Digital Watermarking: From Traditional Techniques to Machine Learning Approaches

R12921067 電機所碩二 林想

Outline

1. Introduction to Digital Watermarking
2. Traditional Digital Watermarking Method
    a. Image watermarking
    b. Spatial-based vs Transform-based
    c. Discrete cosine transform (DCT)
    d. Discrete wavelet transform (DWT)
    e. Proposed watermarking scheme
    f. Experiment result
3. ML Digital Watermarking method
    a. Supervised GAN watermarking
4. Challenges and Limitations
5. Conclusions

1. Introduction to Digital Watermarking

- Definition
  - Digital watermarking: Hidden information embedded within digital media
  - Technique for protecting intellectual property and ensuring data integrity
- Key Applications
  - Copyright protection
  - Authentication
  - Integrity verification
  - Content tracking
- Core Challenges
  - **Robustness**: Survive various transformations
  - **Imperceptibility**: Invisible to human perception
  - **Security**: Resistant to intentional attacks

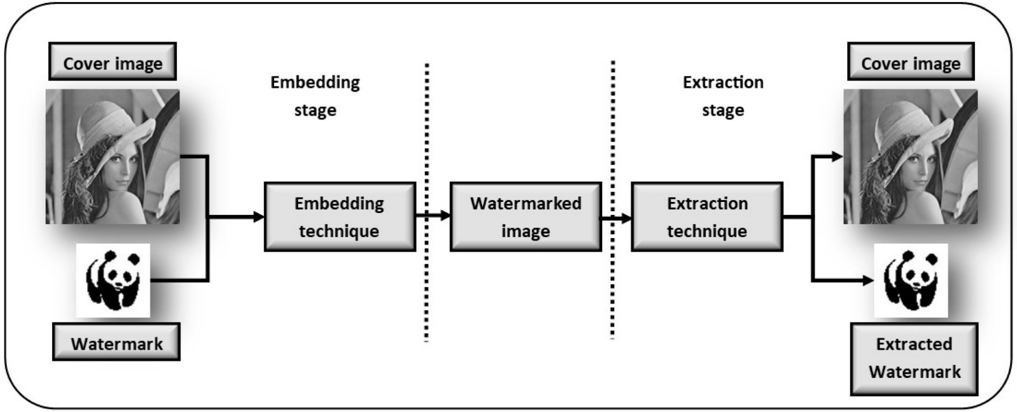
2. Traditional Digital Watermarking Method:

A blind and robust color image watermarking scheme based on DCT and DWT domains [1]

[1] Mohammed, Abdulhakeem O., et al. "A blind and robust color image watermarking scheme based on DCT and DWT domains." *Multimedia Tools and Applications* 82.21 (2023): 32855-32881.

2.a Image watermarking

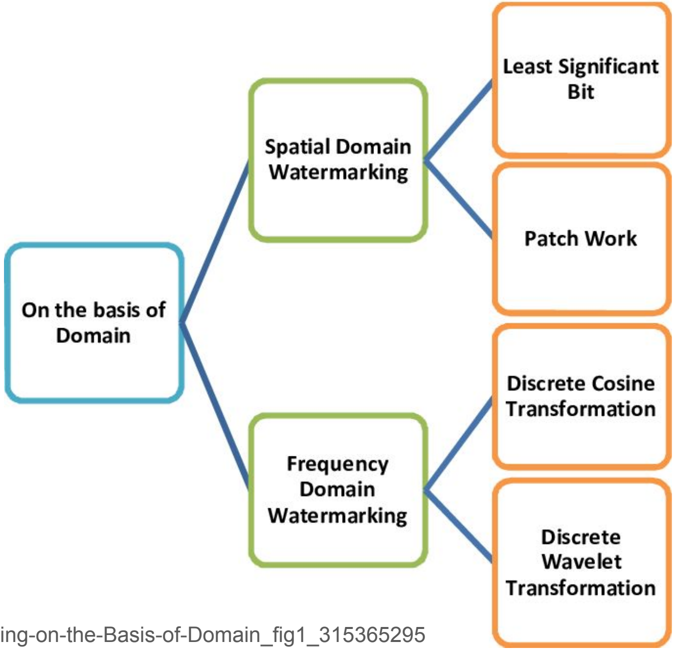
- Two main processes: embedding and extraction
- Robust vs Fragile
  - Robust watermarking is developed to withstand attacks, ensuring that the embedded watermark survives modification attempts. Used in copyright protection.
  - Fragile watermarking is devised to be vulnerable to any modification. Used in tamper detection and authentication.



https://www.mdpi.com/2076-3417/13/10/6105

2.a Image watermarking

- Blind vs Non-blind
  - Blind approaches require neither the original host image nor the watermark.
  - Non-blind require either or both
- Spatial-based vs Transform-based


On the basis of Domain
Spatial Domain Watermarking
Frequency Domain Watermarking
Least Significant Bit
Patch Work
Discrete Cosine Transformation
Discrete Wavelet Transformation

https://www.researchgate.net/figure/Types-of-Watermarking-on-the-Basis-of-Domain_fig1_315365295

2.b Spatial-based vs Transform-based

- Spatial-based [2], [3]
    - Histogram shifting, Least Significant Bits (LSBs), prediction error expansion, and difference expansion
    - LSB: modifies the least significant bits of the image’s pixel data(0-255) to embed the watermark
    - Pros:
        - Simple implementation
        - High capacity
    - Cons:
        - Low robustness
        - Vulnerable to image processing

[2] Craver, Scott, et al. "Resolving rightful ownerships with invisible watermarking techniques: Limitations, attacks, and implications." *IEEE Journal on Selected areas in Communications* 16.4 (1998): 573-586.
[3] Yeung, Minerva M., and Fred Mintzer. "An invisible watermarking technique for image verification." *Proceedings of international conference on image processing*. Vol. 2. IEEE, 1997.

