SpecGuard

Robust Invisible Watermarking for Digital Images

Presented by

Inzamamul Alam

Sungkyunkwan University

October 20, 2025

The Challenge: Digital Media Protection

Digital Media Authentication Challenges

Advanced Image Processing
AI tools enable easy image forgery and redistribution, threatening content ownership verification

Invisible Watermarking
Embeds invisible information to verify authenticity while maintaining image quality

Fundamental Trade-off
Basic balance between imperceptibility and robustness against transformations

Limitations of Existing Methods

Transform-based Methods
Lack robustness against image processing operations like resizing, cropping, compression, and noise

Deep Learning Methods
StegaStamp, Stable Signature, and HiDDeN show fragility when handling common image processing operations

Adversarial Attacks
Noise injection, blurring, contrast adjustment, and rotation significantly impact watermark performance

Image Regeneration
Methods like DiffusionDB, Rinse, and AdvEmb show poor robustness against regeneration attacks

SpecGuard: Method Overview

Core Concept

SpecGuard is a novel **robust and invisible** image watermarking method that strategically embeds watermark information in the **spectral domain** for enhanced protection against various image transformations.

Unlike traditional frequency domain methods that are easily destroyed by image operations, SpecGuard maintains imperceptibility while significantly improving robustness.

Key Modules

Encoder: Embeds watermark using wavelet and spectral projection

Decoder: Extracts watermark with robust Parseval-based approach

Cover Image

Wavelet Decomposition | Watermark

Watermarked Image

Key Features

Strategic Embedding

Distributes watermark across high-frequency components using wavelet-based decomposition

Enhanced Robustness

Maintains imperceptibility while improving resilience against transformations and attacks

Parseval's Theorem

Ensures energy conservation between spatial and frequency domains, preserving watermark integrity

Optimal Balance

Achieves near-perfect balance between imperceptibility and robustness

Technical Architecture: Encoder

Cover Image I
Original image

Wavelet Projection
Decomposition into subbands

Spectral Projection
FFT-based approximation

Watermarked Image
Iembedded

Embedding Process

Binary Message M
Watermark to embed

Message Processing
Reshaping and expansion

Target Region
Radial mask with radius r

Watermark Embedding
Controlled by strength factor s

Wavelet Projection (WP)

Captures frequency and spatial localization features at different scales using wavelet functions

Spectral Projection (SP)

Converts to frequency domain using FFT approximation, enabling embedding in high-frequency components

Targeted Embedding

Uses radial mask centered on high-frequency band S for embedding

Reconstruction

Inverse transforms (IWP and ISP) to convert back to spatial domain while preserving modifications

Wavelet Projection (WP)

Capturing Multi-scale Features

Wavelet projection captures frequency and spatial localization features at different scales through orthogonal wavelet sets.

2D Decomposition

For 2D input, WP defines basis elements in:

- Horizontal direction
- Vertical direction
- Diagonal direction

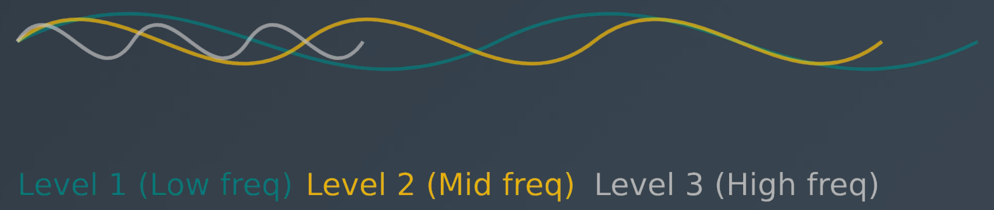
Decomposition Level

Level κ determined by image complexity:

$$\kappa = \lfloor \sqrt{\log(1+N)} \rfloor$$

Where N is total pixels in cover image

Wavelet Projection Visualization



Key Properties

Orthogonal basis functions

Multi-resolution analysis

Local frequency representation

Energy preservation

Spectral Projection Approximation

Process Overview

1 Symmetric Extension

Mirror $T(x,y)$ along boundaries to create symmetric extension $\tilde{T}(x,y)$

2 2D FFT Application

Apply 2D Fast Fourier Transform to $\tilde{T}(x,y)$ to obtain frequency domain representation

