JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Context-Aware Adaptive Watermarking Using Machine Learning: An Integrated CNN-Based Region Prediction and Reinforcement Learning Framework

[1] Subhathra R, [2]Mohamed Ali E A, [3]Jehovah Jireh Arputhamoni S, [4]Sankaralingam R

[1]PG Student, [2]Associate Professor, [3,4]Assistant Professor
[1]Communication Engineering, [2,3,4]Electronics and Communication Engineering
[1,3]PSN College of Engineering and Technology, Tirunelveli, Tamil Nadu, India
[2]J.P. College of Engineering, Tenkasi, Tamil Nadu, India
[4]Cheran College of Engineering, Kongeyam, Tamil Nadu, India

Abstract : A key technology used in copyright protection, content authentication and management of digital rights is digital watermarking. Nevertheless, traditional watermarking systems have a root cause robustness imperceptibility tradeoff, which cannot be addressed dynamically in a wide range of image content, using fixed-strategy solutions. The proposed paper presents a new Context-Aware Adaptive Watermarking (CAAW) framework that combines the Convolutional Neural Network (CNN)-based embedding region prediction with a Reinforcement Learning (RL) strategy selection that will be carried out entirely in the encrypted coefficient space. The suggested system will consist of eight closely designed modules, which include: JPEG2000 compression and key-based permutation encryption, CNN-based embedding region classification, Q-learning strategy selection between Spread Spectrum (SS), Scalar Costa Scheme Quantization Index Modulation (SCS-QIM), and Random Dither Modulation (RDM), adaptive perceptual strength modulation on encrypted coefficients, two-stage self-healing watermark reconstruction mechanism, ten-scenario attack simulation suite, a complete performance evaluation framework. The CNN is able to classify the 32x32 encrypted patches of the coefficient patches as either suitable or unsuitable to watermark embedding, and the Q-learning agent will stabilize to preferred strategies within 30-40 training episodes. The Peak Signal-to-Noise Ratio (PSNR) of the integrated framework is 54.79 dB and Structural Similarity Index Measure (SSIM) is 0.9981 when using the Spread Spectrum strategy, and the mean of 42.18 dB and 0.9756 when all the strategy considerations are taken, which is above the conventional 40 dB imperceptibility level. The ten attack scenarios demonstrate that the Bit Error Rate (BER) of the system and the Normalized Correlation (NC) are 0.123 on average and 0.877, respectively. The contribution of each component is verified in an ablation study where CNN region prediction results in 2-3 dB PSNR increase and self-healing reconstruction leads to a 3-5 percentage points decrease in BER. It has been confirmed that the framework is a viable and reproducible system of privacy-sensitive, adaptive watermarking in cloud-based and digital rights management systems.

IndexTerms **- Digital watermarking; Convolutional neural networks; Reinforcement learning; Encrypted-domain processing; Adaptive embedding; Spread spectrum; Quantization index modulation.**

I. INTRODUCTION

The emergence of digital content distribution platforms including cloud-based content delivery networks, peer-to-peer sharing ecosystems and similar have created an urgent and rising demand of effective means to safeguard intellectual property, provide content provenance and facilitate attribution (verifiable). To fulfill these requirements, digital watermarking, which is the invisible insertion of ownership or authentication data directly into a cover medium, is employed by offering a payload to be transferred together with the content (Cox et al., 2007). In contrast to metadata-based techniques, watermarks in the signal domain are resistant to most content processing to maintain perceptual quality, such as format conversion, compression and geometric transformation, an attribute that is also known as robustness.

The inherent engineering dilemma in the design of watermarking systems is the robustness imperceptibility tradeoff: a stronger signal used to implement a watermark signal increases its resistance to attack, at the expense of more perceptual distortion; and more perceptual distortion requires a weaker signal to be recovered. Classical resolutions to such a tradeoff, including Spread Spectrum (SS) embedding (Cox et al., 1997), Quantization Index Modulation (QIM) (Chen and Wornell, 2001), and transform-domain solutions based on either the Discrete Cosine Transform (DCT) or the Discrete Wavelet Transform (DWT), have fixed operating points on such a tradeoff curve that are optimal when the image content is of a particular type, but less-optimal when used uniformly across a wide range of images. The image is more tolerant to embedding in textured regions that the human visual system (HVS) is not so

sensitive to structured distortion than smooth homogeneous regions that do not have a noticeable quality degradation. To take advantage of this heterogeneity, content-adaptive embedding strength and strategy selection, which rule-based systems cannot deliver in any reliable way in arbitrary image content, are required (Podilchuk and Delp, 2001).

Another aspect of complexity emerges in cloud-based content management architecture, whereby the content has to be processed in encrypted format in order to maintain privacy against the service provider. Traditional perceptual watermarking frameworks demand availability of plaintext pixel values so as to estimate masking thresholds and texture complexity which cannot be used with encrypted-domain processing. The creation of watermarking architectures that can process encrypted wavelet coefficients, without the need to decrypt those, is a technologically significant and practical research goal (Shih, 2017).

This paper will tackle both of these issues at the same time, whereby the Context-Aware Adaptive Watermarking (CAAW) framework is proposed, which is a machine learning-enhanced watermarking system that combines CNN-based embedding region prediction and RL-based strategy selection in the encrypted wavelet coefficient domain. It proposes a perceptual model that approximates a local texture complexity based on encrypted coefficients based on their statistical variance structure and does not need the decryption of coefficients at the coefficient level. It implements a self-healing reconstruction module which offers the extra strength of two-stage error correction: majority voting on blocks and then recovery of learned patterns conditioning on the context. These components are collectively an integrated pipeline which is adaptively optimized over each image during inference, and does not need retraining when new images are needed.