2.b Spatial-based vs Transform-based

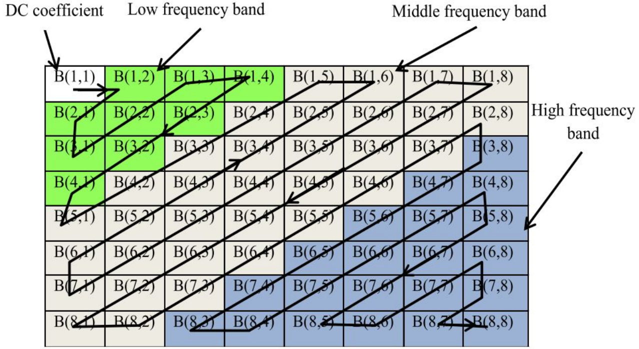
- Transform-based
  - In the transform domain, watermark embedding involves converting the host image into the frequency domain, modifying its coefficients according to predefined rules, and subsequently reconstructing the watermarked image through an inverse transform.
  - Discrete Cosine Transform (DCT), Singular Value Decomposition (SVD), Discrete Fourier Transform (DFT), and Discrete Wavelet Transform (DWT)
  - Greater robustness compared to spatial-based methods
  - Robust against various attacks, including compression, filtering, blurring, and rotation

2.c Discrete cosine transform (DCT)

- Low frequency: Image quality deteriorates significantly during embedding.
- High frequency: Vulnerable to typical attacks like JPEG compression.
- Middle frequency: Best suited for embedding due to resistance to lossy compression and minimal impact on the cover image.

$$F(p,q) = \alpha(p)\alpha(q) \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} I(m,n) cos\left[\frac{\pi p(2m+1)}{2M}\right] cos\left[\frac{\pi q(2n+1)}{2N}\right], \quad \frac{0 \le p \le M-1}{0 \le q \le N-1}$$

$$I(m,n) = \sum_{p=0}^{M-1} \sum_{q=0}^{N-1} \alpha(p)\alpha(q) F(p,q) cos\left[\frac{\pi p(2m+1)}{2M}\right] cos\left[\frac{\pi q(2n+1)}{2N}\right], \quad \frac{0 \le m \le M-1}{0 \le n \le N-1}$$

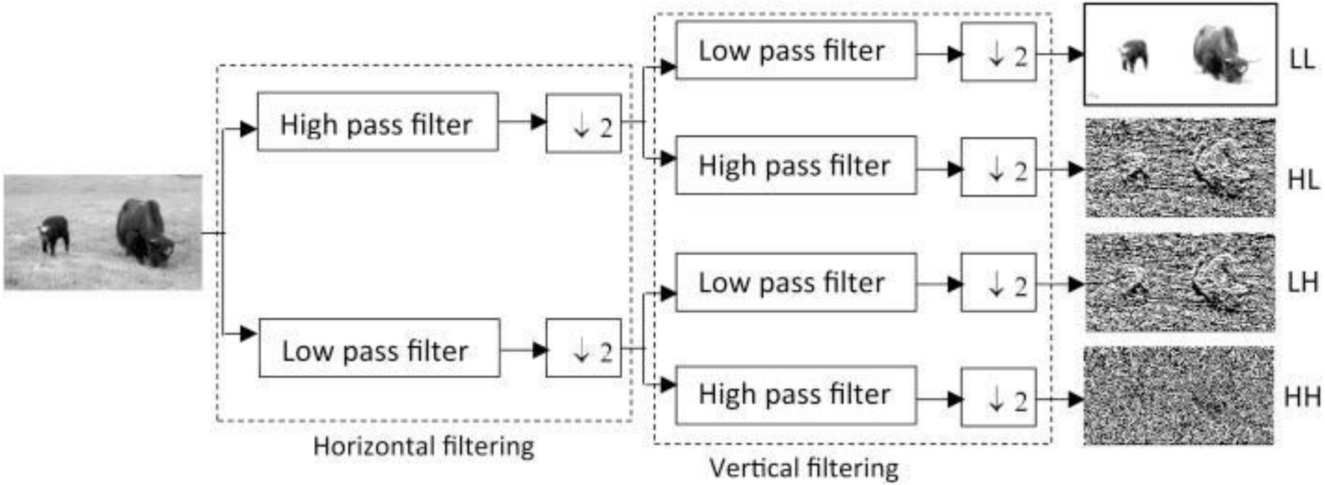


B(1,1)	B(1,2)	B(1,3)	B(1,4)	B(1,5)	B(1,6)	B(1,7)	B(1,8)
B(2,1)	B(2,2)	B(2,3)	B(2,4)	B(2,5)	B(2,6)	B(2,7)	B(2,8)
B(3,1)	B(3,2)	B(3,3)	B(3,4)	B(3,5)	B(3,6)	B(3,7)	B(3,8)
B(4,1)	B(4,2)	B(4,3)	B(4,4)	B(4,5)	B(4,6)	B(4,7)	B(4,8)
B(5,1)	B(5,2)	B(5,3)	B(5,4)	B(5,5)	B(5,6)	B(5,7)	B(5,8)
B(6,1)	B(6,2)	B(6,3)	B(6,4)	B(6,5)	B(6,6)	B(6,7)	B(6,8)
B(7,1)	B(7,2)	B(7,3)	B(7,4)	B(7,5)	B(7,6)	B(7,7)	B(7,8)
B(8,1)	B(8,2)	B(8,3)	B(8,4)	B(8,5)	B(8,6)	B(8,7)	B(8,8)

[4] Parah, Shabir A., et al. "Robust and blind watermarking technique in DCT domain using inter-block coefficient differencing." *Digital Signal Processing* 53 (2016): 11-24.

