IWRN:A Robust Blind Watermarking Method for Artwork Image Copyright Protection Against Noise Attack

Feifei Kou[1,2], Yuhan Yao[1,2], Siyuan Yao[1], Jiahao Wang[1,2], Lei Shi[4, 5*], Yawen Li[3], Xuejing Kang[1]

[1]School of Computer Science (National Pilot School of Software Engineering), BUPT, Beijing 100876, China
[2]Key Laboratory of Trustworthy Distributed Computing and Service, BUPT, Ministry of Education, Beijing 100876, China
[3]School of Economics and Management, BUPT, Beijing 100876, China
[4]State Key Laboratory of Media Convergence and Communication, CUC, Beijing 100024, China
[5]State Key Laboratory of Intelligent Game, Yangtze River Delta Research Institute of NPU, Taicang 215400, China
koufeifei000@bupt.edu.cn; yaoyuhan@bupt.edu.cn; yaosiyuan@bupt.edu.cn; wjh2680@bupt.edu.cn; leiky_shi@cuc.edu.cn; warmly0716@126.com; kangxuejing@bupt.edu.cn

Abstract

Adding imperceptible watermarks to artwork images, such as paintings and photographs, can effectively safeguard the copyright of these images without compromising their usability. However, existing blind watermarking techniques encounter two major challenges in addressing this task: imperceptibility and robustness, particularly when subjected to various noise attacks. In this paper, we propose a blind watermarking method for artwork image copyright protection, **IWRN**, which can ensure both the **I**mperceptibility of the **W**atermark and **R**obustness against **N**oise attacks. For imperceptibility, we design a Learnable Wavelet Network (LWN) to adaptively embed the watermark into the high-frequency region where the watermark has better invisibility. For robustness, we establish a Deform-Attention based Invertible Neural Network (DA-INN) with a decoding optimization, which offers the advantage of computational reversion, and combines the deform-attention mechanism and decoding optimization to enhance the model's resistance against noises. Additionally, we design a Joint Contrast Learning (JCL) mechanism to improve imperceptibility and robustness simultaneously. Experiments show that our IWRN outperforms other state-of-the-art blind watermarking methods, achieves an average performance of 46.74 PSNR and 99.91% accuracy across three datasets when facing 12 kinds of noise attacks.

Code — https://github.com/BUPT-SN/IWRN

Introduction

Artwork images play a crucial role in the dissemination and display of art, but they are also faced with risks of copyright infringement. Blind watermarking methods are widely used in protecting the copyright of artwork images (Wan et al. 2022; Xiao, Zhao, and Li 2022). However, there are two main challenges in this task: (1) Imperceptibility problem, that is, how to make the embedded watermark as hidden as possible so as not to affect the viewer's visual experience of the artwork; (2) The robustness problem, which refers to the ability to accurately extract watermarks in the face of various noise attacks such as Gaussian noise and affine transformation.

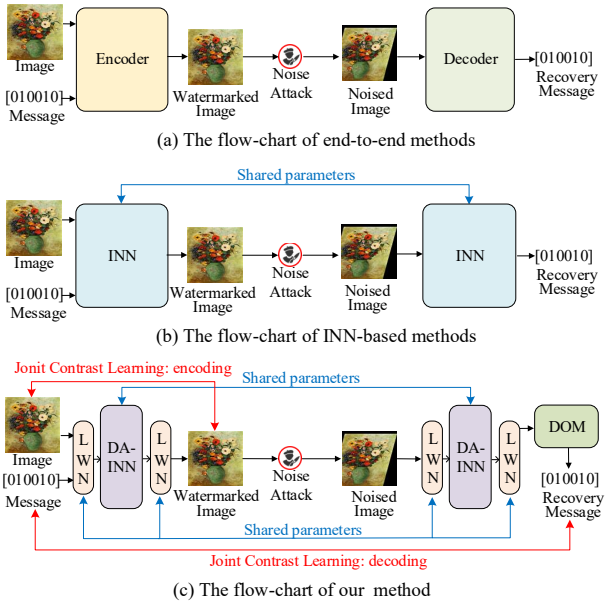

Figure 1: The flow-charts of our method and other methods.

