

Research on Innovation and Application of Digital Image Watermarking Technology Driven by Deep Learning

Jiawei Ding^{1, a}

¹Department of statistics and Data science, Beijing Normal University-Hong Kong Baptist University United International Colledge, Zhuhai, China

^a3136140736@qq.com

Abstract: Digital image watermarking technology realizes copyright traceability and content authentication by embedding copyright identification and authentication data in the host image, but traditional methods rely on manual design features, and it is difficult to balance robustness, invisibility and embedding capacity. In this paper, the deep learning watermarking technology in recent years is classified and summarized, and the advantages and disadvantages of representative methods are compared and analyzed from three dimensions: network architecture, training strategy and application scenario. Aiming at the problem that it is difficult to balance the robustness and invisibility of the existing methods, this paper proposes an improved end-to-end watermarking network of U-Net encoder which integrates the SE (Squeeze-and-Exclusion) attention mechanism. By embedding the SE channel attention module in the U-Net hop connection path, the watermark is guided to be embedded in the texture-rich area, and the reconstruction loss, watermark extraction loss and confrontation loss are jointly optimized. The experiment on the COCO dataset shows that the proposed method achieves a peak signal-to-noise ratio (PSNR) of 41.56 dB and a structural similarity (SSIM) of 0.9812, which is 2.61 dB higher than HiDDeN. Under attacks such as JPEG compression (QF=50), the bit error rate (BER) is reduced by about 24% compared to the baseline, achieving a better balance between invisibility and robustness. The ablation experiment further verified the effectiveness of SE module and anti-loss, and provided a new technical path for digital copyright protection.

Keywords: Digital image watermark; Deep learning; U-Net; SE attention.

1. Introduction

Digital image watermarking technology embeds copyright identification, author information or authentication data into the host image, which makes the watermark information imperceptible under normal use conditions, but can be reliably extracted when infringement occurs, thus realizing copyright traceability and content authentication. Traditional watermarking methods mainly include spatial domain method and transform domain method. These methods have certain effects in specific scenes, but they generally rely on manual design features and fixed embedding rules, and it is difficult to achieve an ideal balance among robustness, invisibility and embedding capacity [1]. Especially in the face of common attacks such as JPEG compression, geometric transformation and filtering noise, the extraction rate of traditional methods often drops sharply, which is difficult to meet the practical application requirements.

In recent years, breakthroughs in deep learning technology have brought a new research paradigm for digital image watermarking. Deep models represented by Convolutional Neural Networks (CNN), Generative Adversarial Networks (GAN), and Autoencoders have powerful automatic feature extraction and nonlinear mapping capabilities [2-3], enabling end-to-end learning of watermark embedding and extraction strategies. In addition, attention mechanism, diffusion model and Transformer architecture are gradually introduced into the field of watermarking, which promotes the continuous evolution of this direction [4-5].

Based on the above analysis, this paper summarizes the deep learning watermarking technology in recent years,

combs the research context from three dimensions: network architecture, training strategy and application scenario, and compares and analyzes the advantages and disadvantages of representative methods; This paper also proposes an end-to-end watermarking network based on improved U-Net encoder and SE (Squeeze-and-Exclusion) attention module. By jointly optimizing reconstruction loss, watermark extraction loss and confrontation loss, a better balance between invisibility and robustness is achieved.

2. Digital Image Watermarking Technology Based on Deep Learning

In recent years, the successful application of deep learning technology in the field of image processing has spawned a number of watermarking methods based on neural networks. According to different network architectures and training strategies, they can be classified into the following categories.

(1) End-to-end watermarking framework based on Autoencoder

Autoencoder structure is the earliest widely used paradigm in deep learning watermarking. Its core idea is to train an encoder-decoder pair. The encoder jointly encodes the watermark information and the host image into a watermarked image, and the decoder recovers the watermark from the watermarked image. HiDDeN proposed by Zhu J and others in 2018 is a pioneering work in this direction [6]. They designed an end-to-end framework based on CNN, and jointly optimized the encoder and decoder by simulating various attacks in the training stage, so that the watermark can be accurately extracted under various attacks. The experimental results show that HiDDeN outperforms

traditional methods in terms of PSNR and bit error rate (BER). Rupavathi et al. introduced attention mechanisms into the Autoencoder structure, guiding the network to focus on regions suitable for embedding in the image, further enhancing invisibility [7].