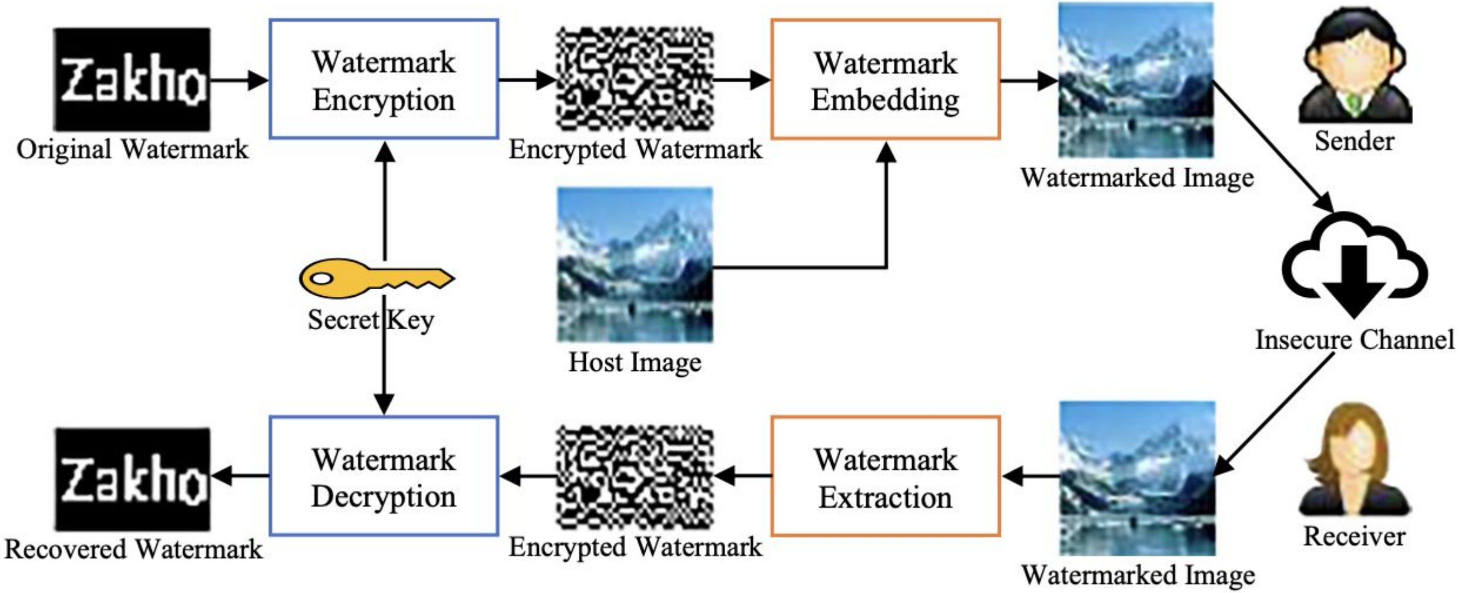
2.d Discrete wavelet transform (DWT)

- Haar wavelet is used here due to its simplicity and low cost.
- LL: Contains most image energy; embedding here reduces image quality.
- HH: Vulnerable to compression attack



https://medium.com/@koushikc2000/2d-discrete-wavelet-transformation-and-its-applications-in-digital-image-processing-using-matlab-1f5c68672de3

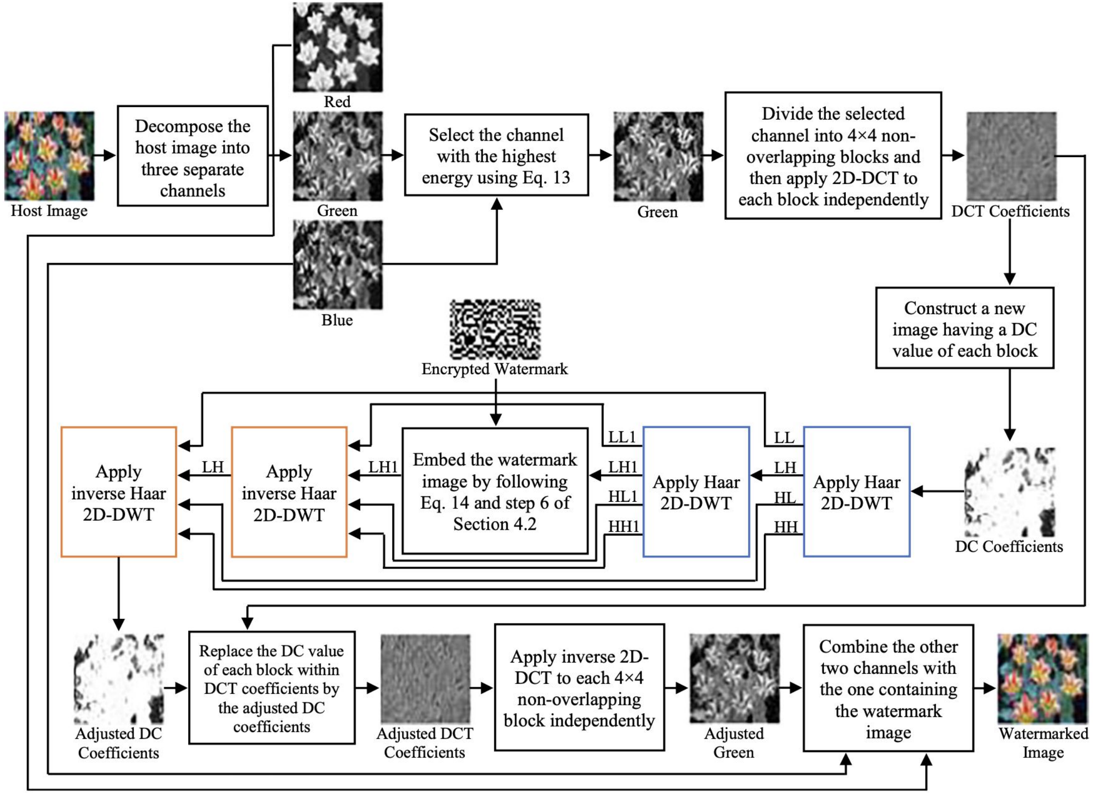
2.e Proposed watermarking scheme



2.e Proposed watermarking scheme

- Combine DCT and DWT
    - DCT
        - DCT is highly resistant to signal processing attacks such as filtering, blurring, and JPEG compression
        - low computational cost and high-energy compaction
    - DWT
        - DWT can decompose an image into different resolutions while capturing its time and frequency representations
        - The DWT allows the decomposition of the image into different frequency sub-bands (LL, LH, HL, HH). By embedding the watermark in the LH sub-band, which is less sensitive to modifications, the scheme can achieve a better balance between robustness and imperceptibility.
        - The further decomposition of the LH sub-band into LH1 provides additional flexibility for watermark embedding

2.e Proposed watermarking scheme - embedding



Decompose the host image into three separate channels

Select the channel with the highest energy using Eq. 13

Divide the selected channel into 4×4 non-overlapping blocks and then apply 2D-DCT to each block independently