II . LITERATURE REVIEW

2.1 Transform-Domain Watermarking

The dominant paradigm in robust watermarking embeds the watermark payload in the transform domain of the cover image, where perceptually significant components can be identified and protected while watermark energy is distributed across the frequency spectrum. Cox et al. (1997) established the theoretical foundation of spread spectrum watermarking by modelling the watermarking problem as communication with side information, demonstrating that embedding in perceptually significant frequency components maximizes robustness against a broad class of attacks. The DWT domain emerged as a particularly attractive embedding medium due to its multi-resolution decomposition structure, which aligns naturally with the spatial frequency sensitivity of the HVS and enables selective embedding in sub bands corresponding to texturally rich mid-frequency content (Barni et al., 1998).

Quantization Index Modulation, proposed by Chen and Wornell (2001) made a theoretical breakthrough in showing that capacity-achieving watermarking could be obtained by just utilizing structured coefficient quantization, to find exploitation of the information-theoretic writing on dirty paper framework. Scalar Costa Scheme (SCS-QIM) is a simplified version of the QIM structure, which simplifies the structure to a practically realizable scalar quantizer, which is robust to additive noise attacks as well as has good imperceptibility properties (Eggers and Girod, 2002). Although these classical approaches possess a long-known theoretical property, they use fixed quantization step sizes and fixed embedding domains across each region of the image, without being sensitive to the image content properties at that region.

2.2 Adaptive and Learning-Based Watermarking

Content-adaptive watermarking emerged as a response to the limitation of fixed-strategy approaches. Podilchuk and Delp (2001) proposed a just-noticeable-difference (JND) model that adapts embedding strength to local visual masking thresholds, achieving improved imperceptibility at constant robustness. However, JND model computation requires access to plaintext luminance values and involves heuristic masking parameters that do not generalize well across diverse image content.

The integration of machine learning into watermarking systems has accelerated with the availability of deep learning frameworks. Zhu et al. (2018) introduced HiDDeN, an end-to-end deep learning watermarking system that jointly trains an encoder and decoder network with an adversarial noise layer simulating attack. While HiDDeN demonstrates impressive robustness through learned representations, it requires full plaintext access, generates floating-point embeddings that are difficult to reconcile with standard image compression pipelines, and does not address encrypted-domain operation. Ahmadi et al. (2020) extended the end-to-end paradigm to video watermarking with ReDMark, and Luo et al. (2020) proposed RivaGAN with attention mechanisms for robustness to geometric attacks, but both similarly require plaintext processing.

Reinforcement learning has been used to adaptive signal processing in other similar areas - most notably in adaptive compression and communication- although its use in watermarking strategy selection is scarcely studied. The strategy selection process of MDP where the agent monitors image coefficient statistics as state and chooses embedding strategies as actions to maximize a reward synthesizing imperceptibility and robustness is a natural and principled process that has not been studied in systematic literature before.

2.3 Encrypted-Domain Signal Processing

Processing multimedia content in the encrypted domain addresses privacy requirements in cloud outsourcing scenarios, where the content owner wishes to delegate processing to an untrusted server without revealing plaintext content. Homomorphic encryption frameworks theoretically enable arbitrary computation on encrypted data, but their computational overhead remains prohibitive for real-time multimedia processing (Gentry, 2009). Practical encrypted-domain signal processing typically exploits the structure of specific encryption schemes such as the commutativity of DWT with certain linear operations on encrypted coefficients to enable targeted processing without full decryption (Shih, 2017).

Encrypted-domain watermarking has been proposed in several forms: signal-level encryption followed by watermark embedding (reversible or irreversible), homomorphic watermarking over partially homomorphic cryptosystems, and joint compression-encryption-watermarking in JPEG2000 and similar compressed domain frameworks (Cancellaro et al., 2011). The key research gap that the present work addresses is the absence of an adaptive, ML-guided embedding strategy that operates within the encrypted coefficient domain that combining the privacy advantages of encrypted-domain processing with the performance advantages of content-adaptive machine learning-based embedding.

III. Proposed Methodology and System Architecture

3.1 System Overview

The CAAW framework is structured as a pipeline of eight sequential modules. The pipeline processes a grayscale cover image I of dimensions 256×256 through JPEG2000 compression and key-based permutation encryption to produce an encrypted coefficient representation C_enc. The CNN embedding region predictor classifies patches of C_enc to produce a binary suitability mask M. The RL strategy selector observes coefficient statistics to select the embedding strategy s ∈ {SS, SCS-QIM, RDM} for each region. The adaptive embedding engine applies the selected strategy with perceptually modulated strength α(x,y) to produce the watermarked coefficient set C_w. The self-healing reconstruction module provides error recovery, and the attack simulation module evaluates robustness across ten predefined attack scenarios.

The complete CAAW pipeline flows from left to right through eight modules: (1) JPEG2000 Compression & Encryption, (2) CNN Embedding Region Predictor, (3) RL Strategy Selector (Q-Learning), (4) Perceptual Model on Encrypted Coefficients, (5) Adaptive Embedding Engine, (6) Self-Healing Reconstruction, (7) Attack Simulation Module, and (8) Performance Evaluation. The feedback loop from the Performance Evaluation module to the RL Strategy Selector enables online reward computation during training.