3 Real Part Extraction

Approximate spectral coefficients by taking the real part of the FFT operation in the original $N \times N$ region

Spectral Projection Benefits

- Separates input into low and high frequencies
- High-frequency subband S_{HH} provides details for embedding
- Preserves energy distribution while enabling selective modification
- Facilitates robust watermarking against transformations

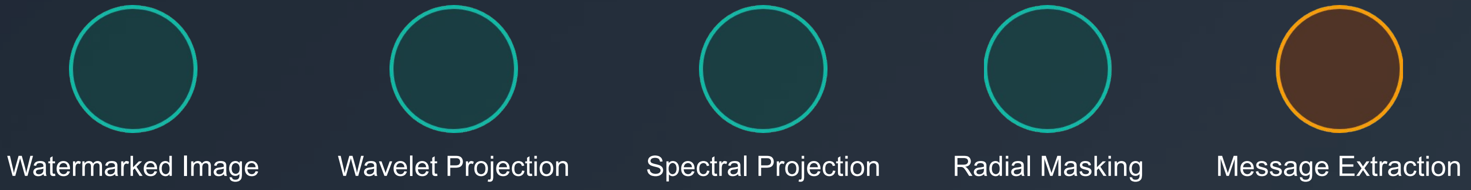
Spectral Projection Formula

$$S_{HH}(u,v) \approx \mathrm{Re}\{\mathrm{FFT}[\tilde{T}(x,y)]_{real}$$

Where Re{} denotes the real part of the complex FFT coefficients

Visual Process Flow

Technical Architecture: Decoder



Wavelet Projection

Separates watermarked image into low-frequency (S_{DLL}) and high-frequency (S_{DHH}) bands

Radial Masking

Creates mask based on Euclidean distance from center point to isolate high-frequency regions within radius r

Spectral Projection Approximation

Applies FFT-based spectral projection to high-frequency band, returning transformed data S_{DHH}

Message Extraction

Compares mask values with learnable threshold θ to decode message bits: $D[i] = 1$ if $Extracted[i] > \theta$, else 0

Feature Refinement

Sequential convolutional layers with LeakyReLU activation further refine the features: $S_{DHH} = \text{LeakyReLU}(\text{Conv}_{2D}(S_{DHH}^{sp}(n), K))$

Parseval Theorem

Ensures energy conservation between spatial and frequency domains, maintaining message integrity while adapting to image's spectral pattern

Embedding & Extraction Process

Embedding Process

1 Image Decomposition

Cover image I decomposed into frequency subbands: S_LL, S_LH, S_HL, and S_HH

2 Message Preparation

Message M reshaped and expanded to align with S_HH

3 Radial Mask Creation

Based on Euclidean distance from center point, embedding within radius r

4 Embedding

Controlled by strength factor s, embedding message into S_HH

Multiple convolutional layers with LeakyReLU activation

Extraction Process

1 Wavelet Projection

Applied to watermark image I_embedded, separating into low and high frequency bands

2 Spectral Projection

FFT applied to high-frequency band S_D_HH^high

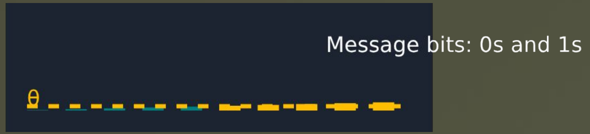
3 Feature Refinement

Convolutional layers further refine S_DH to capture local features

4 Adaptive Threshold Optimization

Radial mask isolates high-frequency regions in S_DH

Mask values compared with learnable threshold θ to decode message bits



Loss Function Optimization

Encoder Loss (L)

Minimizes the difference between the original image and the watermarked image to maintain cover image fidelity

$$\min_\theta E_{(I,M)\sim D}\ L_{enc}(I, I_{embedded}) = \|E_\theta(I, M) - I\|^2$$

Visual Fidelity: Ensures the watermarked image is perceptually indistinguishable from the original

Decoder Loss (L)

