International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V73I4P108

Volume 73 Issue 4, 81-91, April 2025

Original Article

Adapting RivaGAN for Robust Image Watermarking with Attention Mechanisms

Abdelhay Hassani Allaf[1], M'hamed Ait Kbir[2]

[1,2]*Intelligent Automation & BioMedGenomics Laboratory (IABL), STSM Doctoral Studies Center, Abdelmalek Essaadi University, Tangier, Morocco.*

[1]*Corresponding Author : a.hassani@uae.ac.ma*

Received: 21 October 2024 Revised: 15 February 2025 Accepted: 1 March 2025 Published: 26 April 2025

Abstract - *This paper presents the adaptation of the RivaGAN framework for robust image watermarking, specifically targeting image transformations such as JPEG compression, Gaussian noise, scaling, and cropping. Attention mechanisms are employed to improve watermark embedding and extraction robustness and accuracy. The proposed method incorporates a 32-bit watermark into 512 x 512 images from the CIFAR-10 dataset, including pre-and post-processing phases, to further improve performance. The effectiveness of the technique is judged using indicators such as Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) and Recovery Accuracy (RA). The results illustrate strong resilience to attacks, including JPEG compression and scaling, with negligible visual deterioration and excellent accuracy in watermark detection. However, the system demonstrates vulnerability to heavy Gaussian noise and cropping, where recovery accuracy significantly drops. Additionally, we evaluate the effect of pre-and post-processing on system performance under Gaussian noise conditions, highlighting their benefits in mitigating these vulnerabilities.*

Keywords - *Watermarking, RivaGAN, Robustness, Adversarial Networks, Attention Mechanism*.

1. Introduction

With the rapid development of new communication technologies, the use, transmission, and restructuring of digital content are increasing, making the security of multimedia exchanges challenging due to unauthorized use, falsification, or copyright infringement. Digital watermarking has been widely studied to protect media data from illegal use in recent years. Developing advanced protection methods is essential to preserve digital data exchanges. Watermarking is a common solution that embeds hidden messages or information (called a watermark) into the original image (cover data). The watermarked image should resemble the original (imperceptibility), contain as much information as possible (capacity), and be robust against various attacks (transformations). Watermarking applications span various fields, including securing medical images in telemedicine [1, 2], protecting digital media, tracking broadcast content [3], and authenticating IoT network data [4]. Despite these uses, many existing methods struggle to balance robustness, invisibility, and computational efficiency, particularly under complex transformations. The limitations of traditional watermarking methods against transformation attacks have directed research towards more robust systems. Generative Adversarial Networks (GANs) have emerged as a promising tool to improve the robustness of watermarking, providing the ability to learn complex data patterns and generate realistic results. RivaGAN has demonstrated success in video watermarking using convolutional networks, attention mechanisms, and adversarial training. Nevertheless, its application for image watermarking remains underexplored. Unlike video watermarking, which benefits from temporal redundancy, image watermarking relies solely on spatial characteristics, making it more sensitive to distortions. This paper addresses this gap by adapting RivaGAN for robust image watermarking. Using the CIFAR-10 dataset, a 32-bit watermark is embedded into 512x512 images, and the system's performance is evaluated under transformations such as JPEG compression, Gaussian noise, scaling, and cropping. This work integrates attention mechanisms to enhance robustness and employs pre- and post-processing techniques to improve performance under noisy conditions. This work introduces several key contributions:

- Adaptation of RivaGAN with the integration of attention mechanisms for image watermarking to demonstrate its potential to manage single-frame transformations.
- A detailed analysis of the system's performance against various attacks, highlighting its strengths and limitations with the use of metrics such as Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Recovery Accuracy (RA).
- Pre- and post-processing techniques to improve system performance under noisy conditions.

The remainder of this paper is organized as follows. Section 2 reviews the state of the art in watermarking techniques, covering both traditional and deep learning-based methods. Section 3 outlines the methodology, including the pre-processing, watermark embedding, post-processing, and transformations applied to test the system's robustness. Section 4 presents and discusses the results, focusing on the effects of various attacks on PSNR, SSIM, and recovery accuracy. Finally, Section 5 concludes the paper and suggests future directions for improving the system's robustness, particularly against more complex transformations.

2. Related Work

In recent years, image watermarking has evolved significantly, evolving from traditional methods based on simple algorithms to sophisticated strategies based on deep learning. In this section, we examine the main watermarking methods, with a particular emphasis on recent studies dealing with issues of robustness and imperceptibility.

2.1. Traditional Watermarking Methods

Traditional image watermarking methods can basically be divided into two groups: those that rely on the spatial domain and those that rely on the frequency domain. These methods have been used for decades and form the basis of recent advances in watermarking.

2.1.1. Spatial Domain Methods

Least Significant Bit (LSB) replacement is one of the oldest and most basic techniques for inserting watermarks into images [5]. This method embeds the watermark in the least significant bits of the pixel values. While LSB is computationally efficient and produces minimal visual distortion, it is highly vulnerable to basic image transformations such as compression, cropping, and noise addition [6]. Content-based approaches increase their robustness by inserting watermarks in key regions of the image, such as edges, textures or feature points. These techniques use image characteristics, such as corners identified by the Harris detector for example, to embed watermarks in areas less likely to be affected by image changes, thereby ensuring a high recovery accuracy and minimal visual distortion.

2.1.2. Frequency Domain Methods

Stronger than spatial domain-based methods, frequency domain techniques incorporate the watermark into the transformed coefficients of the image. The use of watermarking based on Discrete Cosine Transform (DCT) [7] is widespread due to its robustness against compression, especially for JPEG images. Techniques based on Discrete Wavelet Transform (DWT) [8] are frequently used thanks to their ability to represent multi-resolution features, which makes them more robust to a greater diversity of transformations. DWT combined with Singular Value Decomposition (SVD) has gained popularity for providing robustness while maintaining good image quality [9]. Further refinements, such as the use of Discrete Fourier Transform (DFT) [10], offer additional resilience to geometric distortions like rotation, but these methods still suffer from limited robustness against scaling and cropping. Despite these improvements, traditional methods often face a trade-off between robustness and invisibility. The watermark's resistance to removal usually comes at the cost of noticeable visual degradation, especially when more aggressive embedding is required to withstand geometric attacks [11].