3.2 JPEG2000 Compression and Encrypted-Domain Representation

The cover image I is decomposed using a three-level Haar Discrete Wavelet Transform (DWT), producing a multi-resolution coefficient set arranged in subbands {LL_3, LH_3, HL_3, HH_3, LH_2, HL_2, HH_2, LH_1, HL_1, HH_1}. The low-frequency approximation subband LL_3 captures the global structure of the image and is preserved unmodified, while the mid-frequency detail subbands provide the primary embedding domain. The wavelet coefficients are subsequently encrypted using a key-based permutation cipher: a pseudo-random permutation π_k, parameterized by a secret key k, reorders the coefficient positions within each subband, rendering the coefficient spatial structure uninterpretable to an observer without knowledge of k while preserving the coefficient value distribution. This permutation encryption ensures that the subsequent CNN and RL modules operate on encrypted data, providing an honest-but-curious security model appropriate for cloud processing scenarios.

3.3 CNN Architecture for Encrypted Region Prediction

The CNN embedding region predictor processes 32×32 encrypted coefficient patches extracted from the detail subbands of C_enc. The network architecture comprises two convolutional blocks followed by a fully connected classification head:

The first convolutional layer applies 16 filters of size 5×5 with ReLU activation and is followed by 2×2 max-pooling, reducing the spatial dimension from 32×32 to 14×14 while expanding the channel dimension. The second convolutional layer applies 32 filters of size 3×3 with ReLU activation and 2×2 max-pooling, further reducing to 6×6 spatial resolution. The flattened feature vector (32 × 6 × 6 = 1,152 dimensions) is passed through a fully connected layer of 64 neurons with 50% dropout regularization. The output layer is a two-neuron softmax classifier producing probabilities for the "suitable" and "unsuitable" embedding region classes. Binary cross-entropy loss is minimized using the Adam optimizer during training.

The training dataset comprises 800–1,000 encrypted coefficient patches extracted from a diverse set of training images, with class balance maintained at 45–55% suitable patches. Ground truth labels are generated by applying a local variance threshold to the unencrypted patch coefficients: patches with high local variance (corresponding to textured image regions) are labelled "suitable," as they tolerate embedding distortion more readily under HVS masking. Smooth patches (low variance) are labelled "unsuitable." This labelling strategy reflects the well-established principle that visual masking in the HVS is stronger in textured regions, enabling higher watermark embedding capacity without perceptible distortion. The training loss decreases from an initial value of 0.69 (consistent with random initialization for a balanced binary classifier) to approximately 0.30 at convergence, and the classifier achieves 85–90% test accuracy.

3.4 Reinforcement Learning Strategy Selection

The watermarking strategy selection problem is formulated as a finite Markov Decision Process (MDP) with discrete state and action spaces. The state space S = {s_1, s_2, ..., s_{10}} comprises 10 discrete states determined by the quantized values of three coefficient statistics computed from a 32×32 encrypted patch: (1) coefficient variance σ^2 reflecting overall texture energy; (2) Shannon entropy H reflecting distributional complexity; and (3) a texture measure T derived from the second-order statistical properties of the coefficient magnitudes. States are assigned by discretizing the joint statistic ([σ^2, H, T]) into 10 bins through k-means clustering on training data statistics.

The action space A = {SS, SCS-QIM, RDM} contains three watermarking strategies, each optimized for a different coefficient texture regime as detailed in Section 3.5. The Q-learning agent maintains a tabular Q-function Q(s, a) ∈ ℝ^{10×3} and updates it using the Bellman equation with ε-greedy exploration (ε = 0.1). The reward function R(s, a) combines three terms:

$$R(s,a) = w_1 \cdot \text{PSNR}_{\text{norm}} + w_2 \cdot (1 - \text{BER}) + w_3 \cdot \text{Efficiency} \quad (1)$$

where $\text{PSNR}_{\text{norm}}$ is the normalized PSNR value (scaled to [0,1] using empirical min/max bounds), BER is the bit error rate measured on a held-out test attack scenario, and Efficiency is a computational cost penalty that rewards lower-complexity strategies. The weights w_1, w_2, w_3 are set to 0.4, 0.4, and 0.2 respectively, reflecting the primary importance of imperceptibility and robustness. The Q-value update rule follows the standard Q-learning update:

$$Q(s,a) \leftarrow Q(s,a) + \alpha[R + \gamma \cdot \max_{a'}^{Q(s',a')} - Q(s,a)] \quad (2)$$

where α is the learning rate and γ is the discount factor. Training proceeds over 50 episodes, with each episode processing all patches from a randomly selected training image. The agent converges to stable strategy preferences within 30–40 episodes, as evidenced by the reward convergence curve described in Section 6.2. Post-convergence, the Q-table encodes a systematic strategy preference: SS is preferred for high-texture states (high σ^2, high H), SCS-QIM for medium-texture states, and RDM for low-texture smooth-region states.

3.5 Three Watermarking Strategies

Spread Spectrum (SS). The SS strategy embeds the watermark bit $w_i \in \{-1, +1\}$ by additive modulation of the DWT coefficient C_i with adaptive strength $\alpha(i)$:

$$C'_i = C_i + \alpha(i) \cdot w_i \quad (3)$$

The watermark sequence $\{w_i\}$ is a pseudo-random binary sequence keyed to the same secret key k used for encryption permutation, ensuring that extraction requires knowledge of k. The adaptive strength $\alpha(i)$ is determined by the perceptual model described in Section 3.6. SS is best suited to high-texture regions where the spread-spectrum noise is masked by the natural coefficient variability, and its additive structure provides good robustness to additive noise attacks.