Construct a new image having a DC value of each block

Apply Haar 2D-DWT

Apply Haar 2D-DWT

Embed the watermark image by following Eq. 14 and step 6 of Section 4.2

Apply inverse Haar 2D-DWT

Apply inverse Haar 2D-DWT

Replace the DC value of each block within DCT coefficients by the adjusted DC coefficients

Apply inverse 2D-DCT to each 4×4 non-overlapping block independently

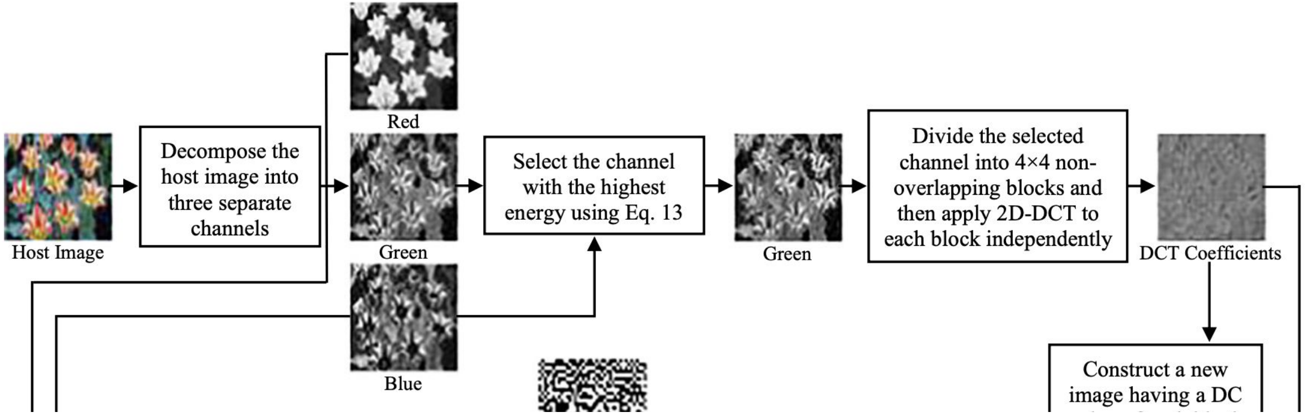
Combine the other two channels with the one containing the watermark image

2.e Proposed watermarking scheme - embedding

1. Decompose RGB Image
2. Calculate Energy:
   a. Formula:

$$E = \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} P(i,j)$$

3. Divide and Transform



Decompose the host image into three separate channels

Select the channel with the highest energy using Eq. 13

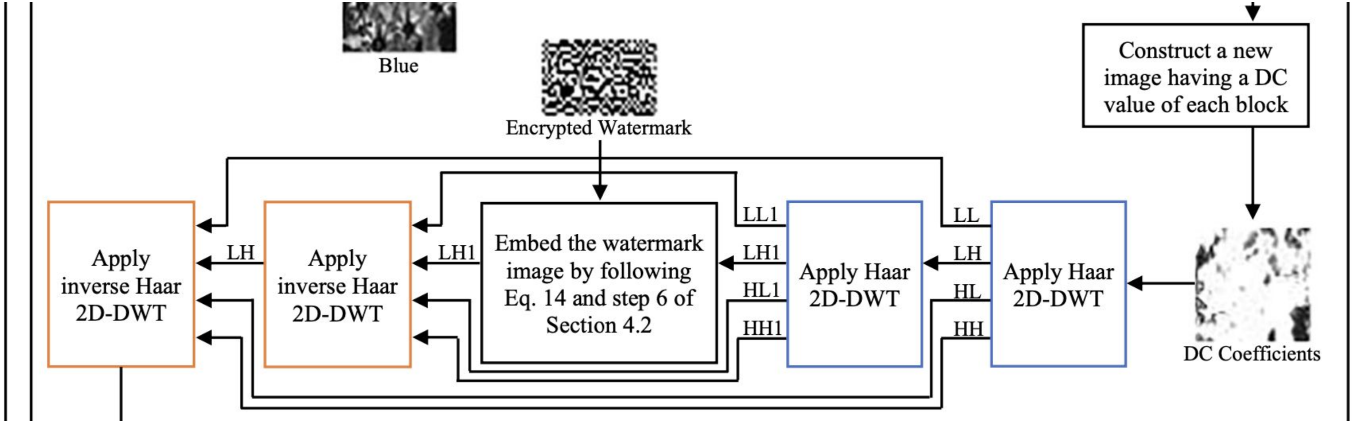
Divide the selected channel into 4×4 non-overlapping blocks and then apply 2D-DCT to each block independently

Construct a new image having a DC

2.e Proposed watermarking scheme - embedding

4. Haar 2D-DWT:

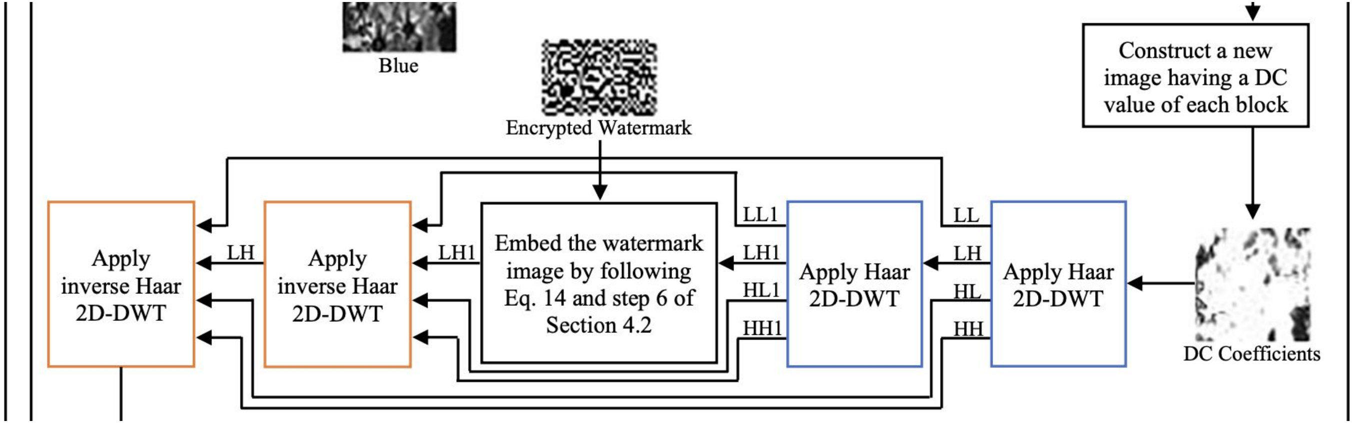
- Apply Haar 2D-DWT to matrix to get sub-bands: LL, LH, HL, HH.
- Reapply Haar 2D-DWT to the LH sub-band to obtain: LL1, LH1, HL1, HH1.


