Invisible Watermarking with DWT-SVD for Safeguarding Copyrighted Images against Unauthorized Generative AI Training*

Seo-Yi Kim, Na-Eun Park, and Il-Gu Lee†

Sungshin Women's University, Seoul, South Korea
{sykim.cse, nepark.cse, iglee19}@gmail.com

Abstract

With advancements in artificial intelligence (AI) technologies, copyright infringement of digital images has become increasingly severe, leading to growing interest in digital watermarking as a key solution. Digital watermarking is a technique that embeds and detects a unique watermark to protect digital content's copyright and identify and trace tampering and forgery. In this paper, we propose a method that applies a three-level discrete wavelet transform (DWT) on an image to separate its frequency components into multiple levels, followed by singular value decomposition (SVD) across multiple regions to repeatedly embed the watermark into the singular values. The goal is to achieve robust and imperceptible watermarking capable of withstanding signal distortion attacks.

Keywords— discrete wavelet transform, singular value decomposition, digital watermarking, image protection

1 Introduction

The unauthorized reproduction and copyright infringement of digital images have intensified with the rapid growth of digital media. The rise of generative artificial intelligence (AI) has particularly increased instances of image manipulation and misuse, making copyright protection a pressing issue. Lawsuits are emerging from image providers and content creators against AI-based image generation platforms [1], claiming that the unlicensed use of their data for training constitutes infringement.

In 2022, the U.S. Copyright Office ruled that AI-generated images are not eligible for copyright protection [2], reaffirming that only works created by humans can be copyrighted. This decision ignited debate over the legal boundaries of AI-generated content and underscored the need for alternative technological protection methods. Digital watermarking has garnered attention as a vital tool for preserving content authenticity and copyright [3].

Digital watermarking is widely employed to protect image ownership and detect tampering [4]. It must meet two primary conditions: maintaining visual quality and ensuring robustness against attacks [5]. Frequency-domain methods, such as discrete wavelet transform (DWT), are commonly utilized to achieve these objectives. However, even DWT-based methods can be susceptible to certain attacks [6].

To enhance robustness, recent approaches have combined DWT with singular value decomposition (SVD) [7]. SVD modifies singular values that contain essential image features, allowing for watermark insertion with minimal distortion [8]. In the DWT-SVD framework, watermarks

*Proceedings of the 9th International Conference on Mobile Internet Security (MobiSec'25), Article No. 78, December 16-18, 2025, Sapporo , Japan.

†Corresponding author

are embedded into the singular values of DWT-decomposed components, improving resistance to noise and compression [9], [10].

This paper proposes a watermarking method that applies a three-level DWT followed by SVD on both low- and selected high-frequency components. The watermark is embedded repeatedly to ensure recovery even under partial distortion. Embedding in the low-frequency region preserves imperceptibility, while embedding in high-frequency bands enhances resistance to filtering and frequency-domain attacks. During extraction, repeated patterns facilitate error correction and accurate recovery through correlation analysis.

The main contributions of this study are as follows:

- This study proposes an invisible digital watermarking method that integrates three-level DWT and SVD for enhanced copyright protection. By embedding the watermark across multiple frequency components and applying redundancy, the method achieves robust resilience against partial data loss and maintains strong performance under various signal distortion attacks.

- A comprehensive evaluation framework is developed to systematically assess digital watermarking performance under a range of signal distortion conditions. Experimental results using this framework demonstrate that the proposed method significantly enhances both image fidelity and watermark extraction accuracy, outperforming conventional watermarking techniques.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 presents the proposed watermarking method, while Section 4 describes the experimental setup. Section 5 discusses the performance evaluation results. Finally, Section 6 concludes the paper.

2 Related Works

As interest in protecting digital multimedia copyrights grows, digital watermarking has emerged as an effective solution, prompting active research in the field. This section investigates and analyzes various frequency-domain watermarking techniques prioritizing imperceptibility and robustness.

2.1 Digital Image Watermarking Techniques Using DWT

Hosseini et al. [11] proposed a hybrid watermarking method that combines DWT, discrete cosine transform (DCT), and principal component analysis (PCA) to enhance robustness against noise and compression attacks through multi-resolution analysis and data compression. However, this method suffers from high computational complexity and a poor trade-off between imperceptibility and robustness, leading to noticeable degradation in image quality.