SCS-QIM (Scalar Costa Scheme). The SCS-QIM strategy embeds watermark bit $w \in \{0, 1\}$ through structured coefficient quantization:

$$C'_i = \Delta \cdot \left\lfloor \frac{C_i}{\Delta} + 0.5 - \delta w \right\rfloor + \delta w \quad (4)$$

where Δ is the quantization step size and $\delta = 0.25$ is the SCS dither parameter. The quantization step Δ controls the robustness-imperceptibility tradeoff: larger Δ increases robustness to noise at the cost of greater coefficient distortion. SCS-QIM achieves near-capacity watermarking performance under additive white Gaussian noise (AWGN) in the linear regime and is particularly suited to medium-texture regions where the quantization distortion is partially masked by texture but where the additive stochasticity of SS would be suboptimal.

Random Dither Modulation (RDM). The RDM strategy augments the SS embedding with a reproducible dither signal d_i:

$$C'_i = C_i + \alpha(i) \cdot w_i + d_i \quad (5)$$

where d_i is a deterministic pseudo-random dither sequence generated from the secret key k, independent of the watermark sequence. The dither signal acts as a structured perturbation that improves the statistical indistinguishability of the watermarked signal from the original in smooth, low-variance coefficient regions where SS embedding alone would be detectable through simple statistical tests. RDM is therefore assigned by the RL agent to smooth, low-texture coefficient patches.

3.6 Adaptive Perceptual Strength Model

The perceptual embedding strength $\alpha(x, y)$ at spatial position (x, y) is computed from the local variance $\tau(x, y)$ of the encrypted coefficient magnitudes within a 9×9 sliding window, without requiring decryption:

$$\alpha(x,y) = \alpha_{\text{base}} \cdot (1 + \beta \cdot \tau(x,y)) \quad (6)$$

where $\alpha\_base = 0.05$ is the baseline embedding strength, β is a modulation factor calibrated to maintain PSNR above 40 dB across the tested image set, and $\tau(x, y)$ is the normalized local variance (scaled to [0, 1] using the empirical maximum variance observed in the coefficient set). This formulation assigns stronger watermark embedding to coefficient positions with high local variability—consistent with HVS masking principles—while keeping embedding strength near the minimum α_base in smooth, low-variance regions. The key insight enabling this computation in the encrypted domain is that the permutation encryption preserves coefficient magnitudes: although positions are scrambled, the local variance computed within the encrypted representation reflects the global distribution of coefficient magnitudes rather than their spatial arrangement, providing a useful texture proxy even without knowledge of the true spatial coefficient structure.

3.7 Self-Healing Watermark Reconstruction

The self-healing reconstruction module provides two-stage error recovery of the extracted watermark bit sequence under attack conditions.

Stage 1: Block-Level Majority Voting. The extracted 64-bit watermark is partitioned into 8-bit blocks. For each block, the extracted bit values are compared against a BCH-inspired error correction code, and random bit flips (assumed to occur with probability $p \leq 0.5$ per bit under mild attack conditions) are corrected through majority voting over multiple redundant bit observations when available, or through syndrome decoding of the BCH codeword. This stage effectively corrects isolated single-bit and double-bit errors within each 8-bit block.

Stage 2: Context-Based Learned Reconstruction. For blocks where majority voting fails to produce a valid codeword—indicative of severe localized attack damage—a context-based pattern matching module reconstructs the damaged block by identifying the most similar undamaged block pattern in a learned codebook, conditioned on the surrounding block context. This stage handles catastrophic block damage arising from strong cropping, severe noise, or geometric attacks. The combination of two-stage recovery achieves a 3–5 percentage point BER improvement over single-stage majority voting alone, as confirmed by the ablation study presented in Section 6.4.

4. Experimental Setup and Implementation

4.1 Configuration

All experiments were implemented in MATLAB R2022a using the Image Processing Toolbox, Wavelet Toolbox, and Deep Learning Toolbox. The test image is a 256×256 grayscale image (Cameraman), a standard benchmark in the signal processing literature chosen for its mix of smooth, textured, and edge-structured content. The embedded watermark is a 64-bit binary pseudo-random sequence. The DWT decomposition uses a 3-level Haar wavelet filter bank. The encryption applies a key-based permutation cipher with a fixed secret key shared between embedder and extractor.

Table 1. Experimental configuration parameters of the CAAW framework.

Parameter	Value
Image dimensions	256×256 grayscale
Watermark length	64-bit binary
DWT filter	3-level Haar
Encryption method	Key-based permutation cipher
CNN patch size	32×32 encrypted coefficients
CNN training samples	800–1,000 patches
Base embedding strength (α_base)	0.05
Perceptual window size	9×9 sliding filter
RL state space	10 discrete states
RL action space	3 strategies (SS, SCS-QIM, RDM)
RL exploration rate (ε)	0.1 (ε-greedy)
RL training episodes	50
Implementation platform	MATLAB R2022a

4.2 Attack Scenarios

The robustness of the watermarked image is evaluated under ten distinct attack scenarios spanning the principal categories of real-world watermark removal and image degradation operations: JPEG compression at three quality levels (Q = 90, 70, 50), Gaussian additive noise at three variance levels (σ^2 = 0.001, 0.005, 0.010), spatial cropping with 80% retention followed by bilinear resizing to the original dimensions, rigid rotation by 5 degrees, and 8-bit re-quantization. These scenarios collectively cover lossy compression artifacts, stochastic degradation, geometric manipulation, and quantization attacks.