In order to deal with the imperceptibility problem, traditional blind watermarking methods (Hsu and Wu 1999; Barni, Bartolini, and Piva 2001) commonly embed the watermarks in the frequency domain of the image. However, when such methods encounter noise attacks, the watermarks will be corrupted and cannot be recovered normally (Huang et al. 2023; Arab, Ghorbanpour, and Hefeeda 2024). In recent years, deep learning has greatly promoted the development of blind watermarking. Typical methods are end-to-end methods (Zhu et al. 2018; Liu et al. 2019; Jia, Fang, and Zhang 2021; Yu 2020; Zhang et al. 2020a; Tancik, Mildenhall, and Ng 2020; Guo et al. 2023; Wang et al. 2024b,a; Khan, Wong, and Baskaran 2024), their flow-chart is shown in Figure 1(a). These methods include encoding and decoding networks, which are trained in an end-to-end manner. Optimizing the model structure or adding suitable loss

*Corresponding Author.

functions will yield better results. However, it is worth noting that the imperceptibility and robustness of such methods may not always achieve satisfactory results across all scenarios. The architectures between encoding and decoding of these methods are fundamentally different, making it difficult to achieve a balance between imperceptibility and robustness. Especially when the encoding ability is excessively strong, ensuring the robustness of their decoding becomes difficult, and robustness is a top priority in watermarking methods. For this reason, in order to ensure robustness, imperceptibility has to be sacrificed.

For the robustness problem, the INN-based methods (Jing et al. 2021; Ma et al. 2022a; Xu et al. 2022; Guan et al. 2022; Fang et al. 2023; Luo et al. 2023) are effective, and their flow chart is shown in Figure 1(b). They use a shared reversible network for watermark embedding and extraction, and decoding can be seemed as the inverse process of encoding, so they can robustly extract watermarks in an exact form. However, they also possesses inherent limitations. Firstly, INN-based digital watermarking methods struggle to maintain robustness against complex noise (such as Affine, Rotation, etc.). In addition, there is a lack of specialized optimization strategies to improve the models' encoder-decoder ability. Therefore, there is still a large space to enhance robustness and Imperceptibility, especially when dealing with complex noise.

In this paper, to address the aforementioned issues, we propose a blind watermarking method for artwork image copyright protection, **IWRN**, which can ensure both the imperceptibility of the watermark and its robustness against noise attacks. As shown in Figure 1(c), our IWRN mainly contains three core components: Learnable Wavelet Network (LWN) , Deform-Attention based Invertible Neural Network (DA-INN), and a Decoding Optimization Module (DOM). Firstly, to enhance the imperceptibility, we design the LWN to adaptively embed the watermark into the high-frequency region where the human eye is less sensitive to watermark. Secondly, we propose the DA-INN, which offers the advantage of computational reversion, and combines the deform-attention mechanism to enhance the model's resistance against noises. To further improve the robustness of DA-INN, we introduce the DOM to make up for its shortcomings in optimization. Additionally, we established a joint contrast learning mechanism (JCL), through which the similarity between the image and the watermarked image, as well as the similarity between the watermark and the decoded watermark, was contrastively learned, thus further improving the encoding and decoding ability of the model, and further enhancing the imperceptibility and robustness of the watermark.

The main contributions are summarized as follows:

- We propose a novel robust blind watermarking method, IWRN, with the proposed joint contrast learning, it can maintain high imperceptibility and achieve high robustness against all kinds of noise.
- The proposed LWN, can adaptively learn the parameters of the wavelet transform, so that the watermark can be placed in the high-frequency region where the human eye is less sensitive to watermark, so as to improve the imperceptibility.
- The proposed DA-INN and DOM joint decoding can greatly improve the robustness from three aspects: 1) It inherits reversibility advantages of INN; 2) With the deform-attention, it is designed to potentially offer better resistance against complex noise attacks, such as affine transformation; 3) By introducing the DOM module, the decoding ability of the model can be further improved.
- Extensive experiments demonstrates that our IWRN achieves better performance than the state-of-art blind watermarking methods.

Related Work

Digital watermarking

Digital watermarking is frequently employed in image protection, as well as other information protection applications (Begum and Uddin 2020), such as audio protection (Qu et al. 2023; Bassia, Pitas, and Nikolaidis 2001) and model protection (Yang, Wang, and Wang 2022; Zhang et al. 2020b; Darvish Rouhani, Chen, and Koushanfar 2019). Watermarking technology helps to deal with the problem of copyright protection of artwork images. Traditional watemarking methods, (Van Schyndel, Tirkel, and Osborne 1994; Hsu and Wu 1999; Kundur and Hatzinakos 1997; Das, Samaddar, and Keserwani 2018; Wang et al. 2023), commonly embed watermarks into frequency domain of images, thus improving the invisibility of the watermark. However, its robustness under noise attacks is poor. Recently, The deep learning-based methods have achieved better performance than traditional methods, as they can better learn the depth features of the watermark and the image. They can be divided into end-to-end methods and INN-based methods.