Lidyawati et al. [12] applied a three-level DWT and embedded the watermark into the LL3 low-frequency sub-band with reduced strength to maintain high visual quality. The method was tested against Gaussian noise, Salt & Pepper noise, and blurring. While it ensured good imperceptibility, the lack of consideration for attack intensity led to weak performance in some scenarios.

2.2 Digital Image Watermarking Techniques Combining DWT and SVD

Yasmeen et al. [13] utilized a four-level DWT on the host image and a three-level DWT on the watermark image, applying SVD to the LL (low-frequency) and HH (high-frequency) sub-bands. The watermark was embedded by merging the singular values. The watermarked image achieved an average peak signal-to-noise ratio (PSNR) of 40 dB, while the extracted watermark maintained PSNRs of 23 dB, 27.5 dB, and 30 dB under various noise attacks. Although the method demonstrated good imperceptibility and robustness, it lacked evaluation under varying attack intensities.

Kusumaningrum et al. [14] proposed a watermarking method that combined a two-level DWT with SVD. The watermark was embedded in the LL2 sub-band following the two-level DWT and SVD process, and a non-blind extraction method, which required the original image, was employed. The results were compared with DWT-only and SVD-only approaches under attacks such as Salt & Pepper noise, Gaussian filtering, and JPEG compression. The findings indicated superior extraction performance overall, emphasizing the advantages of embedding in low-frequency components and combining DWT with SVD. However, the experiments did not define attack intensities, and the method exhibited limited robustness under certain conditions. Despite this, the study effectively demonstrated the performance benefits of this hybrid approach.

3 Image Watermarking Method Based on DWT and SVD

This study proposes an imperceptible watermarking method that is robust against signal distortion attacks. The proposed approach applies a three-level DWT to decompose the image into frequency sub-bands, then performs SVD on the low-frequency and selected high-frequency components to repeatedly embed the watermark into the singular values. This method helps prevent watermark degradation under various attacks and enables successful reconstruction by integrating the watermarks extracted from multiple frequency regions.

High-frequency components represent edge details and are less perceptible to the human eye, making them suitable for watermark embedding. In contrast, low-frequency components contain the overall structure and major features of the image, so minimizing quality degradation during embedding is essential. Since JPEG compression primarily removes high-frequency content, watermarks embedded in the low-frequency domain demonstrate greater robustness. Additionally, low-frequency components are more stable under noise attacks such as Gaussian noise or resolution reduction. However, the embedded watermark may still be affected by high compression rates or formats such as JPEG2000, which also compress low-frequency bands.

By leveraging both low- and selected high-frequency components, this study achieves an imperceptible watermarking method with enhanced robustness against various noise-based attacks.

3.1 Watermark Embedding Process

While the general structure of image watermarking methods that combine DWT and SVD is similar, the specific steps may vary depending on the research objectives and methodology. Typically, the host image undergoes a DWT to generate sub-bands (LL, LH, HL, HH), followed by the application of SVD to a selected sub-band. The watermark is embedded by modifying the singular values obtained from SVD, although different studies may adopt various strategies

for this step. Various approaches can be employed to enhance robustness and imperceptibility or to improve computational efficiency, depending on the intended goals.