(2) GAN based watermarking method

GAN can generate more realistic watermarked images through adversarial training of the generator and discriminator. The StegaStamp proposed by Tancik M et al. in 2020 is a representative work in this direction [8]. They use the adversarial training mechanism of GAN to enable the generator to learn to generate visually more natural images while embedding watermarks, while the discriminator attempts to distinguish between watermarked images and real images, both of which work together to improve watermark quality in the game.

(3) Method based on attention mechanism and Transformer

Attention mechanism can help the network focus on the key areas in the image and avoid the visual distortion caused by embedding watermarks in sensitive areas. Manish R et al. embedded the channel attention and spatial attention into the watermark network [9], and the experiment showed that the PSNR of the watermarked image was improved by about 1-2 dB. In recent years, Transformer architecture has also been introduced into the field of watermarking due to its global

modeling capability. For example, Zhiwei Z et al. proposed a watermarking scheme based on Vision Transformer (ViT) [10], which utilizes self attention mechanism to capture long-range dependencies of images and exhibits good generalization performance on large-scale datasets.

(4) Emerging methods based on diffusion models

The diffusion model, as a cutting-edge technology in the field of image generation in recent years, has also begun to be applied to watermark research. Wang Y et al. explored invisible watermarking techniques based on semantic aware text image diffusion models [11], which restored watermarks through denoising processes and demonstrated potential on high-resolution images. However, this direction is still in the preliminary exploration stage, with limited relevant literature and high computational costs, and has not yet formed a mature framework.

3. Comparison and Deficiency of Existing Methods

Table 1 compares and summarizes the above representative methods from five dimensions: network architecture, robustness, invisibility, reversibility and computational complexity.

Table 1. Comparison of representative deep learning watermarking methods

Method	Network architecture	Robustness	Invisibility	Reversibility	Computational complexity
HiDDeN (2018)	CNN Autoencoder	Medium high	Middle	No	Middle
StegaStamp (2020)	GAN	High	High	No	High
Attention enhancement methods	CNN+SE/CBAM	High	High	No	Middle
Transformer method	ViT	Medium high	Middle	No	High
Diffusion model method	U-Net+DDPM	Preliminary verification	Preliminary verification	To be researched	Very high
Traditional DCT/DWT	None/Handmade design	Middle	Middle	No	Low

Through comprehensive analysis, it can be seen that the existing deep learning watermarking methods are significantly superior to the traditional methods in performance, but there are still the following shortcomings: First, the balance between robustness and invisibility is still not optimal. Many methods inevitably introduce visual distortion when improving robustness, especially under high-intensity attacks, it is difficult to have both invisibility and robustness. Second, the generalization ability of unknown attack types is limited. Most methods are optimized only on the preset attack set during training, and the performance may drop sharply in the face of new attacks not included in the training set. Third, the research on reversibility is insufficient. Most methods are irreversible watermarking, that is, the original image can not be completely recovered after embedding the watermark, which constitutes an application obstacle in medical images, legal evidence and other scenes that require high integrity of the original data. Fourth, the exploration of application scenarios is relatively simple. Most of the existing research focuses on general natural images, but the adaptability research in specific fields such as medical images, remote sensing images and social media content is still insufficient, lacking a unified analysis framework for multiple scenes. Based on the above analysis, this paper puts forward an improved watermarking network combining

attention mechanism and confrontation training in section 4, aiming at making up for some shortcomings of existing methods, and verifies it through experiments in section 5.

4. Improving U-Net Encoder and SE Attention Module

This study adopts an end-to-end architecture of encoder noise layer decoder, and optimizes the embedding and extraction process of watermarks through joint training [12]. The overall process is shown in Figure 1. The carrier image I_c and watermark information w are fused by an encoder to generate a watermarked image I_w . I_w simulates channel attacks such as JPEG compression, Gaussian noise, and geometric transformations through a noise layer; The decoder blindly extracts the watermark $\hat{w}$ from the attacked image. On this basis, this study improves the U-Net backbone network of the encoder by embedding SE attention modules in the skip connection path and adopting a multi loss joint optimization strategy in training.