2.2. Deep Learning-Based Watermarking

The advent of deep learning has introduced new possibilities for image watermarking, offering more adaptive and sophisticated models that can balance robustness and invisibility more effectively than traditional methods.

2.2.1. CNN-Based Watermarking

One of the first uses of deep learning in the field of image watermarking was based on Convolutional Neural Networks (CNN), which were used for both embedding and extraction operations. The HiDDeN framework [12] was one of the first to show how deep neural networks can be employed to hide information in high-quality images. However, this model faced robustness challenges to common image transformations like resizing and rotation, which continue to pose difficulties for CNNs trained on pixel-level data [13]. Recent studies have extended the use of CNNs in image watermarking by combining them with traditional methods like DWT or DCT to improve robustness [14]. These hybrid approaches use CNNs to learn features more resistant to noise and compression, while the frequency domain techniques handle geometric distortions.

2.2.2. Autoencoders and Unsupervised Learning

Autoencoders, another type of deep neural network, have been employed in image watermarking to automatically learn compact representations of both the image and watermark data. Autoencoder-based models have shown promise in reducing computational complexity while maintaining robustness [15]. Furthermore, they have been applied in conjunction with CNNs to improve watermark recovery in scenarios where the image has been compressed or attacked [16].

2.2.3. GAN-Based Watermarking

Implementing adversarial learning has transformed the watermarking field through Generative Adversarial Networks (GANs). GANs consist of two competing networks: a generator that inserts the watermark and a discriminator dedicated to its detection. This conflicting configuration forces the generator to create more undetectable and resilient watermarks. Zhang et al. [17] employed GANs for video watermarking, demonstrating notable improvements in robustness to compression and noise. GANs have also been used to improve image watermarking. In [18], the authors

combined GANs with an encoder-decoder architecture, achieving superior robustness against various image manipulations while maintaining imperceptibility. Despite these successes, GAN-based approaches often require extensive training and are computationally expensive, limiting their practical application [19].

More current research has delved deeper into GAN-based watermarking techniques. Guangyong Gao et al., (2024) [20] proposed an enhanced GAN model designed for improving resilience against adversarial attacks, while Debolina Mahapatra et al., (2023) [21] presented a hybrid GAN-autoencoder approach that strengthens the watermark embedding process, ensuring robustness against noise and compression attacks. These studies highlight the continuous development of GAN-based watermarking techniques, improving their practical use in real-world scenarios.

2.3. Attention Mechanisms in Watermarking

In computer vision, attention mechanism techniques are increasingly common due to their ability to focus attention on specific areas of an image. In the watermarking domain, attention mechanisms can improve robustness and invisibility by precisely embedding the watermark in areas less likely to experience distortion. The first works introducing attention mechanisms into video watermarking were those of Zhang et al. [22]. They demonstrated that attention-based embedding can increase resistance to common video transformations like compression and resizing. Attention-based watermarking for images is still new, but initial research indicates that it could strengthen the strength of the watermark by emphasizing textured areas of the image where the watermark is less likely to be affected by manipulations like resizing and cropping [23]. More recent work by Yimeng Zhao et al. (2022) [24] and Jiren Zhu et al. (2017) [14] has explored attention mechanisms in high-resolution watermarking, showing that attention maps can significantly improve robustness by embedding watermarks into areas of an image that are less likely to be affected by transformations. These studies are highly relevant for integrating attention mechanisms in image watermarking systems, such as the adaptation of RivaGAN in this paper.

2.4. Adversarial Training for Robust Watermarking

Adversarial training has established itself as an effective lever in the deep learning sector, particularly to strengthen the resilience of models against attacks. In the context of watermarking, adversarial training consists of simulating authentic attacks such as noise, compression or cropping during the training phase, thus pushing the model to develop more robust integration strategies. In recent work by Vukotić et al. [25], adversarial networks were employed to simulate attacks aimed at removing the watermark. The generator network learns to embed the watermark in a way that resists these attacks while the adversary tries to modify the watermarked image without leaving visible traces. This setup has been shown to improve the robustness of the watermark without sacrificing image quality. Recent research by Javni Thakkar et al. (2022) [26] and Jianbo Chen et al. (2023) [27] looked at better adversarial training methods for watermarking, focusing on improving resistance to cropping, noise, and digital tampering. This progress highlights the importance of adversarial networks in reinforcing the sustainability of marked content in different attack contexts.

2.5. Limitations of Existing Approaches

Despite these advancements, several challenges remain unsolved in the field of image watermarking.

2.5.1. Fragility to Geometric Transformations

While frequency domain techniques are robust against compression, their resistance to geometric attacks such as scaling and rotation is still limited. Deep learning models, though more adaptable, are often vulnerable to these transformations unless specifically trained to handle them [28].

2.5.2. Computational Complexity

While highly effective in creating imperceptible and robust watermarks, GAN-based approaches often require significant computational resources. The training process can be slow and expensive, making these models impractical for real-time applications [29].

2.5.3. Under-Explored Attention Mechanisms

While attention mechanisms have shown great promise in other computer vision tasks, their application in image watermarking remains underexplored. More research is needed to fully leverage their ability to dynamically focus on robust regions of the image, particularly for resisting advanced adversarial attacks [30].