Blue
Encrypted Watermark
Construct a new image having a DC value of each block
Apply inverse Haar 2D-DWT
LH
Apply inverse Haar 2D-DWT
LH1
Embed the watermark image by following Eq. 14 and step 6 of Section 4.2
LL1
LH1
HL1
HH1
Apply Haar 2D-DWT
LL
LH
HL
HH
Apply Haar 2D-DWT
DC Coefficients

2.e Proposed watermarking scheme - embedding

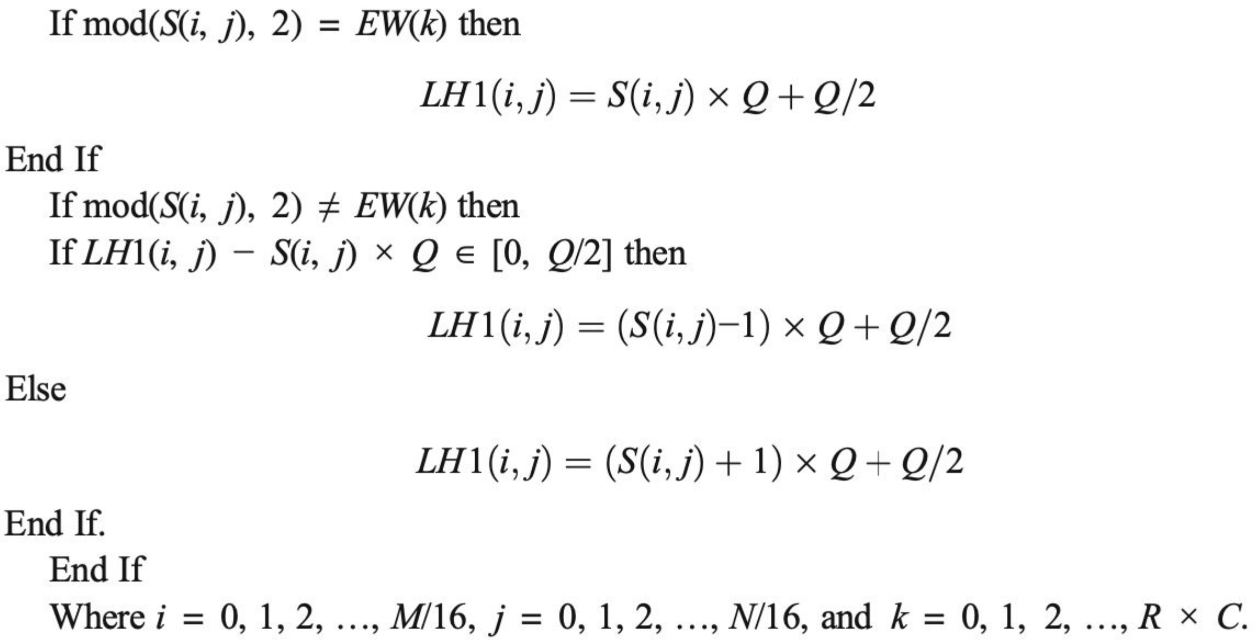
5. Quantization and Matrix Construction:

- Use quantization value Q and $S(i,j) = \left\lfloor \frac{LH1(i,j)}{Q} \right\rfloor$
- Construct matrix S with adjusted LH1 coefficients.


Blue
Encrypted Watermark
Construct a new image having a DC value of each block
Apply inverse Haar 2D-DWT
LH
Apply inverse Haar 2D-DWT
LH1
Embed the watermark image by following Eq. 14 and step 6 of Section 4.2
LL1
LH1
HL1
HH1
Apply Haar 2D-DWT
LL
LH
HL
HH
Apply Haar 2D-DWT
DC Coefficients

2.e Proposed watermarking scheme - embedding

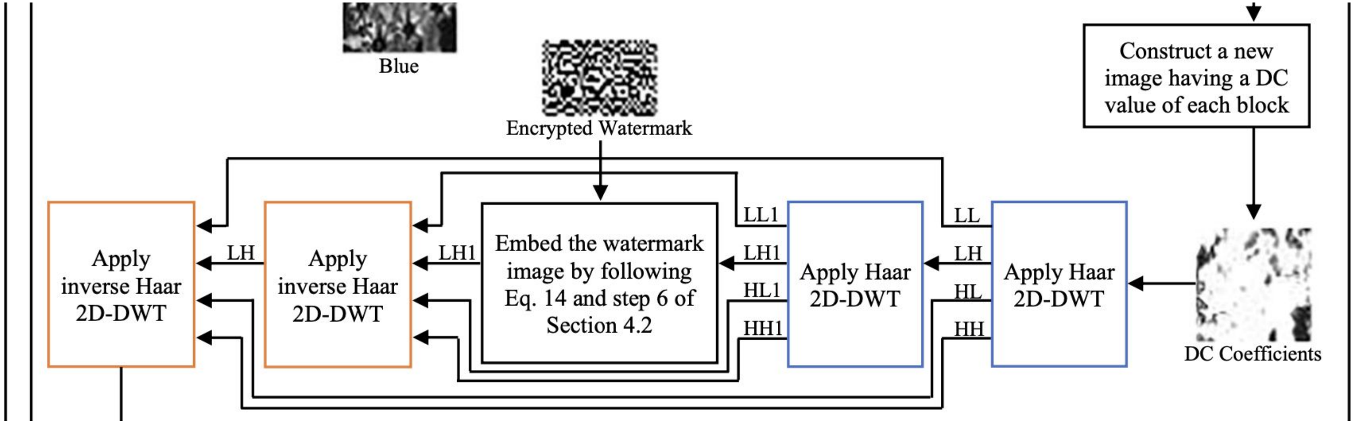
6. Embed the encrypted watermark image (EW) into the LH1 sub-band



```
  If mod($S(i, j)$, 2) = $EW(k)$ then
            $LH1(i,j) = S(i,j) \times Q + Q/2$
End If
  If mod($S(i, j)$, 2) $\neq$ $EW(k)$ then
  If $LH1(i, j) - S(i, j) \times Q \in [0, Q/2]$ then
            $LH1(i,j) = (S(i,j){-}1) \times Q + Q/2$
Else
            $LH1(i,j) = (S(i,j) + 1) \times Q + Q/2$
End If.
  End If
  Where $i = 0, 1, 2, \ldots, M/16$, $j = 0, 1, 2, \ldots, N/16$, and $k = 0, 1, 2, \ldots, R \times C$.
```