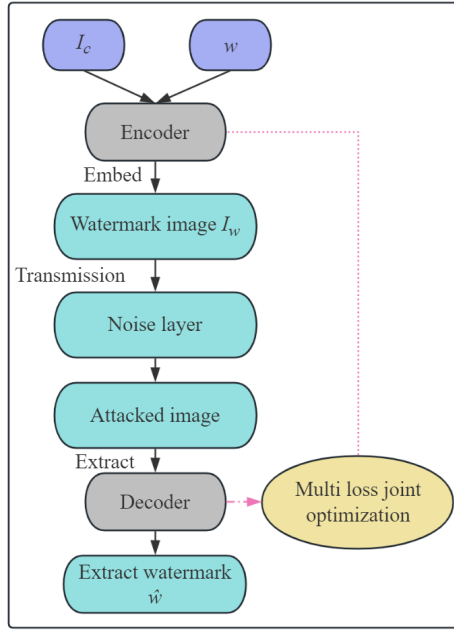


Figure 1. Encoder noise layer decoder end-to-end architecture overall process

The encoder takes U-Net as the backbone network, and the input is the mosaic result of the carrier image I_c and the watermark tensor $W \in \mathbb{L}$ in the channel dimension. The encoder consists of a four-level down-sampling encoding block and a four-level up-sampling decoding block, and the skip connection transfers the shallow spatial features of the encoding stage to the corresponding decoding layer to alleviate the information loss caused by repeated down-sampling [13]. The final output of the encoder is the residual image Δ , and the watermarked image is generated according to $I_w = I_c + \alpha\Delta$, where $\alpha \in (0,1)$ is the embedding strength factor, controlling the disturbance amplitude within the imperceptible range at the pixel level.

SE channel attention module is embedded in each hop connection path of U-Net encoder. The core function of this module is to adaptively weight the multi-channel features extracted by the encoder, so that the network can automatically focus on the areas with rich texture and suitable for embedding watermark, thus reducing the visual distortion caused by embedding in flat areas.

The calculation of SE module is divided into two steps:

Squeeze: Perform global average pooling on the input feature map $U \in \mathbb{R}^{H \times W \times C}$ along the spatial dimension, compress the two-dimensional features of each channel into a scalar, and obtain the channel description vector $z \in \mathbb{R}^C$:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W u_c(i,j), \quad c=1,2,\dots,C \quad (1)$$

Among them, $u_c(i,j)$ represents the activation value of the c -th channel at spatial position (i,j) , H and W are the height and width of the feature map, and C is the total number of channels. This operation aggregates the spatial information of each channel into a global scalar, enabling subsequent operations to make decisions based on the global statistical relationships between channels.

Excitation: By using a bottleneck structure (two fully connected layers) to learn the nonlinear dependencies between channels, a channel weight vector $s \in \mathbb{R}^C$ is generated:

$$s = \sigma(W_2 \delta(W_1 z)) \quad (2)$$

Among them, $W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ is the dimensionality reduction matrix, which compresses the number of channels from C to C/r is the dimensionality reduction ratio; $W_2 \in \mathbb{R}^{C \times \frac{C}{r}}$ is an upgraded matrix, restored to the original number of channels; $\delta(\cdot)$ is the ReLU activation function; $\sigma(\cdot)$ is the Sigmoid function, which normalizes the weights to the $(0,1)$ interval. Finally, the weighted feature map is obtained by $\tilde{u}_c = s_c u_c$, where each channel feature is assigned different importance weights.

In the watermark embedding scenario, the SE module assigns higher weights to texture rich channels, guiding the encoder to concentrate watermark perturbations on the spatial regions corresponding to these channels, thereby reducing overall visual distortion while ensuring embedding capacity.