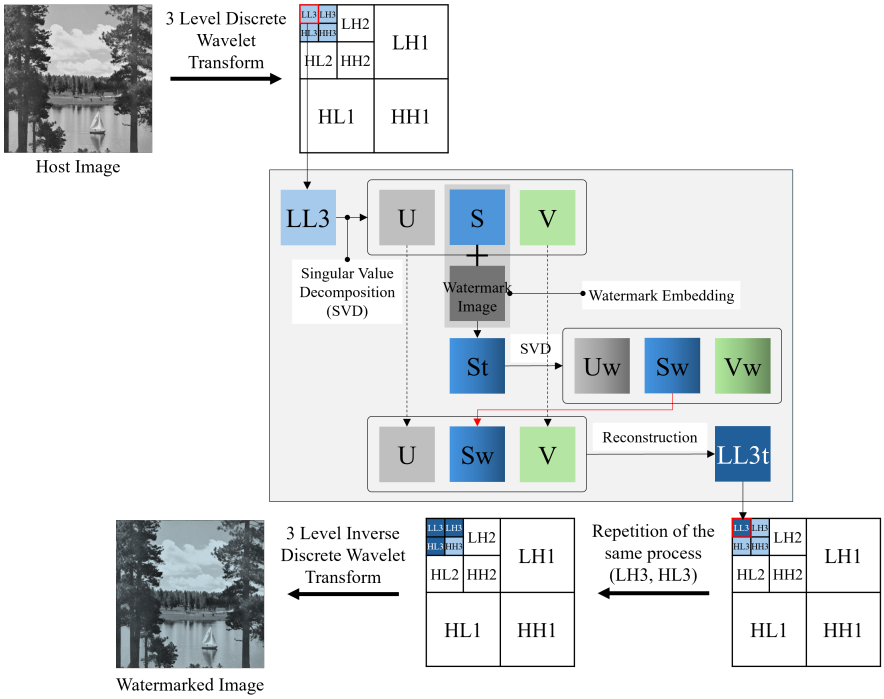

Figure 1: Watermark embedding process

Figure. 1 illustrates the proposed watermark embedding process. In this method, the watermark image must be of a fixed size that corresponds to the size of the host image. The relationship between the sizes of the host image and the watermark image is defined by Equation (1).

$$W = \frac{N}{2^L} \tag{1}$$

Let W be the size of one side of the watermark image, and N be the size of one side of the host image. Let L represent the DWT level. In this work, since the host image has a side length of 512 pixels and a three-level DWT is applied, a 64×64 pixel watermark image is required. When a three-level DWT is applied to the host image to transform it into the frequency domain, four sub-bands are generated: LL3, LH3, HL3, and HH3. Among these, SVD is applied to the low-frequency sub-band (LL3) and selected high-frequency sub-bands (LH3, HL3). The singular values obtained from LL3, LH3, and HL3 are each modified to embed the watermark. After embedding the watermark into these singular values, another round of SVD is performed on the modified singular value matrices (denoted as S_w). The resulting singular values (S_w) are then used to reconstruct the sub-bands LL3 ($LL3_t$), LH3 ($LH3_t$), and HL3 ($HL3_t$). Finally, the watermarked image is reconstructed through the inverse discrete wavelet transform (IDWT). The resulting image contains the same watermark embedded three times across different frequency components.

This study proposes a method that embeds a watermark into the singular values of a host image by applying DWT and SVD, followed by a second SVD step to adjust the modified singular values. Changes to the singular values during embedding may affect the image structure, potentially leading to quality degradation or failed watermark recovery. To address this, the second SVD step serves as a recalibration process that aligns the modified singular values with the original image structure, thereby enhancing both the imperceptibility and robustness of the watermark.

3.2 Watermark Extraction Process

Figure. 2 illustrates the proposed watermark embedding process. The extraction follows a non-blind watermarking approach. A three-level DWT is applied to the watermarked image to transform it into the frequency domain. SVD is then performed on the LL3, LH3, and HL3 sub-bands to extract the singular value matrices where the watermark was embedded. Using the extracted singular value matrices (S_w) along with the corresponding U_w and V_w matrices, the watermark images are reconstructed. Since the watermark was embedded once in each of the LL3, LH3, and HL3 components, three instances of the watermark can be extracted to reconstruct the final watermark image.

The reconstruction of the watermark proceeds as follows: From the watermark arrays extracted from the LH3 and HL3 sub-bands, median fusion is applied by computing the median of the corresponding values at each position. This process combines information from the LH3 and HL3 sub-bands, reducing noise and integrating the watermark data.

The intermediate watermark obtained from median fusion is then combined with the watermark extracted from the LL3 sub-band using a weighted combination. Since the LL3 sub-band contains the most important image information and is least affected by external distortions and noise, it is assigned to be a higher weight. Through this process, the watermark extracted from the LL3 sub-band plays a central role, while the information from LH3 and HL3 complements and enhances the reconstruction.