Typical end-to-end methods, such as HiDDeN (Zhu et al. 2018), achieve watermark embedding and recovering through end-to-end training. More suitable constraints or optimization measures to such methods often achieve more imperceptibility and robustness. Some works (Zhang et al. 2020a) utilize a adversarial network to optimize the model's ability in encoding and decoding. Additionally, several works have focused on watermarking in specific scenarios, such as decoding messages from in-the-wild videos (Tancik, Mildenhall, and Ng 2020), screen-shooting (Fang et al. 2022) or multi-source image detection (Wang et al. 2024a) and embedding watermarks into images of arbitrary resolutions (Guo et al. 2023). They adapt different optimization methods to improve the performance of their models. However, in this kind of method, the decoding output is obtained by approximate likelihood inference rather than exact calculation, which makes the robustness of watermark extraction vulnerable. Usually, the imperceptibility needs to be sacrificed in exchange for robustness. Therefore, the robustness and imperceptibility of this kind of method are often unsatisfactory.

Invertible Neural Network

The Invertible Neural Network (INN) serves as a crucial framework within the normalization flow models. It was

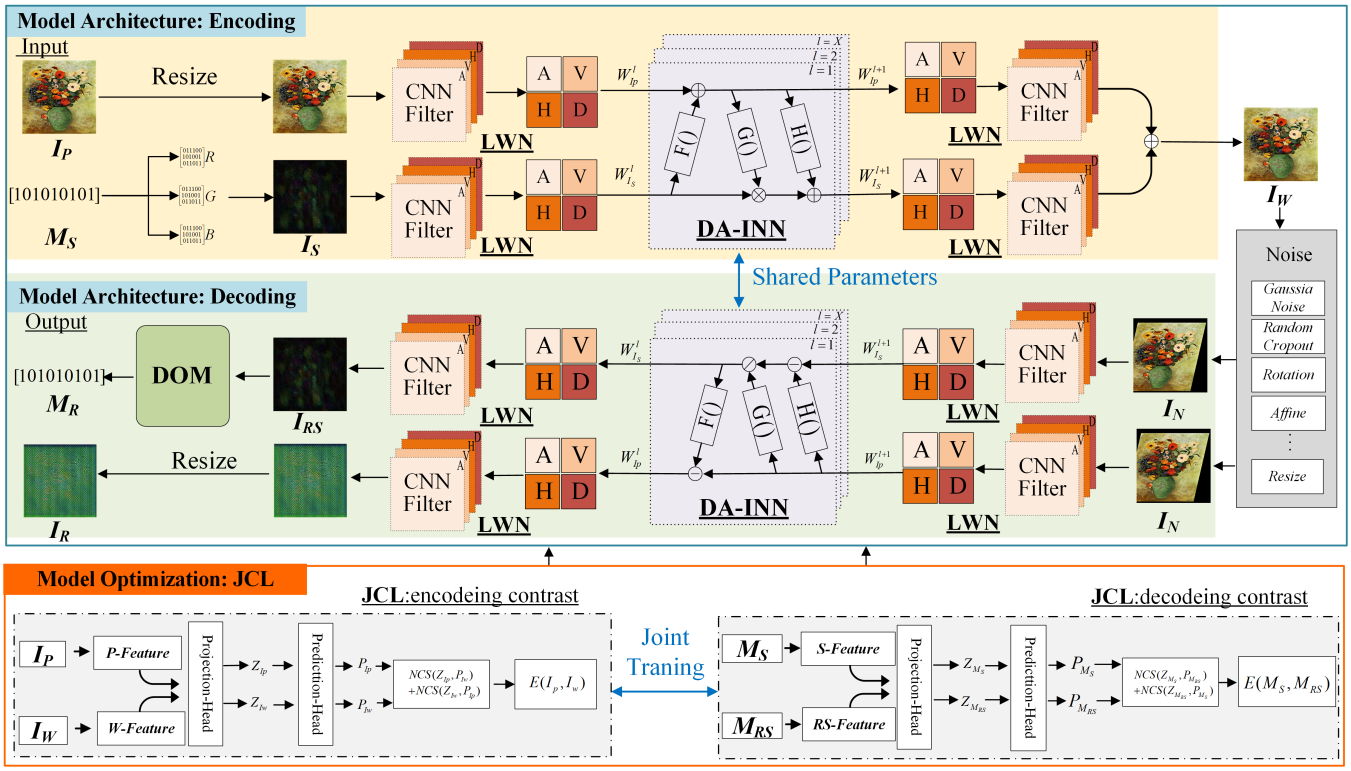

Figure 2: The architecture and the optimization of our IWRN.

originally introduced in NICE (Dinh, Krueger, and Bengio 2014). Due to its characteristic of preserving lossless information, INN has become a valuable choice for digital watermarking tasks. In most INN-based approaches, such as HiNet (Jing et al. 2021), CIN (Ma et al. 2022a) and IRWArt (Luo et al. 2023), both the encoding and decoding processes use a shared INN, enabling more precise watermark extraction compared to end-to-end methods. However, as such methods overly rely on their invertibility and lack additional optimization, their robustness will decrease significantly in the face of complex noise such as JPEG compression noise and affine transformation noise. CIN (Ma et al. 2022a) introduces an additional decoding module for JPEG compression noise, and other simple noises are parallel decoded by INN. Therefore, it has good robustness for compression noise, but for other noises such as affine noise, the robustness still needs to be improved.

Method

Overall Architecture