The training objective consists of three weighted loss factors, which respectively constrain the quality of watermarked images, the accuracy of watermark extraction, and visual imperceptibility:

$$L_{total} = \lambda_{rec} L_{rec} + \lambda_{ext} L_{ext} + \lambda_{adv} L_{adv} \quad (3)$$

The reconstruction loss L_{rec} constrains the pixel level difference between I_w and I_c , measured by mean squared error (MSE), to keep the embedding perturbation minimal in the pixel domain. Extract the loss L_{ext} to constrain the consistency between the decoder output $\hat{w}$ and the original watermark w . Due to the fact that the watermark is a binary sequence, binary cross entropy loss is used. Adversarial loss L_{adv} introduces discriminator D to distinguish I_w from I_c , and improves the visual realism of watermarked images through adversarial training, making the perturbations generated by the encoder more difficult to perceive at the perceptual level:

$$L_{adv} = E_{I_c} [\log D(I_c)] + E_{I_w} [\log(1 - D(I_w))] \quad (4)$$

During training, the encoder and decoder minimize L_{total} , while discriminator D maximizes L_{adv} , and the two are optimized alternately. The weight coefficients λ_{rec} , λ_{ext} , and λ_{adv} are used to balance the contributions of the three losses: λ_{ext} is usually taken as the larger value to ensure extraction accuracy, and λ_{adv} is taken as the smaller value to avoid excessive interference with pixel fidelity during adversarial training. The typical configuration is $\lambda_{ext}=10$, $\lambda_{rec}=1$, and $\lambda_{adv}=0.0001$.

Set a differentiable noise layer between the encoder and decoder, and apply a combination of random sampling attacks on I_w during training, including JPEG compression, Gaussian noise, Gaussian blur, and random cropping. Each round of training randomly samples one type from the above attacks and applies it to I_w , allowing the decoder to learn robust watermark extraction capabilities under diverse distortion conditions. For non differentiable operations such as JPEG compression, DCT domain quantization approximation is used to ensure that gradients can propagate end-to-end.

5. Experimental Analysis

The experiment was carried out on the COCO 2017 data set. 10,000 images were randomly selected as the training set

and 1,000 images as the test set, and scaled to a resolution of 256×256 . The watermark information is set as a 30-bit binary sequence. Adam optimizer is used in network training, the learning rate is set to 1×10^{-3} , and the batch size is 32, and a total of 300 epoch are trained. The noise layer randomly samples JPEG compression, Gaussian noise, Gaussian blur and random clipping with equal probability. The evaluation indicators use peak signal-to-noise ratio (PSNR) and

structural similarity (SSIM) to measure invisibility, and BER to measure robustness. The higher the PSNR, the smaller the pixel level difference between the watermarked image and the carrier image. The closer the SSIM is to 1, the higher the structural fidelity. The lower the BER, the more accurate the watermark extraction. Table 2 lists the invisibility indicators under different embedding strengths α , as well as the comparison between our research method and baseline methods such as HiDDeN and Stegastamp.

Table 2. Invisible comparison

Method	PSNR(dB)	SSIM
HiDDeN	38.21	0.9521
Stegastamp	39.07	0.9684
Our ($\alpha=0.3$)	41.56	0.9812
Our ($\alpha=0.5$)	40.82	0.9867

When $\alpha=0.5$, this research method achieved PSNR of 40.82 dB and SSIM of 0.9867, which were 2.61 dB and 0.0346 higher than HiDDeN, respectively. When α drops to 0.3, PSNR increases to 41.56 dB, but BER slightly increases, indicating a trade-off between embedding strength and extraction accuracy. It can be observed from the residual

diagram in Figure 2 that the disturbance generated by this research method is mainly concentrated in the texture-rich area, and the residual in the flat area is close to zero, while the fuzzy artifacts along the edge of the object can be observed in the HiDDeN residual.