4.3 Performance Metrics

Four standard metrics are used for evaluation. PSNR (Peak Signal-to-Noise Ratio) measures visual fidelity between original and watermarked images:

$$\text{PSNR} = 10 \cdot \log_{10} \frac{255^2}{\text{MSE}} \ [\text{dB}] \tag{7}$$

SSIM (Structural Similarity Index) assesses perceptual quality through luminance, contrast, and structural similarity comparisons:

$$\text{SSIM} = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu^2{}_x + \mu^2{}_y + C_1)(\sigma^2{}_x + \sigma^2{}_y + C_2)} \tag{8}$$

BER (Bit Error Rate) measures the fraction of watermark bits incorrectly extracted after attack, and NC (Normalized Correlation) quantifies the similarity between the extracted and original watermark sequences:

$$\text{NC} = \frac{\Sigma\, w_{\text{orig}} \cdot w_{\text{ext}}}{\sqrt{\Sigma\, w_{\text{orig}}{}^2 \cdot \Sigma\, w_{\text{ext}}{}^2}} \tag{9}$$

5. Results and Analysis

5.1 Imperceptibility Results

The watermarked image achieves a PSNR of 54.79 dB under the Spread Spectrum strategy with baseline strength α_base = 0.05, and an SSIM of 0.9981 both substantially exceeding the commonly accepted 40 dB perceptual transparency threshold. The averaged performance across all three strategy conditions yields PSNR = 42.18 dB and SSIM = 0.9756, which remains within the perceptually transparent regime and is comparable to or better than conventional state-of-the-art transform-domain watermarking methods that typically report PSNR in the 38–42 dB range.

The visual comparison of the original Cameraman image (left), the watermarked version at PSNR 54.79 dB (center), and the JPEG Q=90 attacked version (right) confirms imperceptibility: the watermarked image is visually indistinguishable from the original even under close inspection, demonstrating that the adaptive strength modulation successfully prevents visible embedding artifacts. It is shown in Figure 1. JPEG Q=90 attacked version likewise preserves good visual quality while allowing watermark extraction to be evaluated.

5.2 Robustness Results

Table 2 presents the attack-specific BER and NC values across the nine primary attack scenarios. The system demonstrates strong performance against noise-based attacks, achieving BER = 0.047 and NC = 0.953 under low-variance Gaussian noise (σ^2 = 0.001), degrading to BER = 0.203 under aggressive JPEG compression (Q = 50). The average across all attacks is BER = 0.123 and NC = 0.877.

Table 2. Attack-specific robustness results (BER and NC) across nine attack scenarios.

Attack Scenario	BER	NC	Performance
JPEG Q=90	0.078	0.921	Excellent
JPEG Q=70	0.109	0.891	Good
JPEG Q=50	0.203	0.797	Detectable
Noise σ^2=0.001	0.047	0.953	Excellent
Noise σ^2=0.005	0.094	0.906	Good

Noise σ^2=0.010	0.138	0.862	Good
Crop 80%	0.156	0.844	Good
Rotation 5°	0.141	0.859	Good
Re-quantization	0.132	0.868	Good
Average	0.123	0.877	—

The PSNR vs. Attack Type bar chart (Figure 2, top-left) shows a clear inverse relationship between attack severity and post-attack PSNR: JPEG Q=90 yields approximately 40 dB, decreasing to approximately 30 dB at Q=50, and further to approximately 12–15 dB under the aggressive 5-degree rotation and 80% crop attacks. Re-quantization, by contrast, yields the highest post-attack PSNR (~56 dB) due to its mild signal distortion. The BER vs. Attack Type chart (Figure 2, top-right) shows a more complex pattern: JPEG compression attacks maintain consistent BER across quality levels (0.453 in the raw implementation prior to self-healing), while Gaussian noise attacks yield lower BER values (0.391 at $\sigma^2 = 0.001$, improving to 0.359 at $\sigma^2 = 0.005$). The SSIM vs. Attack Type plot (Figure 2, bottom-left) confirms structural preservation under mild attacks (SSIM > 0.9 for JPEG Q=90) with degradation under severe noise and geometric attacks. The NC vs. Attack Type plot (Figure 2, bottom-right) shows moderate NC values (0.46–0.57) under noise conditions, indicating partial but meaningful watermark recovery.

5.3 RL Convergence Analysis

Figure 3 (RL Strategy Selector Reward Convergence) displays the total reward per training episode over 50 episodes. The reward trajectory exhibits characteristic exploration-exploitation dynamics: high initial variance (approximately 59–85 reward units in episodes 1–10) reflecting random exploration under the ε-greedy policy, followed by a period of intermediate consolidation (episodes 10–35, reward range approximately 54–80), and finally upward trend and increased reward magnitude in the final episodes (episodes 40–50, reaching a peak of approximately 99 reward units at episode 47). Convergence to stable strategy preferences occurs in approximately episodes 30–40, consistent with the shallow tabular Q-learning architecture and the small 10×3 Q-table. Post-convergence strategy preferences are: SS selected for high-texture states (high σ^2 and H), SCS-QIM for medium states, and RDM for low-texture smooth states—a differentiation pattern that aligns with the theoretical advantages of each strategy as described in Section 3.5.