2.e Proposed watermarking scheme - embedding

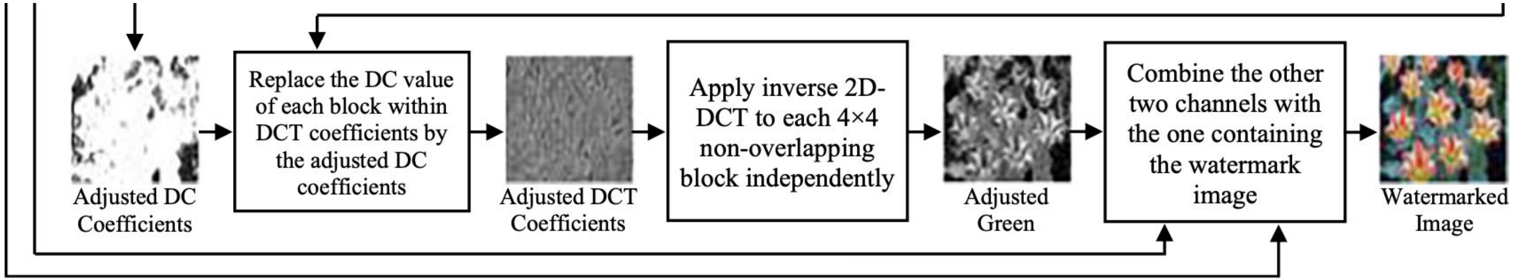
7. Apply inverse Haar 2D-DWT twice to obtained the adjusted matrix D


Blue
Encrypted Watermark
Construct a new image having a DC value of each block
Apply inverse Haar 2D-DWT
LH
Apply inverse Haar 2D-DWT
LH1
Embed the watermark image by following Eq. 14 and step 6 of Section 4.2
LL1
LH1
HL1
HH1
Apply Haar 2D-DWT
LL
LH
HL
HH
Apply Haar 2D-DWT
DC Coefficients

2.e Proposed watermarking scheme - embedding

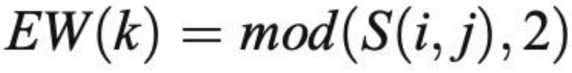
8. Replace DC Values and apply inverse 2D-DCT

9. Create watermarked image


Adjusted DC Coefficients
Replace the DC value of each block within DCT coefficients by the adjusted DC coefficients
Adjusted DCT Coefficients
Apply inverse 2D-DCT to each 4×4 non-overlapping block independently
Adjusted Green
Combine the other two channels with the one containing the watermark image
Watermarked Image

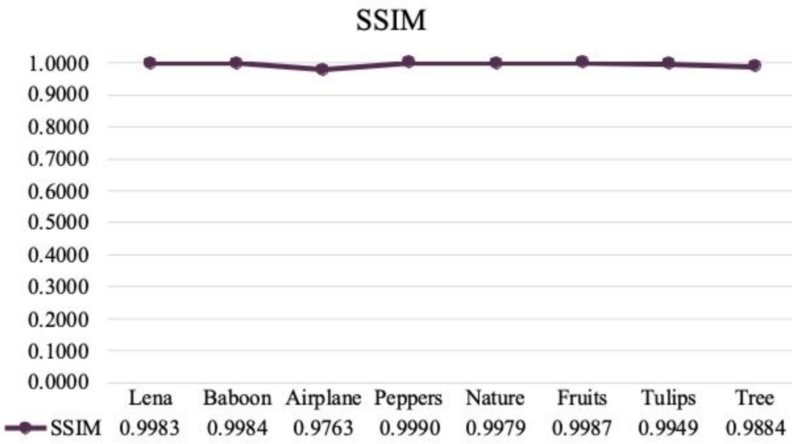
2.e Proposed watermarking scheme - extraction

1. Decompose the watermarked color image (W) into three components as follows: R(Red) = W[:, : , 0], G(Green) = W[:, : , 1], and B(Blue) = W[:, : , 2].
2. Implement steps 2 to 5 of embedding to obtain matrix (S). Where S represents the adjusted coefficients of the LH1 sub-band.
3. Finally, extract the encrypted watermark image (EW) using the eq. below.