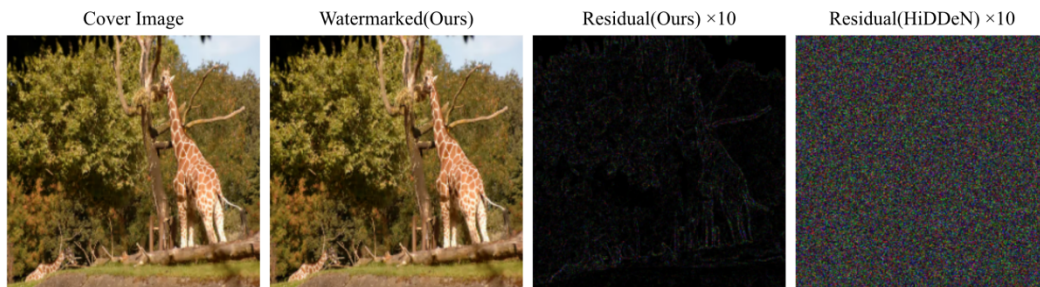


Figure 2. Watermarked image and residual visualization

Table 3 shows the BER comparison of two methods under different attack conditions.

Table 3. Comparison of BER under different attacks

Attack Type	Parameter	HiDDeN	Our
Unassailed	—	0.1696	0.1538
JPEG compression	QF=70	0.1852	0.1421
JPEG compression	QF=50	0.2347	0.1786
Gaussian noise	$\sigma=0.03$	0.2013	0.1624
Gaussian noise	$\sigma=0.05$	0.2581	0.2109
Gaussian blur	5×5	0.1764	0.1352
Random crop	Retain 70%	0.1938	0.1476

Under no attack conditions, this research method reduced the BER from 0.1696 to 0.1538, a relative decrease of 9.3%. Under the strong compression condition of JPEG compression QF=50, the BER is 0.1786, which is about 24% lower than HiDDeN. When Gaussian noise $\sigma=0.05$, the BER is 0.2109, which is still higher than the ideal value but has decreased by 18% compared to the baseline. The SE attention module has weaker resistance to Gaussian noise than Gaussian blur, which is related to its ability to bias watermark embedding towards low-frequency regions. Low frequency regions are more robust to Gaussian blur, but slightly more sensitive to additive Gaussian noise.

Figure 3 shows the BER variation curves under different JPEG quality factors. This research method has a stability advantage over HiDDeN in the medium compression strength range (QF=60-80), while both BER significantly increase

under extreme compression with QF<40.

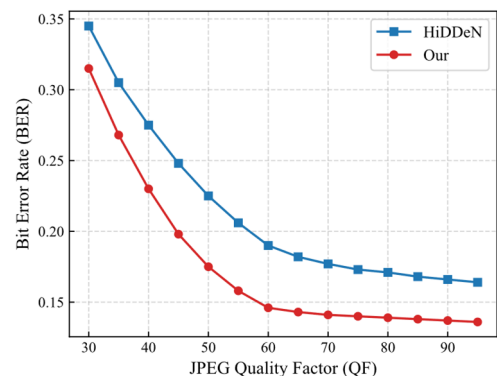


Figure 3. BER variation curve under different JPEG quality factors

To verify the effectiveness of the SE attention module, three variant models were trained under the same training configuration: Baseline (original U-Net encoder),+SE

(embedding SE module in skip connections), and+SE+Adv (adding adversarial loss to the former). The results are shown in Table 4.

Table 4. Results of ablation experiment

Model configuration	PSNR (dB)	SSIM	BER (JPEG QF=70)
Baseline	38.94	0.9602	0.2013
+SE	40.21	0.9756	0.1684
+SE+Adv	41.56	0.9812	0.1621

The introduction of SE module makes PSNR increase by 1.27 dB and BER decrease by 0.033, which verifies the effectiveness of channel attention in optimizing embedding position. After further adding confrontation loss, PSNR is increased by 1.35 dB, SSIM is increased by 0.0056, and BER is improved slightly, indicating that confrontation loss mainly affects the improvement of visual quality, and its contribution to extraction accuracy decreases marginally.