3. Methodology

In this paper, we adapt the RivaGAN framework, originally designed for robust video watermarking, to an image watermarking task using the CIFAR-10 dataset. RivaGAN leverages adversarial networks and attention mechanisms to embed data securely, ensuring the watermark is robust against various image transformations. We focus on embedding a 32-bit watermark into images resized to 512x512 pixels and evaluate the system's resilience against common image distortions and attacks.

3.1. RivaGAN Architecture Overview

RivaGAN's architecture for image watermarking consists of the following components:

3.1.1. Encoder

Responsible for embedding the watermark into the image. The encoder generates a residual mask applied to the image, embedding the watermark while keeping it visually imperceptible.

3.1.2. Decoder

Extracts the watermark from the watermarked image, using a convolutional network and attention mechanism to identify and decode the watermark from the image.

3.1.3. Attention Mechanism

Focuses on specific regions of the image for the integration and decoding process, thereby strengthening resistance by directing watermark integration towards less vulnerable areas. This helps minimize the effect of changes like cropping and scaling.

3.1.4. Adversarial Training

In the original RivaGAN model, adversarial training is employed to improve robustness. A discriminator tries to detect watermarked images, forcing the encoder to hide the watermark more resiliently. In this adaptation, adversarial training is deferred to future work.

3.2. Image Watermarking Process

The RivaGAN framework was adapted to embed a 32-bit watermark into images from the CIFAR-10 dataset (Figure 1). The images are resized to 512x512, ensuring uniformity across the test cases. The following stages summarize the process:

3.2.1. Pre-Processing

Before embedding the watermark, the images are pre-processed:

Image Resizing: All CIFAR-10 images are scaled to 512 x 512 pixels to accommodate the integration of the watermark.

Gaussian Filtering: A Gaussian smoothing filter is applied to the resized image to reduce high-frequency noise and prepare the image for watermark embedding. This pre-processing step improves the system's resilience against noisy conditions.

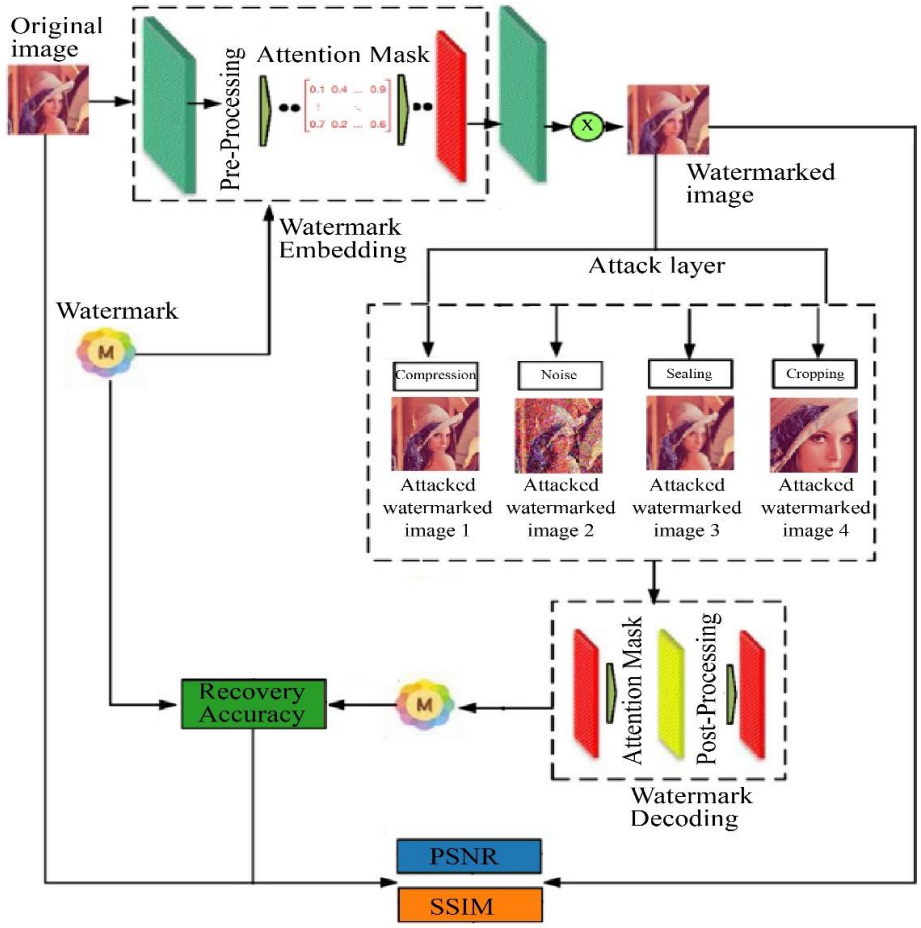

Fig. 1. General scheme of the watermarking system using the adapted RivaGAN Framework

This figure illustrates the watermark embedding and decoding processes with attention mechanisms and pre- and post-processing steps, demonstrating the system's robustness against various attacks such as compression, noise, scaling, and cropping. The figure also highlights the evaluation metrics, PSNR, SSIM, and Recovery Accuracy, used to assess the system's performance after the watermark extraction.

3.2.2. Watermark Embedding

Watermark embedding uses the RivaGAN encoder, a convolutional network enhanced with an attention mechanism. The watermark is embedded through a residual mask, ensuring minimal distortion of the original image:

$$I_w = I + \alpha \cdot tanh\big(f(I,W)\big) \quad (1)$$

Where:

- I_w is the watermarked image, and I is the original image.
- W is the 32-bit watermark.
- $f(I,W)$ is a function applied by the encoder's convolutional layers.
- α is a small scaling factor to keep distortions minimal.

3.2.3. Watermark Decoding

The watermark extraction process uses the decoder, which recovers the watermark even after transformations:

$$W' = D(I_w) = Pool\big(f(I_w) \cdot Attention(I_w)\big) \quad (2)$$

Where:

- W' is the extracted watermark, and D is the decoder network.
- Pooling aggregates spatial features to recover the watermark, with the attention mechanism guiding the decoding.