Experiments were conducted to evaluate the performance of the proposed image watermarking method, which combines three-level DWT and SVD. This section describes the experimental environment, procedures, and performance evaluation metrics.

4 Methodology

Experiments were conducted to evaluate the performance of the proposed watermarking method. This section describes the experimental environment, procedures, and performance evaluation metrics.

4.1 Experiment Environments

Figure. 3 shows the host and watermark images used in the experiment. A 512×512 gray-scale image served as the host image, while a 64×64 grayscale image was used as the watermark.

To evaluate the robustness of the proposed watermarking method, experiments were conducted using seven types of signal distortion attacks, including noise, compression, and filtering. The robustness was assessed across five levels of attack intensity, ranging from mild to severe. The parameters and settings defining the attack strengths are summarized in Table 1.

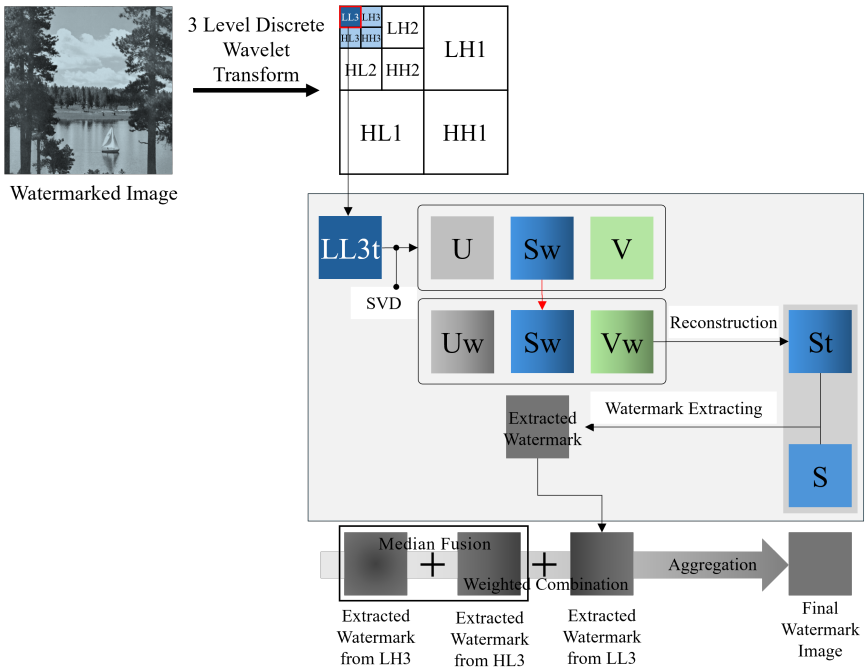

Figure 2: Watermark extraction process

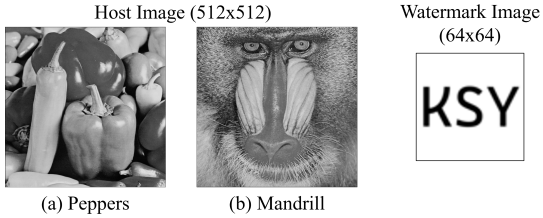

Figure 3: Host and watermark images

4.2 Experimental Procedure

Figure. 4 illustrates the flowchart of the experimental procedure. After uploading a gray-scale host image, it is resized to 512×512 pixels, and a three-level DWT is performed to extract the LL3, LH3, HL3, and HH3 sub-bands. The DWT uses the Daubechies 4 (db4) wavelet with periodization as the boundary handling mode. SVD is then applied to the low-frequency band (LL3) and selected high-frequency bands (LH3, HL3). The watermark is embedded into the singular value matrices using a scaling factor alpha (α). Another SVD is performed on the modified matrices to generate the watermarked singular value matrices, which are then used to reconstruct LL3t, LH3t, and HL3t. The final watermarked image is obtained by replacing the original LL3, LH3, and HL3 with these modified components and applying the inverse three-level DWT.

