade the watermark, including cropping, rotation, or filtering operations. **Signal Processing Attacks** involve technical modifications to the image's signal, such as compression or noise, which affect the image's structure and can disrupt watermark detection

JPEG compression is one of the most common real-world transformations applied to digital images – whether through social media platforms, cloud storage, or content delivery systems. Since JPEG is a lossy compression algorithm, it removes information, especially in high-frequency components, which can damage or remove watermarks embedded in subtle pixel patterns.

- $Q = 40$ is chosen as it represents a moderately aggressive compression level:
- $Q > 75$ = low compression, little distortion
- $Q \approx 40$ = noticeable compression artifacts, real-world scenario
- $Q < 30$ = high compression, severe quality loss

Thus, testing at $Q = 40$ provides a realistic and moderately difficult challenge to evaluate how well the watermark survives common file saving and sharing scenarios [5], [6].

Gaussian noise simulates natural image distortions caused by camera sensors, scanning artifacts, or transmission errors. It introduces small, random pixel fluctuations across the image.

$\sigma = 0.1$ means the standard deviation of the noise is 10% of the normalized pixel range ([0, 1]).

It creates visually noticeable, but not extreme, degradation – enough to challenge the watermark without completely destroying the host image.

Using $\sigma = 0.1$ allows us to test the model's robustness to mild-to-moderate stochastic noise, which is common in real-world imaging pipelines.

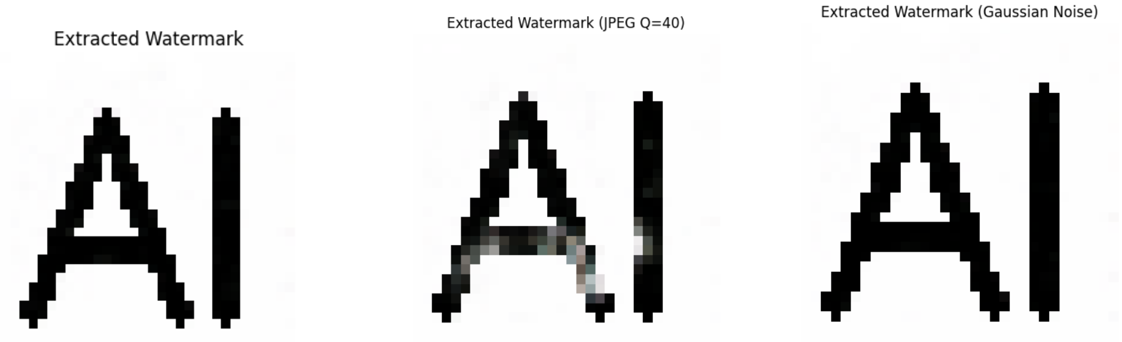

Fig. 4. Extracted watermarks before and after attacks

Table 2

The quality of the extracted watermark using correlation with the original

Attack Type	Correlation (Extracted vs Original)
Clean	1.0000
JPEG (Q=40)	0.9726
Gaussian (σ=0.1)	1.0000

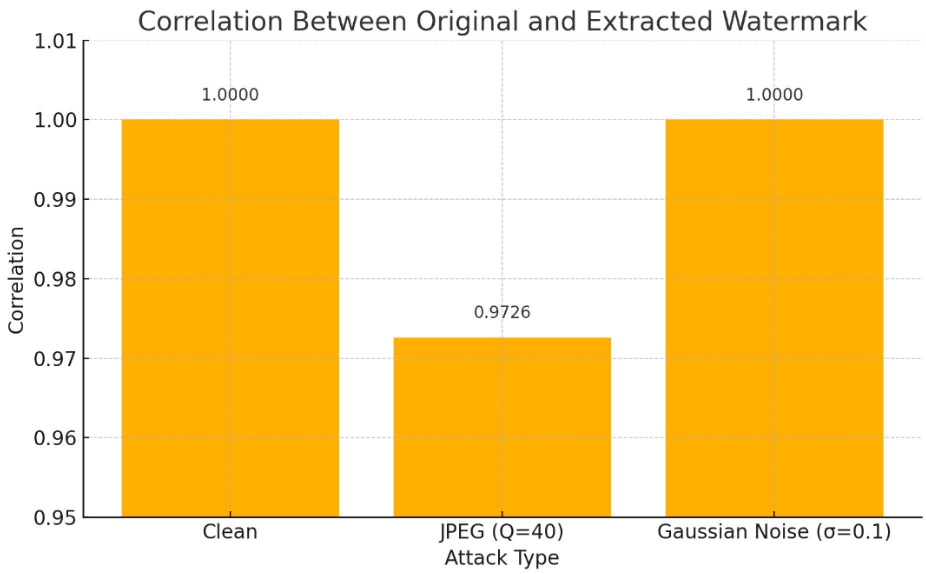

Fig. 5. Correlation between original and extracted watermark

The model demonstrates strong robustness, especially against Gaussian noise, and performs reliably even under lossy compression.

5. Discussion

This experiment confirms that even a minimally trained autoencoder can learn to imperceptibly embed and reliably recover a watermark. The exceptionally high correlation after both JPEG and Gaussian attacks suggests the autoencoder learns robust spatial features for watermark recovery. Although the model is currently overfitted to one image, it establishes a solid foundation for more scalable learning when generalized across larger datasets.

Furthermore, the network's ability to reconstruct a watermark from noisy or compressed inputs positions this approach as a viable candidate for next-generation neural watermarking systems. Future work will focus on training with multiple host images, improving generalization, and testing under more complex attack scenarios such as resizing, cropping, and adversarial perturbations.

6. Conclusion

We proposed and validated a deep learning-based invisible watermarking system using autoencoders in TensorFlow. Despite training on a single image pair, the model achieved high imperceptibility and robustness, recovering the watermark even after aggressive image distortions. This preliminary work showcases the potential of autoencoder-based watermarking and sets the stage for broader dataset training and deployment in real-world content authentication systems.

References

[1] Ayu Astridefi, Rudy A.G. Gultom,Yudistira Dwi Wardhana Asnar, H.A. Danang Rimbawa, Audio, text, image and video digital watermarking techniques for security of media digital, International journal of progressive sciences and technologies. 42 No.1 (2023) pp 389-398.

[2] Priyanka Vashist, Roop Singh, Dr.Rajeev Ratan Arora, Digital watermarking using DCT&DWT technique, International journal of research. 4 No.6 (2017).

[3] Elchaimae Rebahi, Mustapha Hemis, Bachir Boudraa, Image watermarking technique using convolutional autoencoder, 2023 International conference on advances in electronics, control and communication systems (ICAECCS).

[4] Dujan B.Taha, Taha B.Taha, Towards a new metric for watermarked image assesment, 2nd International conference on electrical, communication, computer, Power and control engineering, (2019).

[5] Mohammad Rizwan Khan, Ankur Goyal, Gaussian noise attack analysis of non-blind multiplicative watermarking using 2D-DWT, International journal of computer science and information technologies. 7 No.6 (2016) pp.2481-2486.

[6] Yuepeng Hu, Zhengyuan Jiang, Moyang Guo, Neil Gong, A transfer attack to image watermarks. In International Conference on Learning Representations (ICLR), 2025. https://openreview.net/pdf?id=UchRjcf4z7