3.3. Robustness Testing with Image Transformations

We subjected the watermarked images to various transformations simulating real-world distortions to evaluate the system's robustness. The following transformations were applied:

- JPEG compression of quality 90, 70 and 50.
- Gaussian noise was added with variations of 10, 25 and 50.
- Scaling using factors 0.9, 0.8 and 0.7.
- Crop with proportions of 10%, 20% and 30%.

These transformations are commonly used to simulate potential attacks on watermarked images. By evaluating the system's performance under these scenarios, we aimed to determine its robustness.

3.4. Metrics for Evaluation

We based ourselves on the following criteria to judge the effectiveness of our watermark system:

3.4.1. Peak Signal-to-Noise Ratio (PSNR)

PSNR evaluates the quality of the marked image in comparison with the original. Equation 3 illustrates how to calculate it.

$$PSNR = 10 \cdot log_{10}\left(\frac{MAX_I^2}{MSE}\right) \quad (3)$$

Where: MAX_I is the maximum possible pixel value.

- MSE is the mean squared error between the original and watermarked image.

3.4.2. Structural Similarity Index Measure (SSIM)

SSIM compares the structural similarity between the original and watermarked images, accounting for luminance and contrast changes. SSIM is given by:

$$SSIM(I, I_w) = \frac{(2\mu_I\mu_{I_w} + c_1)(2\sigma_{I\,I_w} + c_2)}{\big(\mu_I^2 + \mu_{I_w}^2 + c_1\big)\big(\sigma_I^2 + \sigma_{I_w}^2 + c_2\big)} \quad (4)$$

Where: μ_I and μ_{I_w} are the means of the original and watermarked images.

- σ_I and σ_{I_w} are their variances.

- c_1 and c_2 are constants to stabilize the equation, avoiding division by zero.

3.4.3. Recovery Accuracy

Recovery accuracy measures the percentage of correctly recovered watermark bits:

$$Accuracy = \left(\frac{Number\ of\ Correct\ Bits}{Total\ Bits}\right) \times 100 \quad (5)$$

Where :

- The $Number\ of\ Correct\ Bits$ refers to the number of watermark bits correctly recovered by the decoder.

This evaluates how well the decoder can extract the original watermark after transformations.

3.5. Pre-Processing and Post-Processing for Robustness Enhancement

To make the watermark system more robust, we applied pre-processing before embedding and post-processing after watermark extraction:

Pre-processing (Gaussian smoothing): An original image is processed with a Gaussian filter to reduce high-frequency noise, which helps reduce the impact of noise and optimizes the watermark's embedding capacity. Smoothing is particularly useful for handling transformations such as Gaussian noise.

Post-Processing (Median Filtering): After watermark extraction, a median filter is applied to reduce residual noise. The median filtering helps to restore image quality by removing outliers, such as noise introduced by Gaussian transformations, enhancing the watermark recovery process.

3.6. CIFAR-10 Dataset and Testing Setup

The CIFAR-10 dataset was selected to evaluate the performance of the proposed watermarking system. CIFAR-10 consists of 60,000 color images in 10 classes; in this work, all images were resized to 512x512 to fit the input requirements of the watermarking system.

For testing, 1000 images were randomly selected from the CIFAR-10 test set and used for initial watermark embedding and decoding without any transformations. Following this, a separate set of 100 images was selected for robustness testing, where the aforementioned transformations (JPEG compression, Gaussian noise, scaling, and cropping) were applied to assess how well the system could extract watermarks under different attack conditions.

4. Results and Discussion

In this section, we evaluate the performance of our watermarking system based on the RivaGAN architecture for embedding a 32-bit watermark in images from the CIFAR-10 dataset.

The images used for the evaluation are sized at 512x512 pixels, and we assess the system's robustness using three metrics: PSNR (dB), SSIM, and Recovery Accuracy (%). The testing was conducted in two main phases: first, without any attacks, and second, with a set of common image transformations that act as attacks.

4.1. Performance Without Attacks

In the first testing phase, we evaluated the system's performance on a dataset of 1000 images from CIFAR-10 without applying any attacks. After embedding the watermark, we measured the PSNR, SSIM, and Recovery Accuracy upon decoding the watermark.

Below are the average results of the system's performance without attacks.

- PSNR: 40.72 dB
- SSIM: 0.9734
- Recovery Accuracy: 98.42%

The PSNR value of 40.72 dB demonstrates that the watermarked images maintain excellent visual quality, while the SSIM of 0.9734 reveals very low structural distortion. A high recovery rate of 98.42% demonstrates the system's proficiency in accurately deciphering the embedded watermark.

It should be clarified that the average recovery accuracy was determined considering only those cases where a total recovery of 100% was obtained. Additionally, the distribution of PSNR, SSIM, and Recovery Accuracy across the dataset is depicted in the distribution graphs below.

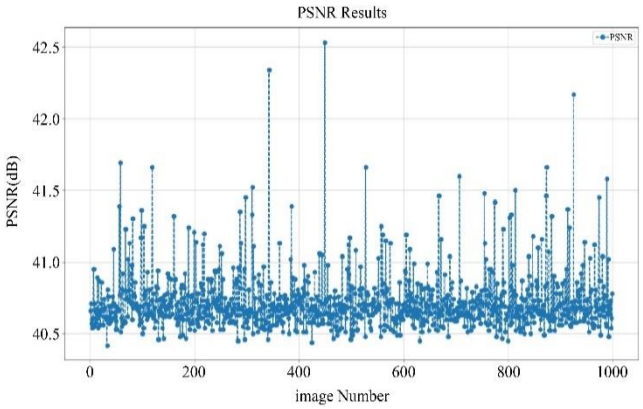

Fig. 2 Distribution of PSNR values across 1000 images without any applied attacks

This graph shows the retention of visual quality after embedding the watermark.