For robustness evaluation, various signal distortion attacks are applied. The attacked image is then subjected to a three-level DWT, and SVD is applied to LL3t, LH3t, and HL3t to

Table 1: Attack parameters and attack intensity

Type	Attack	Parameter	1	2	3	4	5
Noise attack	Gaussian noise	Variance	0.001	0.005	0.01	0.05	0.1
	Salt and pepper	Density	0.01	0.03	0.05	0.1	0.2
	Speckle noise	Probability	0.01	0.03	0.05	0.1	0.2
Compression attack	JPEG compression	Quality factor	90	70	50	30	10
	JPEG2000 compression	Quality factor	90	70	50	30	10
Filtering attack	Blurring attack	Kernel size	3	5	7	9	11
	Low-frequency filtering	Kernel size	3	5	7	9	11

extract the embedded watermark. The same alpha value (α) used during embedding is applied to reconstruct the singular matrices (St), and the three watermarks extracted from each sub-band are integrated to recover the final watermark image. During reconstruction, the median of the watermarks extracted from LH3 and HL3 is computed as intermediate watermark data. This intermediate data is then combined with the LL3 watermark using a weighted sum, where the intermediate watermark is assigned a weight of 0.3 and the LL3 watermark a weight of 0.7.

4.3 Performance Evaluation Metrics

The following evaluation metrics were used.

Normalized cross-correlation (NCC) measures the similarity between two images, indicating how closely they resemble each other. It is used to evaluate the similarity between the host image and the watermarked image and between the original and extracted watermark.

PSNR is a widely used metric for assessing the quality of a transformed image compared to the original. It quantifies how well the image quality is preserved after watermark embedding.

The structural similarity index measure (SSIM) evaluates the structural similarity between two images by incorporating human visual perception.

While PSNR relies on pixel-wise numeric values, SSIM considers image structure, brightness, and contrast, making it a more perceptually relevant measure.

5 Experiments

A comparative analysis was conducted against a conventional approach to validate the proposed method's performance. The conventional method embeds the watermark into the low-frequency sub-band (LL2) using a combination of two-level DWT and SVD [14]. For both the proposed and conventional methods, the embedding strength was set to $\alpha = 0.1$, and robustness was evaluated under five levels of attack intensity.

5.1 Image Quality Comparison

Figure. 5 compares image quality for the watermarked versions of the Peppers and Mandrill images using the proposed and conventional methods. In both cases, the conventional method showed slightly better preservation of image quality. This is because the conventional method