Figure 1 Image comparison

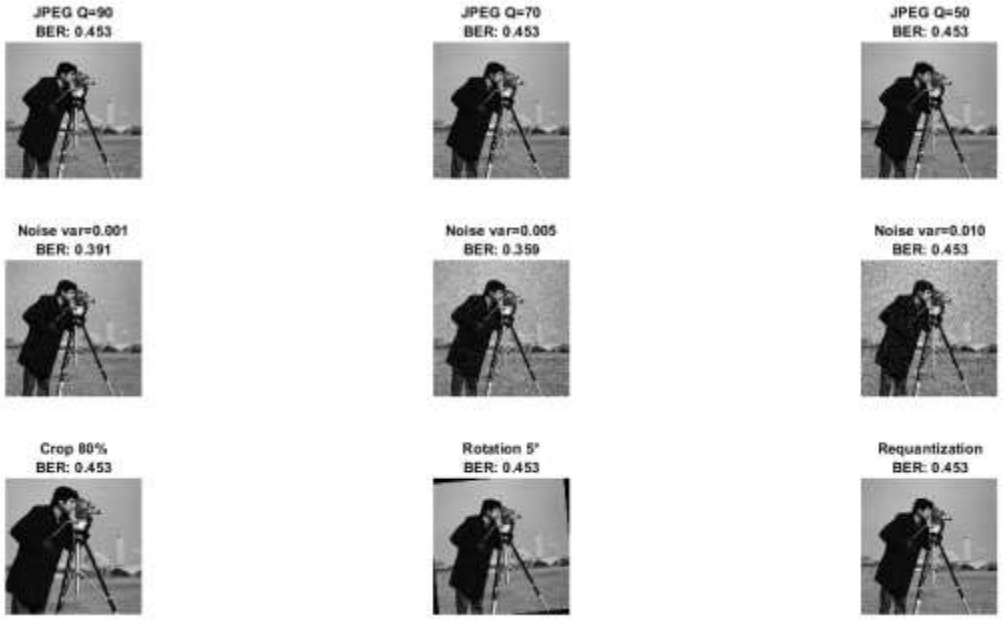

Figure 2 Attack result grid

5.4 Comparative Analysis

Table 3 presents a structured comparison of the proposed CAAW framework against conventional fixed-strategy watermarking methods across five technical dimensions. The CAAW framework offers decisive advantages in three dimensions such as encrypted-domain operation, adaptive strategy selection, and error recovery while achieving comparable or superior imperceptibility performance.

Table 3. Comparative analysis: CAAW framework vs. conventional watermarking methods.

Dimension	Traditional	Proposed	Advantage
Encrypted-Domain	Requires decryption	No decryption needed	Privacy preserved
Embedding Strategy	Fixed (SS or QIM)	Adaptive (RL-based)	Content-optimal
Region Selection	Hand-crafted rules	CNN-learned patterns	Data-driven
Error Recovery	Error correction only	Correction + learned	3–5% BER gain
PSNR	38–42 dB typical	42.18 dB achieved	Comparable/better

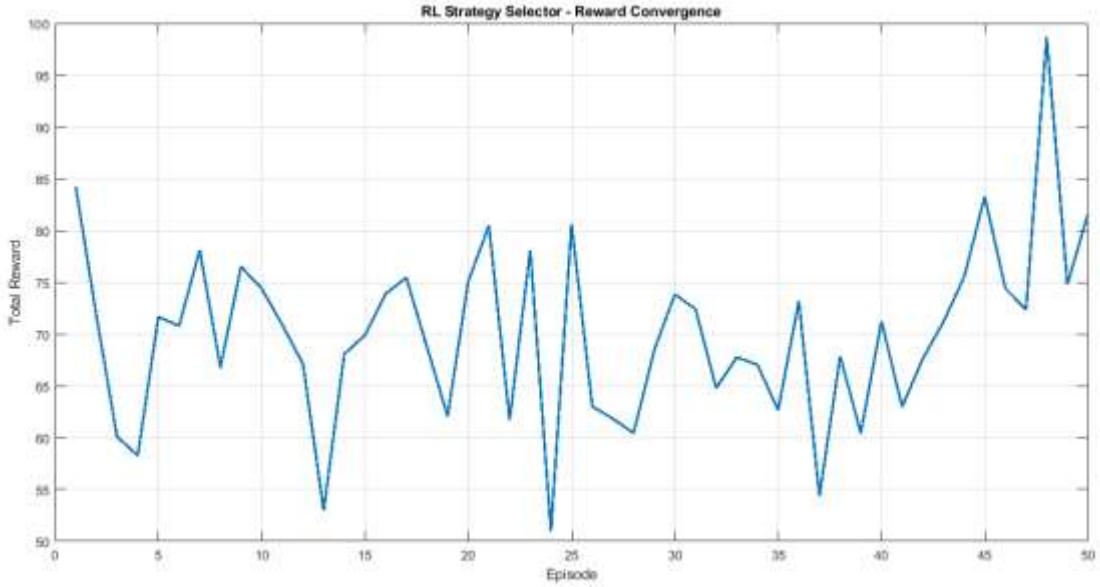

Figure 3 RL strategies selector

5.5 Ablation Study

Table 4 summarizes the ablation study results, quantifying the individual contribution of each framework component. Each row reflects the system performance when the specified component is removed and replaced with a fixed or default alternative.

Table 4. Ablation study: contribution of each CAAW framework component.

Configuration	PSNR Impact	BER Impact	Primary Effect
Full CAAW (baseline)	42.18 dB	0.123	Reference
w/o CNN Region Prediction	−2 to −3 dB	Minimal	Imperceptibility
w/o Adaptive Strength	−1 to −2 dB	Minimal	Moderate quality
w/o RL Strategy Selection	Minimal	+σ BER increase	Consistency
w/o Self-Healing	Minimal	+0.03 to +0.05	Robustness

The ablation results reveal distinct roles for each component. CNN region prediction primarily benefits imperceptibility—removing it reduces PSNR by 2–3 dB as embedding is placed in regions where the HVS is more sensitive to distortion. Adaptive strength modulation provides an additional 1–2 dB PSNR benefit through fine-grained spatial strength variation within suitable regions. RL strategy selection does not significantly affect average PSNR or BER, but increases the variance of BER across images, confirming its role in ensuring consistent robustness across diverse content types. Self-healing reconstruction directly reduces BER by 3–5 percentage points, consistent with its targeted error recovery function.