The architecture and the optimization of our IWRN is shown in Figure 2, and description of symbols are shown in Table 1. As shown in Figure 2, in the encoding and decoding phases, we propose LWN and DA-INN for generating watermarked images I_W and restoration of secret images I_{RS}, and introduce DOM in the decoding process to enhance decoding performance. In addition, we also design the JCL, and utilize it throughout the training process to ensure the imperceptibility of watermark and robustness of the model.

Symbol	Description
I_P	The image to be protected
M_S	The secret message
I_S	The secret image
I_W	The image watermarked by secret message
I_N	The image after noise interference
I_C	The image copied during the inverse process
I_R	The restored protected image
I_{RS}	The restored secret image
M_R	The recovered secret message

Table 1: Description of symbols used in the Method part.

Learnable Wavelet Network(LWN)

Watermark embedded in the high-frequency domain is less perceptible to human eye (Luo et al. 2023), hence, we should insert watermarks preferentially into suitable high-frequency regions. We propose LWN to capture high-frequency features that contribute to watermark imperceptibility.

Figure 3 shows the forward process of the LWN. Given an image $I_m \in (w, h, c)$, where w, h, c represents its width, height and number of channels respectively, $I_{out} \in (\frac{W}{2}, \frac{H}{2}, 4C)$ will be obtained after LWN, where 4C represents the four channels of image frequency domain features: **A**pproximation, **H**orizontal, **V**ertical, **D**iagonal. The core component of LWN is the DWT-CNN, which consists of a group of 2D-convolution operations designed to simulate the discrete wavelet transform. The difference between

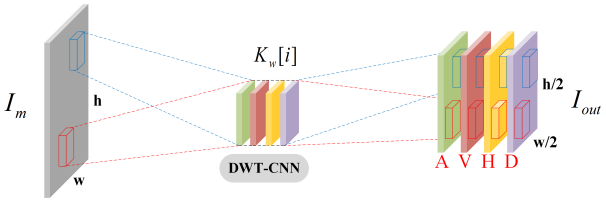

Figure 3: The forward process of LWN

LWN and the traditional discrete wavelet transform (DWT) is that the filter kernel of DWT is fixed. However, in LWN, the DWT is simulated by CNN, and its convolution kernel is constantly learning and updating as the training process of watermark embedding and extraction, so as to select more suitable high-frequency features to embed the watermark, thereby improving imperceptibility. When the input image I_m is input into the forward process of the DWT-CNN, it will be processed by four different DWT-CNN layers, resulting in the final I_{out} composed of four different frequency domains. The forward and inverse processes of this stage are as follows:

$$I_{out} = \mathbf{L}(I_m, w_A, w_V, w_H, w_D) \tag{1}$$

$$I_m = \mathbf{L}^{-1}(I_{out}, w_A, w_V, w_H, w_D) \tag{2}$$

where $\mathbf{L}(\cdot)$ is the symbol of the forward LWN; $\mathbf{L}^{-1}(\cdot)$ represent the inverse LWN; w_A, w_V, w_H, w_D represent the wavelet kernels of corresponding filters of A, V, H, D.