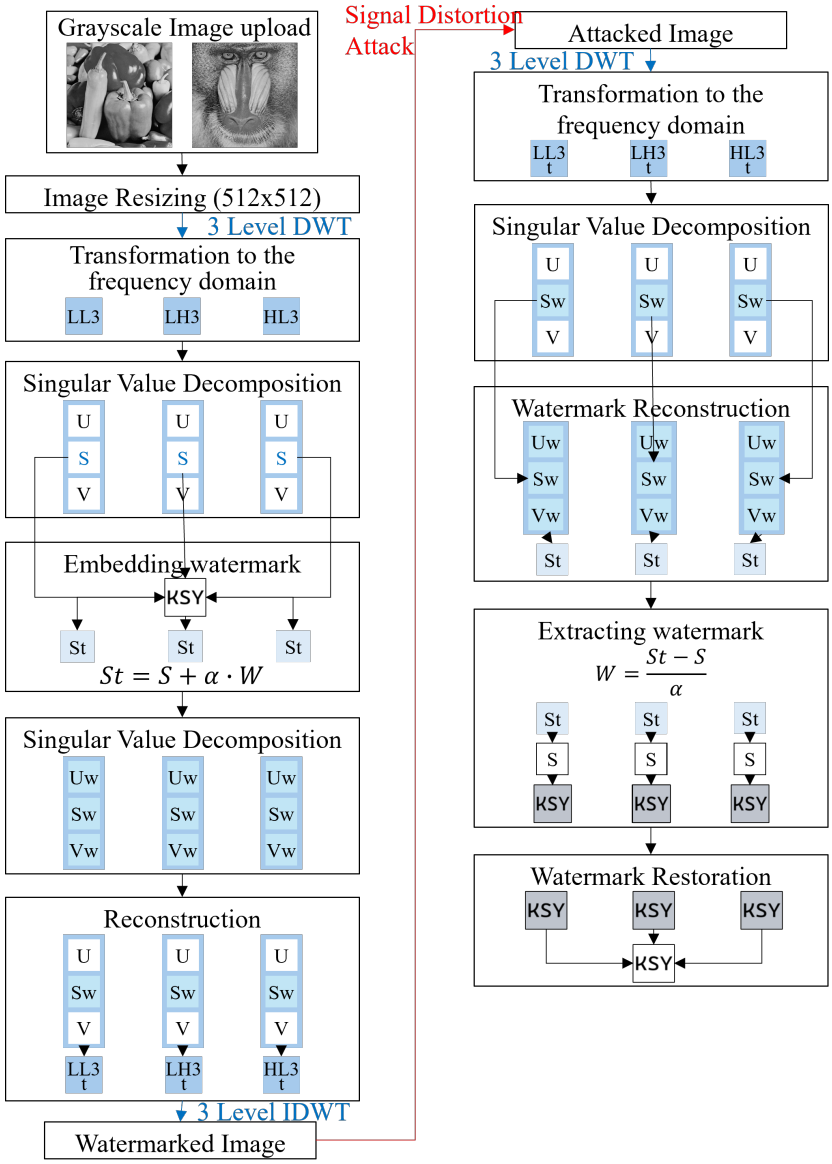

Figure 4: Flowchart of experimental process

embeds the watermark only in the LL2 sub-band, minimizing distortion. Regarding PSNR, the Peppers image showed an 11.5% degradation, while the Mandrill image experienced a 7.28% degradation when using the proposed method. However, the PSNR values remained above 40 dB, and SSIM, which reflects perceptual quality, showed only minor drops of 1.25% and 0.4%, respectively, maintaining high values above 0.98.

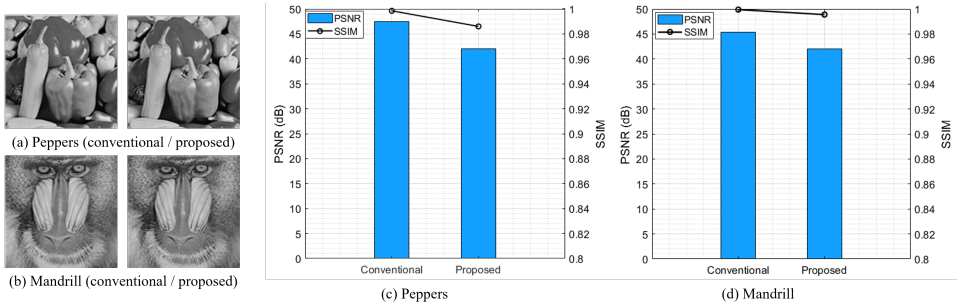

Figure 5: Comparison of image quality between conventional and proposed methods

5.2 Comparison of Watermark Extraction Performance

Signal distortion attacks were applied to the watermarked images to compare the watermark extraction performance between the proposed and conventional methods, followed by watermark extraction. Seven types of signal distortion attacks were considered in the experiments. Each attack was conducted across five intensity levels, ranging from mild to severe, and the watermark extraction performance was evaluated accordingly.

Figure. 6 compares the watermark extraction performance of conventional and proposed methods on the Peppers and Mandrill images under varying levels of signal distortion attacks, evaluated using NCC.

When the conventional method was applied, both images exhibited a significant drop in NCC as the intensity of Gaussian noise, speckle noise, and low-frequency filtering attacks increased. For the Peppers image, performance degraded by 75%, 89.99%, and 82.55%, respectively, as the attack level increased from 1 to 5. In contrast, the Mandrill image showed degradation rates of 65.95%, 90.44%, and 92.34%. Salt & Pepper and blurring attacks also resulted in noticeable decreases in performance.

In contrast, the proposed method demonstrated only minor degradation in performance as attack intensities increased. Under the most damaging low-frequency filtering attack, the NCC dropped by only 14.26% for Peppers and 16.10% for Man-drill, representing a significantly smaller loss than the conventional method. Although the difference in performance was limited under general compression attacks, the proposed method achieved much higher robustness against JPEG2000 compression at level 5, with improvements of 31.47% for Peppers and 94.66% for Mandrill.