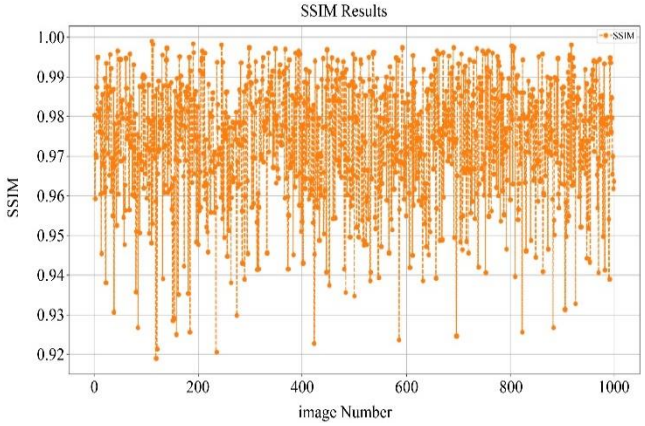

Fig. 3 Distribution of SSIM values across 1000 images without any applied attacks

This graph highlights the structural similarity maintained between the original and watermarked images.

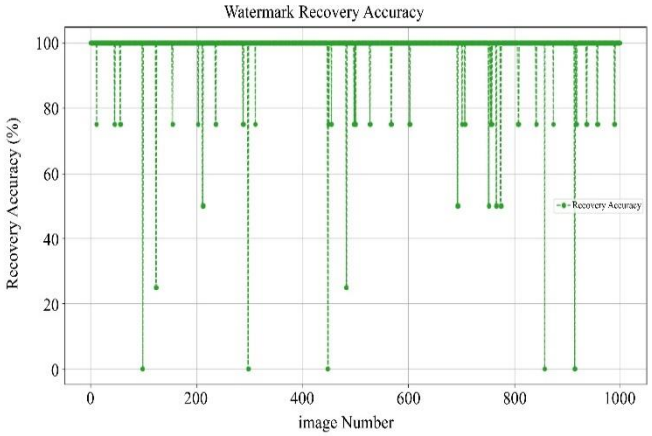

Fig. 4 Distribution of recovery accuracy percentages across 1000 images without attacks

The graph indicates the system's ability to successfully extract the embedded watermark accurately.

4.2. Performance with Attacks

In the second phase, we subjected the system to a series of common image transformations to test its robustness against attacks.

We applied the following four types of attacks to a set of 100 images:

- JPEG compression at three levels: 90, 70, and 50
- Gaussian Noise with variances of 10, 25, and 50
- Scaling factors of 0.9, 0.8, and 0.7
- Cropping with percentages of 0.1, 0.2, and 0.3

We measured the impact on the watermarked images for each transformation by calculating the PSNR, SSIM, and Recovery Accuracy after decoding the watermark.

The following table 1 summarizes the average results for each transformation level.

Table 1. Average PSNR, SSIM, and recovery accuracy for different transformations and levels

Attack Type	Level	PSNR (dB)	SSIM	Recovery Accuracy (%)
JPEG Compression	90	41.31	0.98	95.05
	70	41.13	0.98	68.32
	50	40.66	0.98	4.95
Gaussian Noise	10	27.98	0.71	62.38
	25	20.45	0.39	0.00
	50	14.93	0.19	0.00
Scaling	0.9	40.92	0.98	98.02
	0.8	40.96	0.98	97.03
	0.7	41.01	0.98	98.02
Cropping	0.1	16.92	0.76	74.26
	0.2	13.69	0.62	0.00
	0.3	12.35	0.54	0.00

This table summarizes the performance of the watermarking system under various transformation types (JPEG Compression, Gaussian Noise, Scaling, and Cropping) at different levels.

It illustrates how each attack impacts image quality and the system's ability to recover the embedded watermark, with metrics presented as PSNR, SSIM, and Recovery Accuracy percentages.

To better visualize the impact of the transformations on image quality and watermark recovery, the following graphs illustrate the distribution of PSNR, SSIM, and Recovery Accuracy across different transformations and transformation levels:

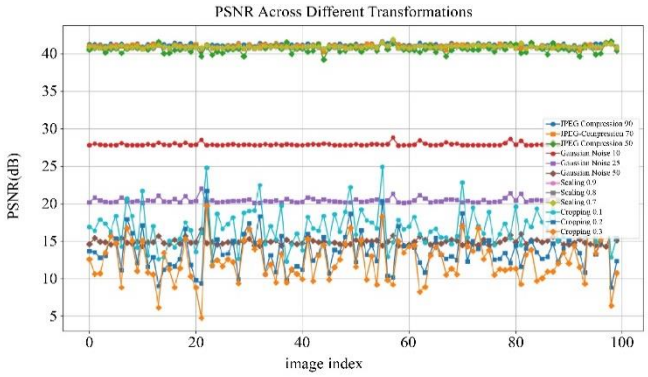

Fig. 5 PSNR versus various transformations

This graph visualizes the degradation in image quality (PSNR) under different transformations, including JPEG compression, Gaussian noise, scaling, and cropping. It highlights the significant drop in PSNR values for transformations involving high levels of Gaussian noise and cropping.