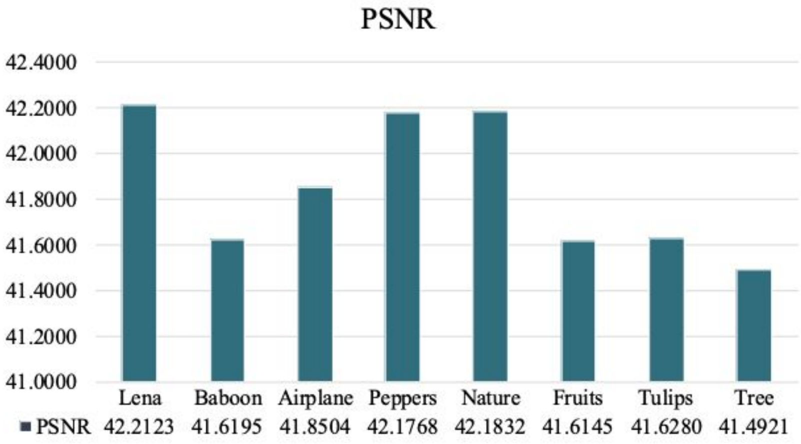


$$EW(k) = mod(S(i,j),2)$$

2.f Experiment result - Interceptibility


SSIM

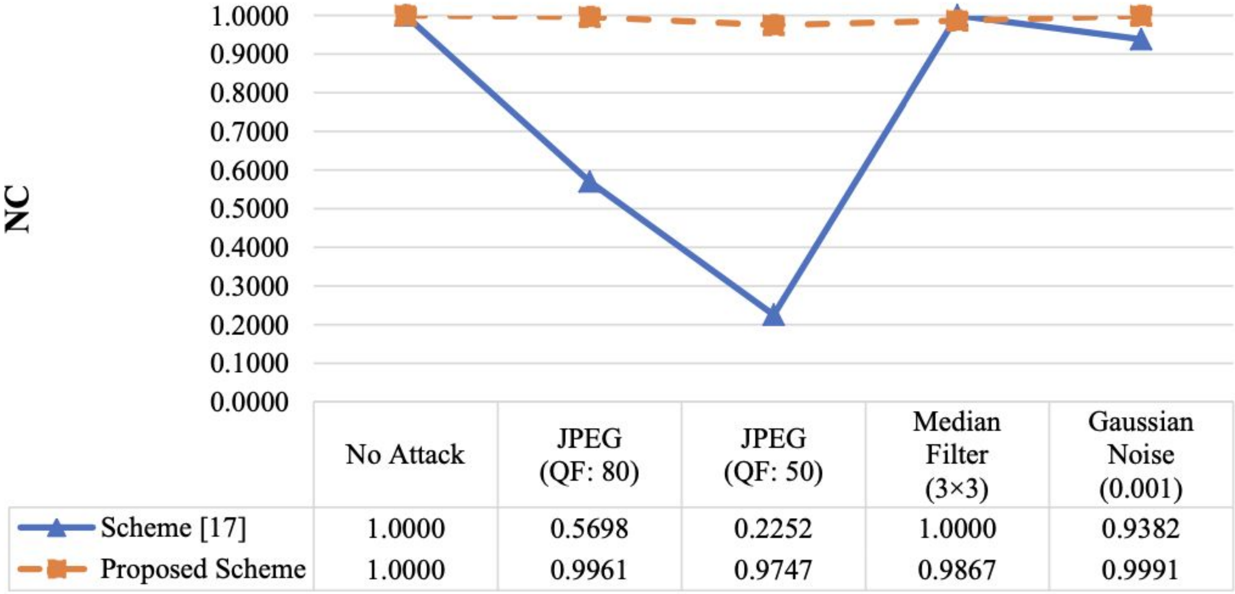
	Lena	Baboon	Airplane	Peppers	Nature	Fruits	Tulips	Tree
SSIM	0.9983	0.9984	0.9763	0.9990	0.9979	0.9987	0.9949	0.9884


PSNR

	Lena	Baboon	Airplane	Peppers	Nature	Fruits	Tulips	Tree
PSNR	42.2123	41.6195	41.8504	42.1768	42.1832	41.6145	41.6280	41.4921

2.f Experiment result - Robustness

- Normalized correlation (NC): similarity ratio between the original watermarked image and the recovered one.



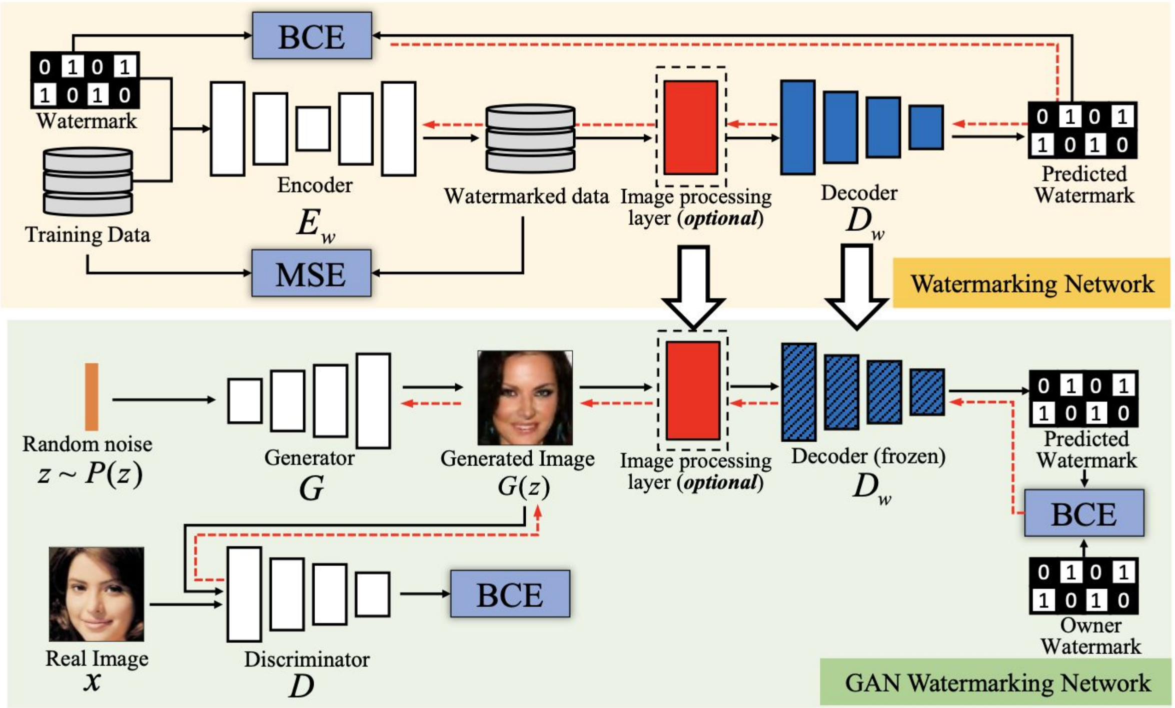
	No Attack	JPEG (QF: 80)	JPEG (QF: 50)	Median Filter (3×3)	Gaussian Noise (0.001)
▲ Scheme [17]	1.0000	0.5698	0.2252	1.0000	0.9382
■ Proposed Scheme	1.0000	0.9961	0.9747	0.9867	0.9991

3. ML Digital Watermarking Method:

Supervised GAN Watermarking for Intellectual Property Protection [5]

[5] Fei, Jianwei, et al. "Supervised gan watermarking for intellectual property protection." *2022 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2022.