6 Conclusion

This paper proposes a digital image watermarking technique that achieves both strong robustness against signal distortion attacks and high imperceptibility. The proposed method combines three-level DWT with SVD, applying SVD to the low-frequency band (LL3) and selected high-frequency bands (LH3, HL3). By repeatedly embedding the watermark into the singular values across these regions, the method maintains image quality while enhancing resistance to various attacks. Repeated embedding in high-frequency regions enables complementary restoration of dam-aged watermark components, effectively addressing the trade-off between imperceptibility and robustness observed in conventional methods.

Experimental results show that the proposed method maintains high PSNR and SSIM values

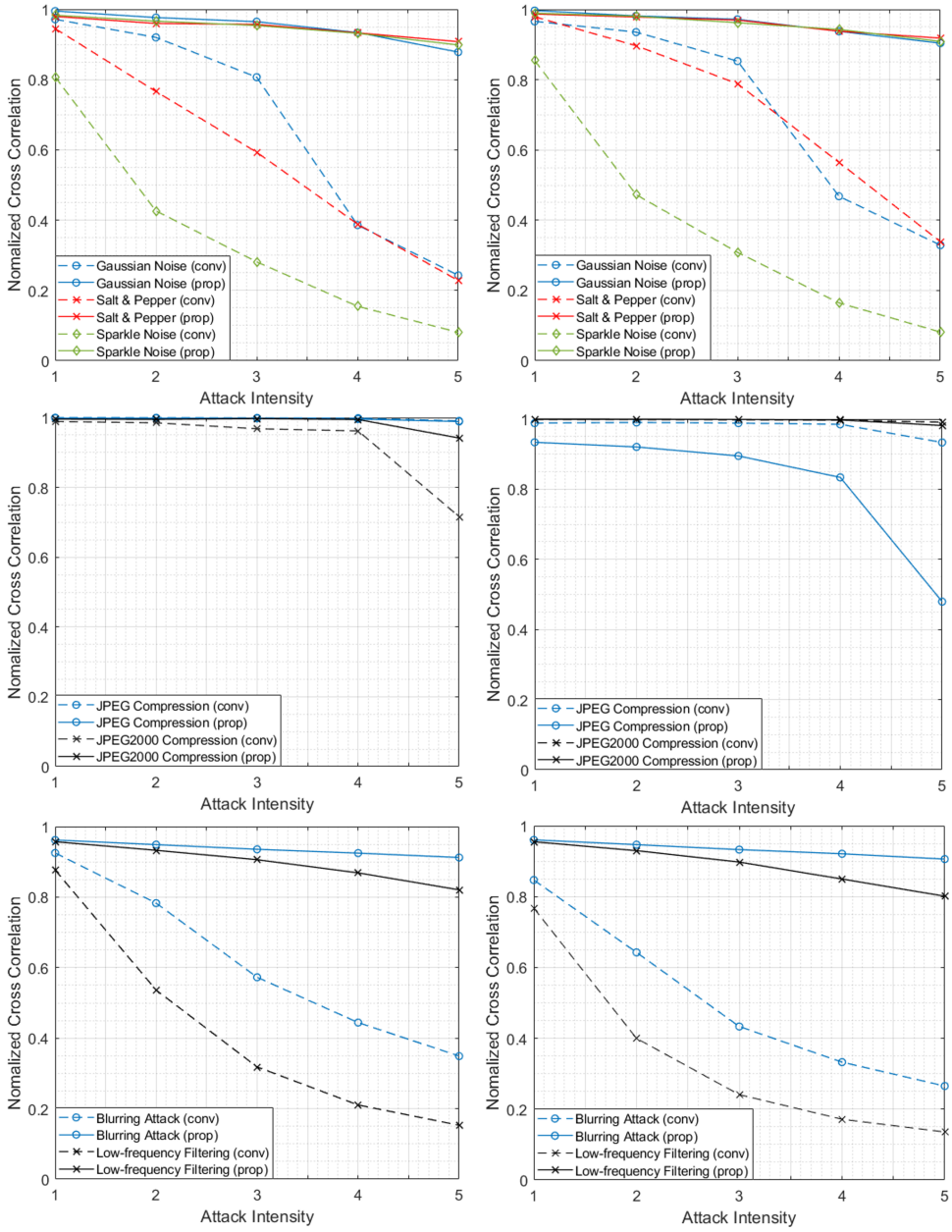

Figure 6: Comparison of extraction performance of conventional and proposed methods based on attack intensity – NCC (left: Peppers, right: Mandrill).

after embedding, thus preserving the image's visual quality. Repeated embedding significantly improves extraction performance compared to single em-bedding. Even when parts of the watermark are degraded, redundant embedding allows for reliable reconstruction, demonstrating strong performance under various noise and compression attacks.