Deform-Attention INN(DA-INN)

INN-based digital watermarking algorithms (Ma et al. 2022b; Jing et al. 2021) have demonstrated that the INN has good performance in imperceptibility and robustness against simple noise. However, INN-based digital watermarking having insufficient robustness encountered with complex noise (such as Affine, Rotation). To address the issue, we propose a novel INN-based network named Deform-Attention INN(DA-INN) in digital watermarking.

As we can see in Figure 2, our DA-INN consists of forward and backward process, each process include three operations: $F(\cdot)$, $G(\cdot)$ and $H(\cdot)$. Since the operations in the forward direction are parameterized by invertible functions, the tensor can be accurately recovered during the backward process with the shared parameters. Taking the forward process of the l^{th} coupling layer as an example, we use $W^l_{I_P}$ and $W^l_{I_S}$ to represent the input watermark and image and the corresponding $W^{l+1}_{I_P}$ and $W^{l+1}_{I_S}$ to represent the output image and secret watermark after through the current layer, respectively. The specific operations of the forward propagation can be mathematically represented as follows:

$$\mathbf{W}^{l+1}_{I_P} = \mathbf{W}^l_{I_P} + F\left(\mathbf{W}^l_{I_S}\right) \tag{3}$$

$$\mathbf{W}^{l+1}_{I_S} = \mathbf{W}^l_{I_S} \odot \exp\left(G(\mathbf{W}^{l+1}_{I_P})\right) + H(\mathbf{W}^{l+1}_{I_P}) \tag{4}$$

The backward process:

$$\mathbf{W}^l_{I_S} = \exp\left(-G(\mathbf{W}^{l+1}_{I_P})\right) \odot \left(\mathbf{W}^{l+1}_{I_S} - H(\mathbf{W}^{l+1}_{I_P})\right) \tag{5}$$

$$\mathbf{W}^l_{I_P} = \mathbf{W}^{l+1}_{I_P} - F(\mathbf{W}^l_{I_S}) \tag{6}$$

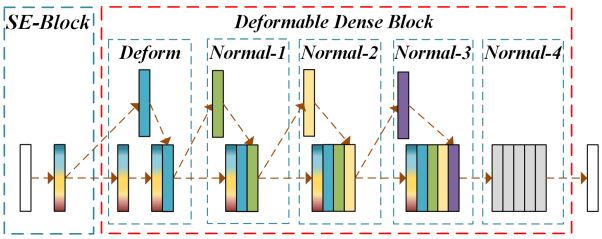

Figure 4: The Deform Attention dense block

Where $\odot$ represents the Hadamard product.

Each operation function $F(\cdot)$, $G(\cdot)$, $H(\cdot)$ in the network is implemented by a dense block structure that integrates deformable convolution and self-attention mechanisms, named as DA-dense block(Deformable Attention dense block). As shown in Figure 4, our DA-dense block is composed of two parts: self attention block(SE-Block) and Deformable Dense Block. Initially, the feature map is fed into the SE-Block, which serves to capture the importance of the channels. Then, the feature map is fed into Deformable Dense Block, which has five convolutions. The first layer is a deformable convolution which is densely connected with four regular convolutions, with the feature channels continually expanding. The final convolution layer reduces the feature map with a high number of channels back to the original channel number.

The attention module within the DA-dense block assigns varying attention weights to different channels, effectively handling feature map that have been differentiated into high and low frequencies by the LWN. Additionally, the adaptive sampling positions of the Deformable Dense Block enhance the robustness of the INN when dealing with spatial transformations.

Decoding Optimization Module(DOM)

As shown in Figure 1(a), the end-to-end watermarking methods use a decoder to recover the message, which gives a great robustness against various of attacks. But the INN-based methods relies solely on the intrinsic invertible characteristic of the INN itself, resulting in lower robustness when confronted with complex noises. In order to improve the decoding ability of INN-based methods. As we can see in Figure 1(c), after the inverse process of INN and LWN, we design a DOM: an attention-based decoder during decoding process to improve the robustness of our IWRN.