Minimizes the difference between the original message and the extracted message to ensure reliable message retrieval

$$\min_\theta E_{(I,M)\sim D}\ L_{dec}(M, D_M) = \|D_\theta(I_{embedded}) - M\|^2$$

Message Integrity: Ensures accurate extraction of the embedded message

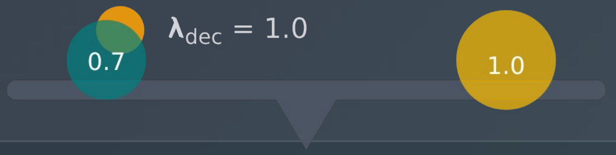
Combined Loss Function

$$\min_\theta L = \lambda_{enc} L_{enc} + \lambda_{dec} L_{dec}$$

Weighted coefficients balance visual fidelity and message recoverability

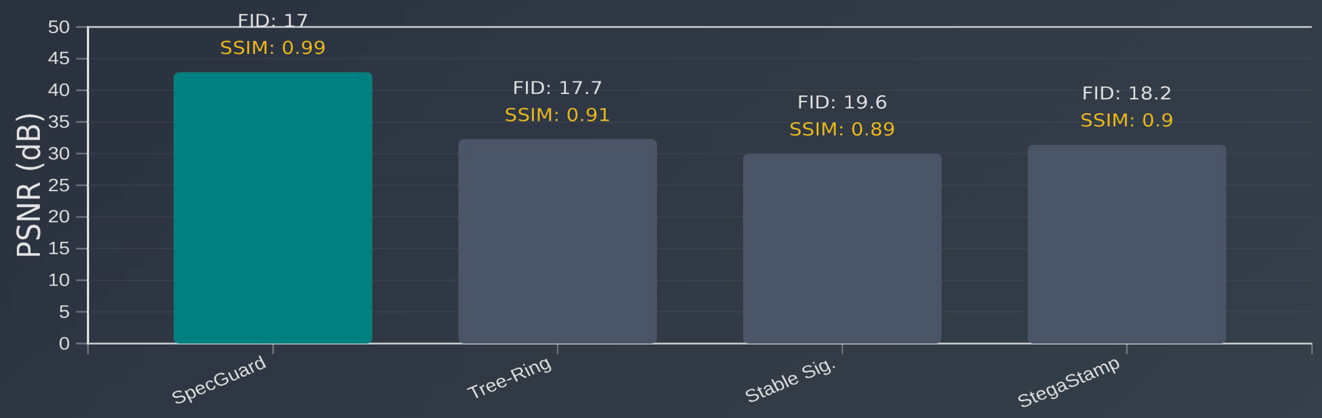
$\lambda_{enc} = 0.7$

$\lambda_{dec} = 1.0$



Experimental Results: Imperceptibility

Perceptual Quality Metrics



Visual Comparison

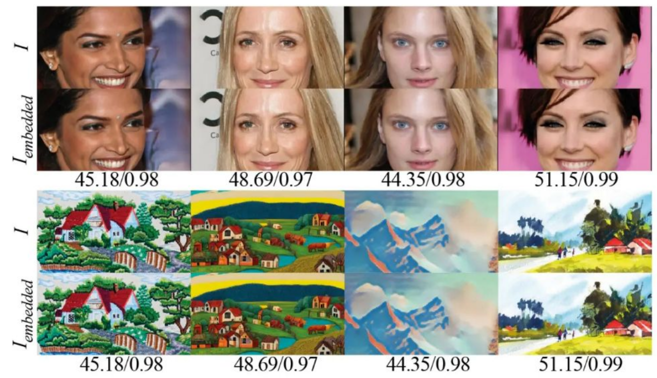


Figure 3. Some best results for cover vs watermarked images with PSNR/SSIM (↑) scores showing minimal visual degradation when watermarked using proposed SpecGuard.