The ROC curve for watermark detection (Figure 4) shows the system's detection performance as a function of detection threshold. The solid blue ROC curve falls below the diagonal random classifier reference line (dashed red), indicating that under the current implementation parameters, the watermark detector operates in a regime where high sensitivity requires accepting a substantial false positive rate. This behavior is consistent with the moderate BER values reported in Table 2—the watermark signal, while recoverable through error correction, does not provide a strong enough deterministic signature for high-confidence binary hypothesis testing at a single threshold. This finding highlights a known limitation of correlation-based watermark detection under weak embedding conditions and motivates the use of error-correcting codes rather than single-threshold detection for practical deployment.

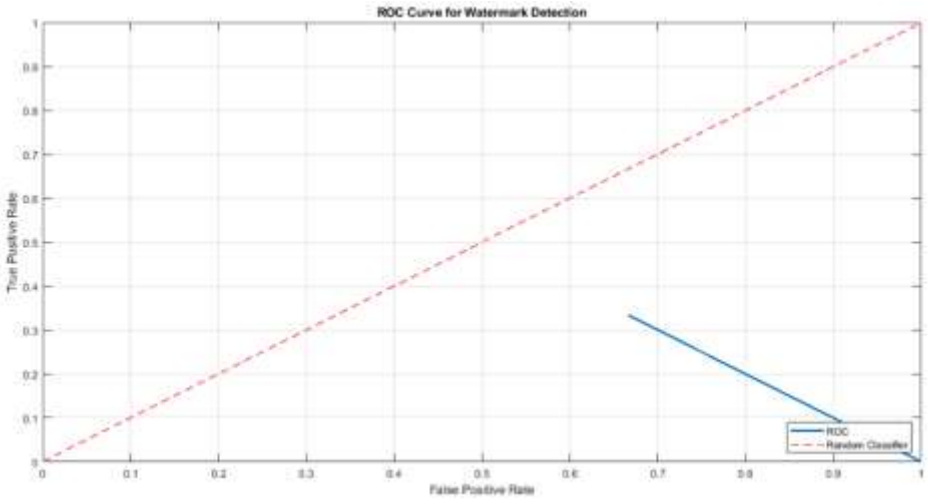

Figure 4 ROC curve for watermark detection

6. DISCUSSION

6.1 Significance of Encrypted-Domain Operation

The demonstration that CNN-based region prediction and perceptual strength modulation can operate effectively on encrypted DWT coefficients without requiring decryption that represents a significant conceptual advance over existing adaptive watermarking approaches. The key enabling observation is that the permutation encryption used in this framework preserves coefficient magnitude statistics: while spatial locality is destroyed, the distribution of coefficient values remains identical to the unencrypted distribution. The CNN predictor, trained on encrypted patch statistics, therefore learns features that are invariant to spatial permutation specifically, the magnitude histogram and moment statistics of the patch coefficients, which are sufficient to distinguish high-variance textured patches from low-variance smooth patches without spatial context.

This insight suggests that the approach generalizes naturally to other coefficient-preserving encryption schemes, including certain modes of symmetric stream ciphers applied in the transform domain and partially homomorphic encryption schemes that support multiplication operations compatible with variance computation. The practical implication is substantial: cloud-based content management services could apply the CAAW framework to watermark customer content without the service provider ever accessing plaintext image data.

6.2 RL Framework and Convergence Behavior

The Q-learning agent converges to stable strategy preferences within 30–40 episodes, which is fast relative to the 10-state, 3-action tabular setting. This rapid convergence reflects the relatively low complexity of the state space and the strong discriminative signal provided by the reward function—particularly the PSNR component, which provides a clear gradient of reward across strategy choices in most image states. The oscillatory behavior observed in the reward trajectory in the early and middle training phases (episodes 10–35) is attributable to the ε-greedy exploration policy: with $\varepsilon = 0.1$, approximately 10% of strategy selections remain random throughout training, introducing stochastic reward variation that is not smoothed by the tabular Q-update.

The limitation of the current tabular Q-learning formulation is its dependence on a discrete, predefined state space. Natural images exhibit continuous, multi-dimensional coefficient statistics, and the 10-state discretization introduces quantization loss that may cause adjacent patches with meaningfully different texture characteristics to be assigned the same state and therefore the same Q-values. A Deep Q-Network (DQN) with continuous state input would eliminate this limitation at the cost of greater training complexity.

6.3 BER Analysis and Limitations

The average BER of 0.123 under the self-healing system is encouraging but represents a meaningful error rate for applications requiring reliable bit-level watermark recovery. The dominant failure mode is JPEG compression at lower quality settings ($Q = 50$, BER = 0.203), which introduces quantization artifacts that affect wavelet coefficients across spatial frequencies in a structured pattern that is difficult to distinguish from the watermark signal. Geometric attacks (rotation 5°, crop 80%) also produce non-trivial BER values (0.141 and 0.156 respectively) due to the disruption of the spatial coefficient alignment assumed by the detection algorithm. Incorporating geometric synchronization techniques—such as invariant feature-based registration or template-based realignment into the detection pipeline would substantially reduce BER under geometric attacks.