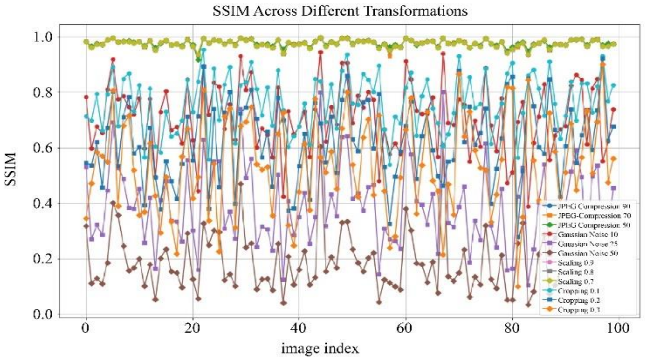

Fig. 6 SSIM versus various transformations

The graph presents the structural similarity index (SSIM) for various transformations, indicating how different levels of JPEG compression, Gaussian noise, scaling, and cropping affect the structural quality of the watermarked images. Gaussian noise and cropping, especially at higher levels, significantly reduce SSIM.

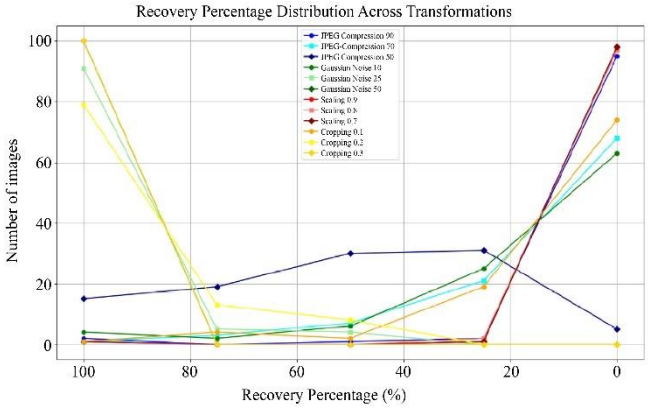

Fig. 7 Recovery Percentage Distribution versus various transformations

This graph shows how the watermark recovery accuracy changes with different transformations. Scaling consistently maintains high recovery accuracy, while Gaussian noise and cropping cause steep declines, particularly at higher intensities of the attacks.

4.3. Performance with Pre- and Post-Processing

In addition to the general test with and without any attacks, we conducted a separate evaluation on six selected images (images below) with and without pre- and post-processing to analyze their effect on robustness against Gaussian Noise (variance 10). The performance was evaluated on normal and processed images using both PSNR, SSIM, and Recovery Accuracy metrics.

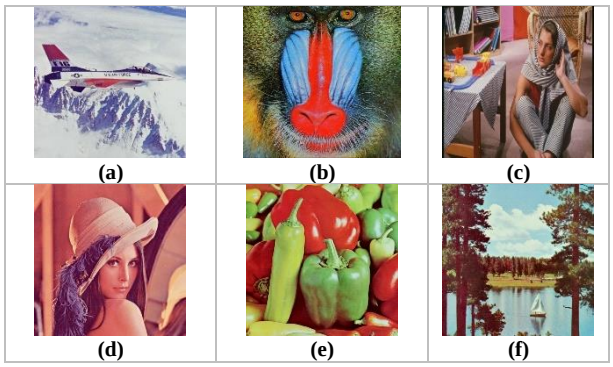

Fig. 8 The host images used in Pre- and Post-Processing Tests: (a) Airplane, (b) Baboon, (c) Barbara, (d) Lena, (e) Pepper, and (f) Sailboat

4.3.1. Without Noise

The PSNR and SSIM values for processed images are slightly lower than the normal images, but the Recovery Accuracy improves to 100% when pre- and post-processing is applied.

4.3.2. With Noise 10 (Variance 10)

The presence of noise leads to a more pronounced degradation of the PSNR and SSIM values of the processed images, highlighting the difficulty of maintaining excellent visual quality in a noisy context.

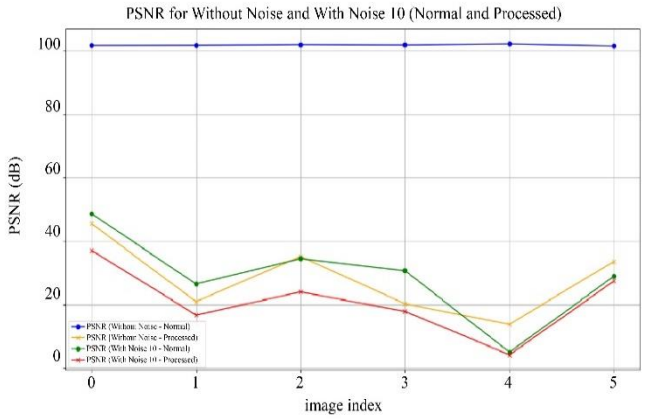

Fig. 9 PSNR comparison for normal and processed images with and without gaussian noise

This graph illustrates the Peak Signal-to-Noise Ratio (PSNR) across six selected images, comparing normal and processed images both with and without Gaussian Noise (variance 10). It highlights how visual quality changes with pre- and post-processing.

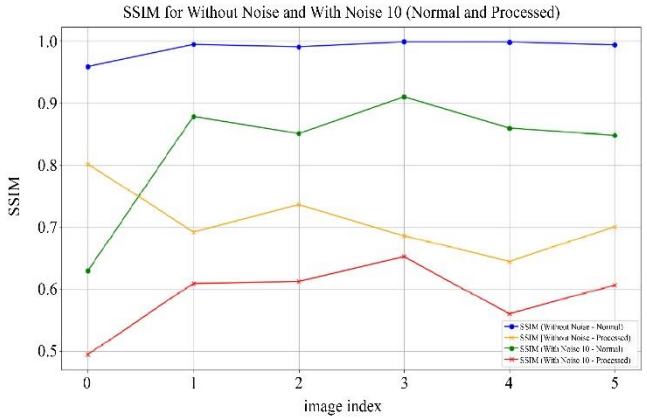

Fig. 10 SSIM comparison for normal and processed images with and without gaussian noise

This graph shows the Structural Similarity Index (SSIM) across the six images, visualizing the impact of Gaussian Noise (variance 10) and the effects of pre-and post-processing on image structure and quality.