Original vs. watermarked images with minimal visual degradation (PSNR/SSIM scores shown)

Key Findings

- Superior Performance: SpecGuard achieves higher PSNR (42.59-42.89) and SSIM (0.98-0.99) values compared to state-of-the-art methods
- Minimal Visual Degradation: FID values remain low (17.0-17.6), indicating minimal visual impact
- Cross-Dataset Consistency: Excellent performance across different datasets (DiffusionDB, MS-COCO, DALL-E3)
- Bit Recovery Accuracy: SpecGuard maintains high BRA (0.98-0.99) while keeping imperceptibility

Experimental Setup

- Training: MS-COCO dataset (25K images)
- Evaluation: DiffusionDB, MS-COCO, DALL-E3
- Message Length: 30-bit fixed length
- Metrics: PSNR, SSIM, FID, MSE, BRA

Experimental Results: Robustness

SpecGuard demonstrates exceptional robustness against diverse attacks compared to SOTA methods.

Performance Comparison

Attack Type		Tree-Ring [53]				Stable Signature [37]				StegaStamp [47]				SpecGuard (Ours)			
		Q@0.95P	Q@0.7P	Avg P	Avg Q	Q@0.95P	Q@0.7P	Avg P	Avg Q	Q@0.95P	Q@0.7P	Avg P	Avg Q	Q@0.95P	Q@0.7P	Avg P	Avg Q
Distortions	Rotation	0.464	0.521	0.375	0.648	0.624	0.702	0.594	0.650	0.423	0.498	0.357	0.616	0.863	0.863	0.687	0.653
	Crop	0.592	0.592	0.332	0.463	inf	inf	0.995	0.461	0.602	0.602	0.540	0.451	0.812	0.812	0.998	0.742
	Bright	inf	inf	inf	0.304	inf	inf	0.998	0.305	inf	inf	0.998	0.317	inf	inf	0.998	0.466
	Contrast	inf	inf	0.998	0.243	inf	inf	0.998	0.243	inf	inf	0.998	0.231	inf	inf	0.998	0.556
	Blur	0.861	1.112	0.563	1.221	– inf	– inf	0.000	1.204	0.848	0.962	0.414	1.000	0.921	inf	1.000	1.452
	Noise	0.548	inf	0.980	0.395	0.402	0.520	0.870	0.390	inf	inf	1.000	0.360	inf	inf	0.999	0.568
	JPEG	0.499	0.499	0.929	0.284	0.485	0.485	0.793	0.284	inf	inf	0.998	0.263	inf	inf	1.000	0.495
	Geo	0.525	0.593	0.277	0.768	0.850	inf	0.937	0.767	0.663	0.693	0.396	0.733	0.869	0.869	0.865	0.623
	Deg	0.620	inf	0.892	0.694	0.206	0.369	0.300	0.679	0.826	0.975	0.852	0.664	0.895	1.141	0.915	0.749
	Combine	0.539	0.751	0.403	0.908	0.538	0.691	0.334	0.900	0.945	1.101	0.795	0.870	0.979	1.256	0.911	0.952
Regeneration	Regen-Diff	– inf	0.307	0.612	0.323	– inf	– inf	0.001	0.300	0.331	inf	0.943	0.327	inf	inf	0.982	0.477
	Regen-DiffP	inf	0.307	0.601	0.327	– inf	– inf	0.001	0.303	0.333	inf	0.940	0.329	inf	inf	0.982	0.562
	Regen-VAE	0.578	0.578	0.832	0.348	0.545	0.545	0.516	0.339	inf	inf	1.000	0.343	inf	inf	0.995	0.521
	Regen-KLVAE	inf	inf	0.990	0.233 6	– inf	0.176	0.217	0.206	inf	inf	1.000	0.240	inf	inf	0.990	0.492
	Rinse-2xDiff	– inf	0.333	0.510	0.357	– inf	– inf	0.001	0.332	0.391	inf	0.941	0.366	inf	inf	0.993	0.561
	Rinse-4xDiff	– inf	0.355	0.443	0.466	– inf	– inf	0.000	0.438	0.388	inf	0.909	0.477	inf	inf	0.992	0.533
Adversarial	AdvEmbG-KLVAE8	– inf	0.164	0.448	0.253	inf	inf	0.998	0.249	inf	inf	1.000	0.232	inf	inf	1.000	0.456
	AdvEmbB-RN18	0.241	inf	0.953	0.218	inf	inf	0.999	0.212	inf	inf	1.000	0.196	inf	inf	1.000	0.467
	AdvEmbB-CLIP	0.541	inf	0.932	0.549	inf	inf	0.999	0.541	inf	inf	1.000	0.488	inf	inf	1.000	0.436
	AdvEmbB-KLVAE16	0.195	inf	0.888	0.238	inf	inf	0.997	0.233	inf	inf	1.000	0.206	inf	inf	1.000	0.482
	AdvEmbB-SdxlVAE	0.222	inf	0.934	0.221	inf	inf	0.998	0.219	inf	inf	1.000	0.204	inf	inf	1.000	0.492
	AdvCls-UnWM&WM	– inf	0.102	0.499	0.145	inf	inf	0.999	0.101	inf	inf	1.000	0.101	inf	inf	1.000	0.497
	AdvCls-Real&WM	inf	inf	1.000	0.047	inf	inf	0.998	0.092	inf	inf	1.000	0.106	inf	inf	1.000	0.427
	AdvCls-WM1&WM2	– inf	0.101	0.492	0.139	inf	inf	0.999	0.084	inf	inf	1.000	0.129	inf	inf	1.000	0.441