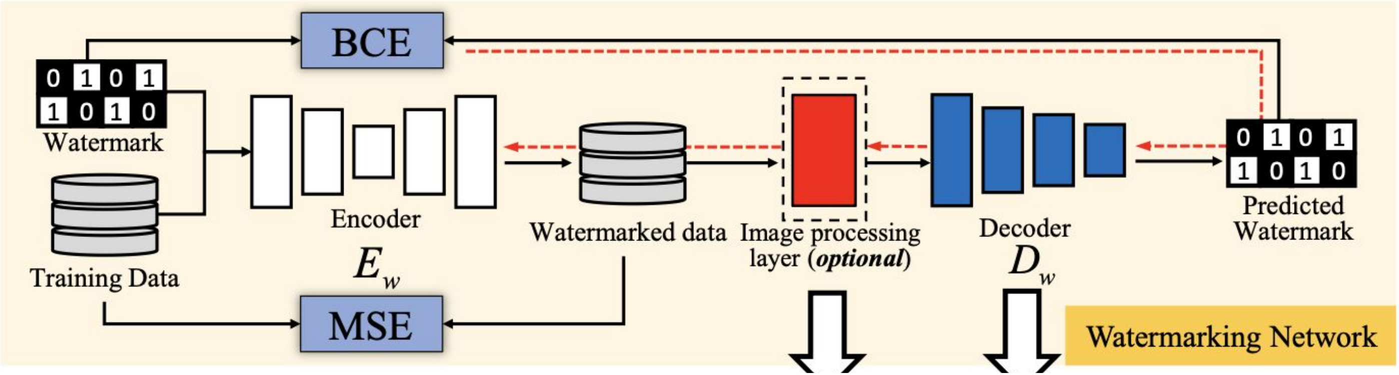
3.a Supervised GAN watermarking



3.a Supervised GAN watermarking

- The goal of watermarking network training stage is to learn a proper encoder and decoder, by minimizing the following loss:

$$L_w = \sum_{i=1}^{n} \mathrm{MSE}(x_i, E_w(x_i, w)) + \lambda \mathrm{BCE}(w, D_w(x_{w,i}))$$

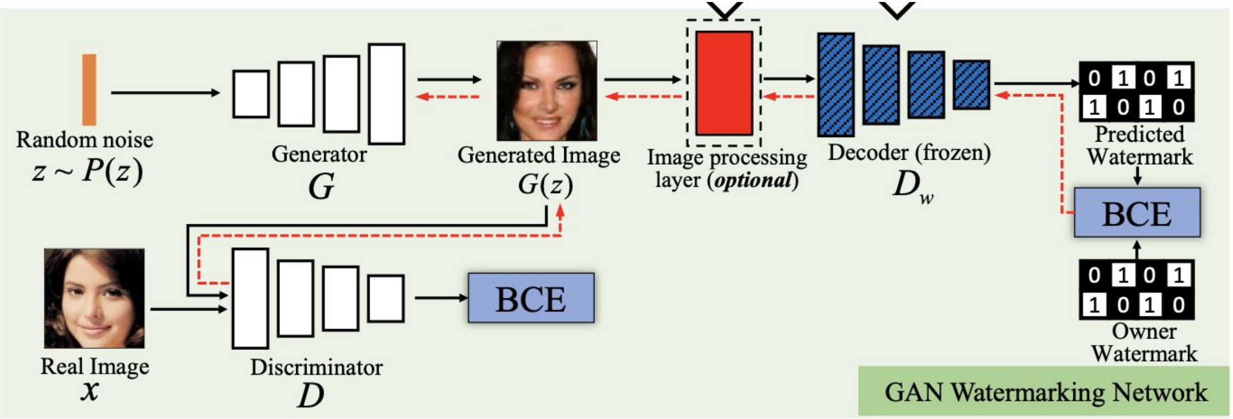


3.a Supervised GAN watermarking

- The discriminator loss:
- The generator loss

$$L_D(D,G) = -\mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] - \mathbb{E}_{z\sim P(z)}[\log(1-D(G(z)))],$$

$$L_G^w(D,G) = L_G(D,G) + \gamma \mathbb{E}_{z\sim P(z)}[BCE(D_w(G(z)), w_{gt})]$$



4. Challenges and Limitations

- Traditional Methods
  - Static embedding strategies
  - Limited adaptability
  - Vulnerability to deep-learning attacks
  - Suitable for medical use
- Machine Learning Approaches
  - High computational overhead
  - Data dependency
  - Potential adversarial vulnerabilities
  - Suitable for LLM, Generative models

5. Conclusions

- Emerging Research Areas
  - Hybrid traditional-ML methods
  - Transformer-based watermarking
  - Lightweight embedded system solutions
- Key Takeaways
  - Digital watermarking evolving from static to adaptive techniques
  - Machine learning offering dynamic, robust solutions
  - Continuous balance between imperceptibility and robustness
- Research Challenges
  - Developing universal watermarking frameworks
  - Ensuring multi-domain generalizability

References

[1] Mohammed, Abdulhakeem O., et al. "A blind and robust color image watermarking scheme based on DCT and DWT domains." Multimedia Tools and Applications 82.21 (2023): 32855-32881.

[2] Craver, Scott, et al. "Resolving rightful ownerships with invisible watermarking techniques: Limitations, attacks, and implications." IEEE Journal on Selected areas in Communications 16.4 (1998): 573-586.

[3] Yeung, Minerva M., and Fred Mintzer. "An invisible watermarking technique for image verification." Proceedings of international conference on image processing. Vol. 2. IEEE, 1997.

[4] Parah, Shabir A., et al. "Robust and blind watermarking technique in DCT domain using inter-block coefficient differencing." Digital Signal Processing 53 (2016): 11-24.

[5] Fei, Jianwei, et al. "Supervised gan watermarking for intellectual property protection." *2022 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2022.

Additional references:
[6] Liang, Yuqing, et al. "Watermarking Techniques for Large Language Models: A Survey." arXiv preprint arXiv:2409.00089 (2024).
[7] Fares, Kahlessenane, Khaldi Amine, and Euschi Salah. "A robust blind color image watermarking based on Fourier transform domain." Optik 208 (2020): 164562.

Thank you!