This work presents a practical approach to digital content protection and is expected to serve as an effective solution in scenarios requiring both high reliability and imperceptibility.

Acknowledgments. This work is supported by the Ministry of Trade, Industry and Energy (MOTIE) under Training Industrial Security Specialist for High-Tech Industry (RS-2024-00415520), supervised by the Korea Institute for Advancement of Technology (KIAT), and the Ministry of Science and ICT (MSIT) under the ICAN (ICT Challenge and Advanced Network of HRD) program (No. IITP-2022-RS-2022-00156310), supervised by the Institute of Information & Communication Technology Planning & Evaluation (IITP).

Disclosure of Intersts. The authors have no competing interests to declare that are relevant to the content of this article.

References

[1] C. T. Zirpoli, "Generative artificial intelligence and copyright law," *CRS Legal Sidebar*, no. LSB10922, pp. 1–3, 2023.

[2] Z. Ü. Kahveci, "Attribution problem of generative ai: a view from us copyright law," *Journal of Intellectual Property Law and Practice*, vol. 18, no. 11, pp. 796–807, 2023.

[3] H. Yao, J. Lou, Z. Qin, and K. Ren, "Promptcare: Prompt copyright protection by watermark injection and verification," in *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 845–861, IEEE, 2024.

[4] D. Awasthi, A. Tiwari, P. Khare, and V. K. Srivastava, "A comprehensive review on optimization-based image watermarking techniques for copyright protection," *Expert Systems with Applications*, vol. 242, p. 122830, 2024.

[5] P. Kadian, S. M. Arora, and N. Arora, "Robust digital watermarking techniques for copyright protection of digital data: A survey," *Wireless Personal Communications*, vol. 118, pp. 3225–3249, 2021.

[6] B. Dharmika, V. Kiran, and A. Muralidhar, "Privacy protection of digital information using frequency domain watermarking technique," in *2021 4th International Conference on Recent Trends in Computer Science and Technology (ICRTCST)*, pp. 123–130, IEEE, 2022.

[7] Z. Zainol, M. A. Hani, and A. Hussain, "Hybrid svd-based image watermarking schemes: a review," *IEEE Access*, vol. 9, pp. 32931–32968, 2021.

[8] N. Zermi, A. Ghozzi, and I. Hamza, "Robust svd-based schemes for medical image watermarking," *Microprocessors and Microsystems*, vol. 84, p. 104134, 2021.

[9] O. Evsutin and K. Dzhanashia, "Watermarking schemes for digital images: Robustness overview," *Signal Processing: Image Communication*, vol. 100, p. 116523, 2022.

[10] N. Zermi, A. Ghozzi, and I. Hamza, "A dwt-svd based robust digital watermarking for medical image security," *Forensic Science International*, vol. 320, p. 110691, 2021.

[11] S. A. Hosseini and P. Farahmand, "An attack resistant hybrid blind image watermarking scheme based on combination of dwt, dct, and pca," *Multimedia Tools and Applications*, vol. 83, no. 7, pp. 18829–18852, 2024.

[12] L. Lidyawati, A. Kricha, and A. Sakly, "Digital watermarking image using three-level discrete wavelet transform under attacking noise," *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 1, pp. 231–238, 2022.

[13] F. Yasmeen and M. S. Uddin, "An efficient watermarking approach based on ll and hh edges of dwt–svd," *SN Computer Science*, vol. 2, no. 2, p. 82, 2021.

[14] D. P. Kusumaningrum, T. Asvini, and W. Prasetyo, "Dwt-svd combination method for copyrights protection," *Scientific Journal of Informatics*, vol. 7, no. 1, p. 311, 2020.