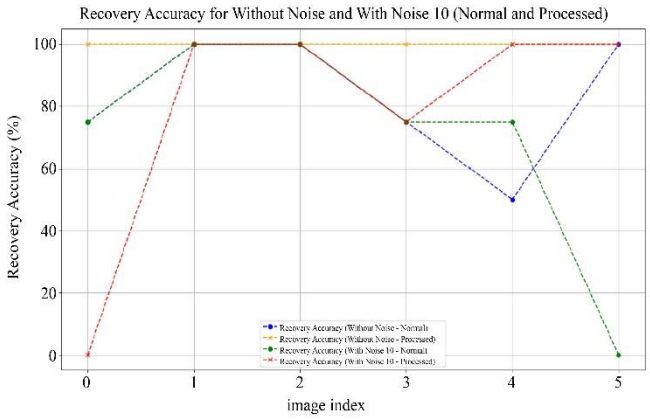

Fig. 11 Recovery accuracy for normal and processed images with and without gaussian noise

This graph depicts the Recovery Accuracy for watermark extraction in normal and processed images, both with and without Gaussian Noise (variance 10). It emphasizes the improvements in Recovery Accuracy after applying pre- and post-processing, especially for cases where the normal method did not achieve full recovery.

Nevertheless, the recovery accuracy significantly improves processed images, reaching up to 100% in situations where normal images have not fully recovered the watermark. More specifically, images that did not achieve 100% retrieval accuracy under normal conditions (such as image e and image

f) managed to achieve 100% retrieval accuracy after applying pre-processing and post-processing. This highlights the value of processing methods to enhance noise resistance, although this may affect visual quality (PSNR and SSIM). The graphs below visually represent PSNR, SSIM, and Recovery Accuracy changes across the six images tested.

4.4. Discussion of Results

The results demonstrate that the watermarking scheme is highly effective in scenarios without attacks, achieving PSNR values above 40 dB and an average Recovery Accuracy of 98.42%. This confirms the minimal visual degradation and high retrieval accuracy in normal conditions. However, when subjected to common image transformations, the system's performance shows varying degrees of robustness depending on the attack type and intensity.

4.4.1. JPEG Compression

The system handles light compressions (90 and 70) well, maintaining high PSNR and SSIM indices and managing to extract the watermark in most situations. However, when the compression ratio reaches 50, the recovery accuracy drops sharply to 4.95%.

4.4.2. Gaussian Noise

The system's tolerance to noise is limited, especially at high levels. With a noise variation of 25 to 50, the accuracy of the recovery drops to zero percent, meaning that the noise significantly impairs the inserted watermark.

4.4.3. Scaling

The watermark model demonstrates high resilience against downscaling, with extraction accuracy remaining high even when the image is downscaled to 70%.

4.4.4. Cropping

Cropping has a significant impact on recovery accuracy, especially at higher cropping rates. While the system achieves a recovery accuracy of 74.26% with 10% cropping, its performance drops to 0% with 20% and 30% cropping.These results suggest that although the watermarking system performs very well under standard conditions and demonstrates high robustness to some modifications (such as JPEG compression and scaling), it is more susceptible to severe noise and cropping. These data are essential to understand the strengths and limitations of the system in practical image protection cases.

5. Conclusion and Perspectives

In this paper, we illustrated the successful adaptation of the RivaGAN framework for image watermarking tasks using the CIFAR-10 dataset. The system demonstrates high robustness against JPEG compression and scaling transformations, recording high PSNR, SSIM and recovery accuracy scores. However, the system performance degrades when faced with more violent attacks, such as high levels of Gaussian noise and extensive cropping. Although the robustness-enhancing RivaGAN architecture incorporates adversarial training, there are still opportunities to improve handling these more complex transformations.

Experiments indicate that before and after processing methods, including Gaussian blurring and median filtering, contribute minimal improvements in sound attack detection accuracy. However, additional actions are required to significantly optimize system performance in these circumstances. To further strengthen the solidity of the watermarking system, future research could focus on the following areas:

5.1. Enhanced Attention Mechanisms

The current attention mechanism in RivaGAN can be refined to selectively focus on texture-rich regions of the image that are less affected by transformations like noise and cropping. Advanced attention models could improve the accuracy of watermark embedding and extraction in these areas.

5.2. Improved Pre-Processing and Post-Processing Techniques

Despite the benefits of Gaussian smoothing and median filtering, further studies could explore more advanced pre- and post-processing methods to enhance robustness against intense noise and geometric deformations.

5.3. Exploration of Hybrid Models

Integrating frequency domain methods, such as Discrete Wavelet Transform (DWT), with RivaGAN's deep learning approach could improve resistance to common attacks, including high-frequency noise and compression artifacts.

5.4. Diverse Image Testing

Extending the evaluation to more diverse image datasets, including higher-resolution and complex real-world images, would better understand the system's generalizability and performance in practical applications

References

[1] Priyanka Singh et al., "Robust and Secure Medical Image Watermarking for Edge-Enabled e-Healthcare," *IEEE Access*, vol. 11, pp. 135831-135845, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[2] Dalel Bouslimi et al., "A Joint Encryption/Watermarking System for Verifying the Reliability of Medical Images," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 5, pp. 891-899, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[3] Souha Mansour, Saoussen Ben Jabra, and Ezzedine Zagrouba, "A Robust Deep Learning-Based Video Watermarking Using Mosaic Generation," *VISIGRAPP*, vol. 5, pp. 668-675, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[4] Aidin Ferdowsi, and Walid Saad, "Deep Learning-Based Dynamic Watermarking for Secure Signal Authentication in IoT Things," *IEEE International Conference on Communications*, Kansas City, MO, USA, pp. 1-6, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[5] J.R. Hernandez, M. Amado, and F. Perez-Gonzalez, "DCT-Domain Watermarking Techniques for Still Images: Detector Performance Analysis and a New Structure," *IEEE Transactions on Image Processing*, vol. 9, no. 1, pp. 55-68, 2000. [CrossRef] [Google Scholar] [Publisher Link]