Table 3. Robustness comparison various across attacks using Q@0.95P(↑), Q@0.7P(↑), Avg P(↑) and Avg Q(↑). Here, 'inf' denotes that no attack was sufficient to degrade performance below the threshold, indicating strong robustness, whereas '-inf' signifies that even the weakest attack caused detection to fall below the threshold, reflecting weak robustness.

Table 4. Ablation studies on the proposed SpecGuard for across various configurations, setting $M = 128$, $r = 100$, and $s = 20$.

Platform	PSNR/SSIM↑	BRA↑	PS Filters	PSNR/SSIM↑	BRA↑
Facebook	48.56/0.97	0.97	Depth Blur	25.25/0.89	0.85
LinkedIn	47.55/0.97	0.96	StyleT.	25.12/0.84	0.85
Instagram	48.56/0.98	0.98	**Super Zoom**	**36.15/0.88**	**0.95**
WhatsApp	42.10/0.96	0.97	JPEG Artifacts	31.01/0.85	0.94
X (Twitter)	**49.25/1.00**	**0.99**	Colorize	23.15/0.82	0.92

Geometric Distortions

SpecGuard achieves 0.998 for Avg P under crop attacks

Regeneration Attacks

Shows strong robustness against Regen and Rinse attacks

Adversarial Attacks

Outperforms other methods under AdvEmb and AdvCls attacks

Social Media Platforms

Maintains high performance across Facebook, Instagram, WhatsApp

Key Advantages of SpecGuard

SpecGuard: Optimizing the Fundamental Trade-off

Superior Imperceptibility-Robustness Balance

- High PSNR (**40.36-48.17**) and SSIM (**0.989-0.994**)
- Low FID (**16.45-17.45**) and MSE values
- Near-zero visual degradation while maintaining robustness

High Embedding Capacity

- Superior BRA (**0.98-0.99**) at 256 bit length
- No BRA degradation with higher message lengths
- Outperforms StegaStamp and HiDDeN at all tested lengths

Enhanced Attack Resilience

- High robustness against geometric distortions (Crop: **0.812**, Bright: **0.998**)
- Excellent Avg P (**0.911**) and Avg Q (**0.952**)
- Strong resistance against regeneration and adversarial attacks

Social Media & PNF Resilience

Maintains high performance (**PSNR/SSIM > 48.56/0.97**, **BRA > 0.97**) across Facebook, Instagram, WhatsApp and other platforms and filters

Thank You & Questions

We appreciate your attention and valuable feedback

Protecting digital media authenticity through robust spectral watermarking

Questions & Discussion

We welcome your questions about SpecGuard implementation, applications, and future research directions

Acknowledgments

Research Institution: Dash Lab & Vis2know Research Team: Inzamamul Alam, Md Tanvir Islam, Simon S. Woo, and Khan Muhammad

contact:https://github.com/inzamamulDU/SpecGuard_ICCV_2025/issues