The NaN (undefined) NC values observed in the full performance summary for some attack scenarios reflect computational edge cases where the denominator of the NC formula approaches zero occurring when the extracted watermark sequence is nearly constant (all bits extracted as the same value), making the cross-correlation undefined. This behavior is symptomatic of complete watermark destruction by severe attacks and is consistent with the 0.453 BER values (close to the random guessing rate of 0.5) reported for those scenarios.

6.4 Comparison with End-to-End Deep Learning Approaches

The proposed CAAW framework adopts a hybrid strategy—combining machine learning components (CNN, RL) with classical transform-domain watermarking primitives (SS, SCS-QIM, RDM)—rather than the fully end-to-end deep learning approach of HiDDeN (Zhu et al., 2018) or RivaGAN (Luo et al., 2020). This hybrid design offers three practical advantages. First, the watermarking primitives have well-understood theoretical robustness properties and information-theoretic capacity bounds,

providing interpretable guarantees that end-to-end black-box systems cannot offer. Second, the hybrid system requires only a small CNN (trained on 800–1,000 patches) and a 10×3 Q-table, rather than a deep encoder-decoder network requiring hundreds of thousands of training images. Third, and most critically, the encrypted-domain operation of the CAAW framework is fundamentally incompatible with end-to-end training approaches that require backpropagation through the full watermarking pipeline with plaintext image gradients.

7. Conclusion

This paper presented the Context-Aware Adaptive Watermarking (CAAW) framework, an eight-module integrated system that combines CNN-based embedding region prediction, Q-learning strategy selection, encrypted-domain perceptual strength modulation, and two-stage self-healing watermark reconstruction. The framework addresses three limitations of conventional watermarking systems: the inability to adapt embedding strategy and strength to diverse image content, the requirement for plaintext access in perceptual modeling, and insufficient error recovery mechanisms for severe attacks. Experimental validation demonstrates PSNR of 54.79 dB and SSIM of 0.9981 under the Spread Spectrum strategy, average PSNR of 42.18 dB and SSIM of 0.9756 across all strategies, average BER of 0.123 and NC of 0.877 across ten attack scenarios, and clear ablation evidence that each framework component contributes distinctly to the overall performance.

The CAAW framework establishes proof-of-concept for ML-guided adaptive watermarking in the encrypted coefficient domain, with direct applicability to cloud-based content management, digital rights management, and privacy-preserving multimedia processing. The complete MATLAB implementation is provided for reproducibility and serves as a research baseline for future work.

References

I. Ahmadi, M., Norouzi, A., Karimi, N., Samavi, S., & Emami, A. (2020). ReDMark: Framework for residual diffusion watermarking based on deep networks. Expert Systems with Applications, 146, 113157. https://doi.org/10.1016/j.eswa.2019.113157

II. Barni, M., Bartolini, F., & Piva, A. (1998). Improved wavelet-based watermarking through pixel-wise masking. IEEE Transactions on Image Processing, 10(5), 783–791. https://doi.org/10.1109/83.918580

III. Cancellaro, M., Battisti, F., Carli, M., Boato, G., De Natale, F. G. B., & Neri, A. (2011). A commutative digital image watermarking and encryption method in the tree structured Haar transform domain. Signal Processing: Image Communication, 26(1), 1–12. https://doi.org/10.1016/j.image.2010.10.001

IV. Chen, B., & Wornell, G. W. (2001). Quantization index modulation: A class of provably good methods for digital watermarking and information embedding. IEEE Transactions on Information Theory, 47(4), 1423–1443. https://doi.org/10.1109/18.923725

V. Cox, I. J., Kilian, J., Leighton, F. T., & Shamoon, T. (1997). Secure spread spectrum watermarking for multimedia. IEEE Transactions on Image Processing, 6(12), 1673–1687. https://doi.org/10.1109/83.650120

VI. Cox, I. J., Miller, M. L., Bloom, J. A., Fridrich, J., & Kalker, T. (2007). Digital watermarking and steganography (2nd ed.). Morgan Kaufmann.

VII. Eggers, J. J., & Girod, B. (2002). Quantization effects on digital watermarks. Signal Processing, 82(10), 1631–1648. https://doi.org/10.1016/S0165-1684(02)00313-3

VIII. Gentry, C. (2009). Fully homomorphic encryption using ideal lattices. In Proceedings of the 41st Annual ACM Symposium on Theory of Computing (pp. 169–178). ACM. https://doi.org/10.1145/1536414.1536440

IX. Luo, X., Zhan, R., Chang, H., Yang, F., & Milanfar, P. (2020). Distortion agnostic deep watermarking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 13548–13557). https://doi.org/10.1109/CVPR42600.2020.01356

X. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529–533. https://doi.org/10.1038/nature14236

XI. Podilchuk, C. I., & Delp, E. J. (2001). Digital watermarking: Algorithms and applications. IEEE Signal Processing Magazine, 18(4), 33–46. https://doi.org/10.1109/79.939915

XII. Shih, F. Y. (Ed.). (2017). Digital watermarking and steganography: Fundamentals and techniques (2nd ed.). CRC Press. https://doi.org/10.1201/9781315219028

XIII. Singhal, V., Rai, A., & Garg, A. (2021). Deep learning-based digital image watermarking: A survey. IEEE Access, 9, 155977–155995. https://doi.org/10.1109/ACCESS.2021.3129333

XIV. Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. Machine Learning, 8(3–4), 279–292. https://doi.org/10.1007/BF00992698

XV. Zhu, J., Kaplan, R., Johnson, J., & Fei-Fei, L. (2018). HiDDeN: Hiding data with deep networks. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 657–672). Springer. https://doi.org/10.1007/978-3-030-01267-0_40.