[6] H. Oktay Altun et al., "Optimal Spread Spectrum Watermark Embedding via a Multistep Feasibility Formulation," *IEEE Transactions on Image Processing*, vol. 18, no. 2, pp. 371-387, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[7] Frank Hartung, and Bernd Girod, "Watermarking of Uncompressed and Compressed Video," *Signal Processing*, vol. 66, no. 3, pp. 283-301, 1998. [CrossRef] [Google Scholar] [Publisher Link]

[8] Sneha Kadu et al., "Discrete Wavelet Transform-Based Video Watermarking Technique," *International Conference on Microelectronics, Computing and Communications*, Durgapur, India, pp. 1-6, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[9] Ashish M. Kothari, and Ved Vyas Dwivedi, "Transform Domain Video Watermarking: Design, Implementation and Performance Analysis," *International Conference on Communication Systems and Network Technologies*, Rajkot, Gujarat, India, pp. 133-137, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[10] S. Pereira et al., "Template-Based Recovery of Fourier-Based Watermarks Using Log-Polar and Log-Log Maps," *Proceedings IEEE International Conference on Multimedia Computing and Systems*, Florence, Italy, vol. 1, pp. 870-874, 1999. [CrossRef] [Google Scholar] [Publisher Link]

[11] Md. Asikuzzaman, and Mark R. Pickering, "An Overview of Digital Video Watermarking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 9, pp. 2131-2153, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[12] Jiren Zhu et al., "HiDDeN: Hiding Data with Deep Networks," *Proceedings of the European Conference on Computer Vision*, pp. 657-672, 2018. [Google Scholar] [Publisher Link]

[13] Xinyu Weng et al., "Convolutional Video Steganography with Temporal Residual Modeling," *arXiv Preprint*, pp. 1-11, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[14] S. Biswas, S.R. Das, and E.M. Petriu, "An Adaptive Compressed MPEG-2 Video Watermarking Scheme," *IEEE Transactions on Instrumentation and Measurement*, vol. 54, no. 5, pp. 1853-1861, 2005. [CrossRef] [Google Scholar] [Publisher Link]

[15] Dajun He, Qibin Sun, and Qi Tian, "A Semi-Fragile Object-Based Video Authentication System," *IEEE International Symposium on Circuits and Systems*, Bangkok, Thailand, vol. 3, pp. 111-111, 2003. [CrossRef] [Google Scholar] [Publisher Link]

[16] Nie Jie, and Wei Zhiqiang, "A New Public Watermarking Algorithm for RGB Color Image Based on Quantization Index Modulation," *International Conference on Information and Automation*, Zhuhai, Macau, pp. 837-841, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[17] Kangli Hao, Guorui Feng, and Xinpeng Zhang "Robust Image Watermarking based on Generative Adversarial Network," *China Communications*, vol. 17, no. 11, pp. 131-140, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[18] Matthew Tancik, Ben Mildenhall, and Ren Ng, "Stegastamp: Invisible Hyperlinks in Physical Photographs," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2117-2126, 2020. [Google Scholar] [Publisher Link]

[19] Iain E. Richardson, *The H.264 Advanced Video Compression Standard*, Wiley Publishing, 2nd ed., pp. 1-352, 2010. [Google Scholar] [Publisher Link]

[20] Guangyong Gao, Tianyou Xu, and Feng Hua, "Robust Image Watermarking Based on Generative Adversarial Networks for Copyright Protection," *Research Square*, pp. 1-27, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[21] Debolina Mahapatra et al., "Autoencoder-Convolutional Neural Network-based Embedding and Extraction Model for Image Watermarking," *Journal of Electronic Imaging*, vol. 32, no. 2, pp. 1-15, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[22] Kevin Alex Zhang et al., "Robust Invisible Video Watermarking with Attention," *arXiv Preprint*, pp. 1-11, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[23] Jamie Hayes, and George Danezis, "Generating Steganographic Images via Adversarial Training," *Advances in Neural Information Processing Systems*, vol. 30, pp. 1-10, 2017. [Google Scholar] [Publisher Link]

[24] Yimeng Zhao et al., "DARI-Mark: Deep Learning and Attention Network for Robust Image Watermarking," *Mathematics*, vol 11, no. 1, pp. 1-16, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[25] Vedran Vukotić, Vivien Chappelier, and Teddy Furon, "Are Deep Neural Networks Good for Blind Image Watermarking?," *IEEE International Workshop on Information Forensics and Security*, Hong Kong, China, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[26] Janvi Thakkar, Giulio Zizzo, and Sergio Maffeis "Elevating Defenses: Bridging Adversarial Training and Watermarking for Model Resilience," *arXiv Preprint*, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[27] Jianbo Chen et al., "Universal Watermark Vaccine: Universal Adversarial Perturbations for Watermark Protection," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2322-2329, 2023. [Google Scholar] [Publisher Link]

[28] Aidin Ferdows, and Walid Saad, "Deep Learning-Based Dynamic Watermarking for Secure Signal Authentication in the Internet of Things," *IEEE International Conference on Communications*, Kansas City, MO, USA, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[29] Santi P. Maity, and Seba Maity, "Multistage Spread Spectrum Watermark Detection Technique Using Fuzzy Logic," *IEEE Signal Processing Letters*, vol. 16, no. 4, pp. 245-248, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[30] Mathias Schlauweg et al., "Self-Synchronizing Robust Texel Watermarking in Gaussian Scale-Space," *Proceedings of the 10th ACM Workshop on Multimedia and Security*, Oxford United Kingdom, pp. 53-62, 2008. [CrossRef] [Google Scholar] [Publisher Link]