6. Conclusion

This research method has achieved a good balance between invisibility and robustness. The SE module guides the watermark to be embedded into the texture redundancy area through channel re-weighting, which improves the visual quality without increasing the embedding strength. The introduction of anti-loss further suppresses the high-frequency artifacts. However, the ablation experiment also reveals a limitation. The SE module makes the watermark embedded in the low-frequency region, which leads to a weaker resistance to Gaussian noise than to compression and fuzzy attacks. The following work can consider introducing channel attention at the decoding end to enhance the ability of extracting degraded signals, or adopting frequency domain mixed attention mechanism to broaden the embedded frequency band.

References

- [1] Wang, Q., Wen, T., Yang, H., et al. (2019). Synchronisation correction-based colour image watermarking using proximal classifier with consistency and multivariate generalised Gaussian model. *Electronics Letters*, 55(24), 1305–1307. <https://doi.org/10.1049/el.2019.2586>.
- [2] Rai, M., Goyal, S., & Pawar, M. (2023). An enhanced digital image watermarking technique using DWT-HD-SVD and deep convolutional neural network. *International Journal of Critical Computer-Based Systems*, 10(4), 269–286. <https://doi.org/10.1504/IJCCBS.2023.136317>
- [3] Shareef, A., Ghodhbani, H., Hamdani, M. T., et al. (2025). Deepfake detection via image watermarking: A generative adversarial model approach with limited data. *Multimedia Tools and Applications*, 84(41), 1–26. <https://doi.org/10.1007/s11042-025-21087-4>
- [4] Koley, S. (2020). Visual attention model based dual watermarking for simultaneous image copyright protection and authentication. *Multimedia Tools and Applications*, 80(5), 1–29. <https://doi.org/10.1007/s11042-020-09918-y>
- [5] Alireza, T., Zahra, H., & Hedieh, S. (2023). Convolutional neural network-based image watermarking using discrete wavelet transform. *International Journal of Information Technology*, 15(4), 2021–2029. <https://doi.org/10.1007/s41870-023-01232-8>
- [6] Zhu, J., Kaplan, R., Johnson, J., et al. (2018). HiDDeN: Hiding data with deep networks. In *Lecture Notes in Computer Science* (pp. 327–342). Springer. https://doi.org/10.1007/978-3-030-01267-0_40
- [7] Rupavathi, N., & Kavitha, M. (2025). Binarized simplicial convolutional neural networks for secure and robust digital image watermark embedding using elite opposite sparrow search optimization algorithm. *Circuits, Systems, and Signal Processing*, 45(7), 1–30. <https://doi.org/10.1007/s00034-025-03340-1>
- [8] Tancik, M., Mildenhall, B., & Ng, R. (2020). StegaStamp: Invisible hyperlinks in physical photographs. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2114–2123). IEEE. <https://doi.org/10.1109/CVPR42600.2020.00219>
- [9] Manish, R., Sachin, G., & Mahesh, P. (2023). An optimized deep fusion convolutional neural network-based digital color image watermarking scheme for copyright protection. *Circuits, Systems, and Signal Processing*, 42(7), 4019–4050. <https://doi.org/10.1007/s00034-023-02299-1>
- [10] Zhiwei, Z., Han, W., & Hui, F. (2023). A convolutional neural network-based blind robust image watermarking approach exploiting the frequency domain. *The Visual Computer*, 39(8), 3533–3544. <https://doi.org/10.1007/s00371-023-02967-y>
- [11] Wang, Y., Xu, H., Wang, Z., et al. (2026). PT-Mark: Invisible watermarking for text-to-image diffusion models via semantic-aware pivotal tuning. *IEEE Transactions on Dependable and Secure Computing*, 23(4), 8909–8922. <https://doi.org/10.1109/TDSC.2026.3689195>
- [12] M. M, D., A. A, F., & M. T, G. E. (2023). Convolutional neural network and 2D logistic-adjusted-Chebyshev-based zero-watermarking of color images. *Multimedia Tools and Applications*, 83(10), 29969–29995. <https://doi.org/10.1007/s11042-023-16649-3>
- [13] R, A., & C, M. (2025). Enhancing the protection of medical images using watermarking technology processed with deep learning algorithms. *Multimedia Tools and Applications*, 84(6), 1–30. <https://doi.org/10.1007/s11042-024-20558-4>