The operation of DOM can be expressed as follows:

$$M_R = \boldsymbol{\mathcal{F}}_{\text{FC}}\Big(\boldsymbol{\mathcal{A}}_{\text{SE}}\big(\boldsymbol{\mathcal{C}}_{\text{conv}}(I_{RS})\big)\Big) \tag{7}$$

where $\boldsymbol{\mathcal{F}}_{\text{FC}}(\cdot)$ is a full connection layer, $\boldsymbol{\mathcal{A}}_{\text{SE}}(\cdot)$ represents the attention mechanism and $\boldsymbol{\mathcal{C}}_{\text{conv}}(\cdot)$ is the convolution layer. It is worth noting that the DOM we established is applicable to all I_{RS}, that is to say, it is not only effective in resisting complex noise, but also improves robustness against simple noise attacks. Therefore, we merge the advantages of end-to-end methods in resisting noise attacks and the advantages of reversible network accurate calculation, which can greatly improve the robustness of the model by DOM.

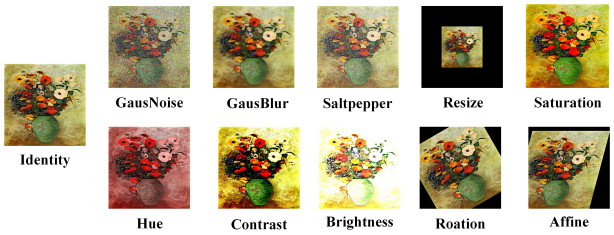

Figure 5: The comparisons of the image under N_{All}.

Joint Contrast Learning(JCL)

Contrastive learning, as a self-supervised, label-free training strategy, is very suitable for the training process of blind watermarking method. Traditional Digital watermarking technology, employs L2 loss or Mean Squared Error (MSE) loss for evaluation and training, which leads to the evaluation metrics being solely based on low-dimensional features. We propose JCL, the method employing a contrast learning network (Chen and He 2021) to evaluate both the encoding and decoding performance of our model in a high-dimensional feature space. As shown in Figure 2, JCL contains two parts: encoding contrast learning and decoding contrast learning. The structure of the contrast learning network can be expressed as follow:

$$Z_T = \rho(T_{feature}) \tag{8}$$

$$P_T = \gamma(Z_T) \tag{9}$$

where T is the tensor in our model, and $T_{feature}$ is the feature extracted from T. The functions $\rho(\cdot)$ and $\gamma(\cdot)$ are the projection head and prediction head, respectively. Z_T and P_T denote the projection and prediction of $T_{feature}$. By comparing the similarity of features between two different inputs after being projected and predicted, unsupervised contrastive learning can be performed.

The corresponding similarity is evaluated through a loss function. The evaluation function for JCL is as follows:

$$\mathbf{E}(I_P, I_W) = \frac{1}{2}\mathbf{D}(P_{I_P}, Z_{I_W}) + \frac{1}{2}\mathbf{D}(P_{I_W}, Z_{I_P}) \tag{10}$$

$$\mathbf{E}(M_S, M_R) = \frac{1}{2}\mathbf{D}(P_{M_S}, Z_{M_R}) + \frac{1}{2}\mathbf{D}(P_{M_R}, Z_{M_S}) \tag{11}$$

where $\mathbf{D}(\cdot)$ represents the function of Negative Cosine Similarity; Z_i and P_i are the projection and prediction of the i tensor.

Noise Pool

During the propagation of an artwork image, it is subjected to a variety of noise attacks. To improve the robustness of our IWRN, the construction of a Noise Pool is a crucial component. The purpose of the Noise Pool is to simulate various potential image processing operations and attacks, ensuring that the watermark algorithm remains effective under different attacks. Ultimately, we selected the identity(no noise attack) and 12 kinds of noises as follows:

$$N_{\text{All}} = \{\text{Identity}, \text{GaussianNoise}, \text{GaussianBlur}, \text{Saltpepper}, \text{Resize}, \text{Saturation}, \text{Hue}, \text{Contrast}, \text{Brightness}, \text{Rotation}, \text{Affine}, \text{Dropout}, \text{Cropout}\}$$

The image quality degradation caused by the noises is shown in Figure 5. Cropout and Dropout are not shown in Figure 5, which are common in watermarking works (Ma et al. 2022b), (Arab, Ghorbanpour, and Hefeeda 2024), (Huang et al. 2023). The reason is that these attacks involve replacing portions of the watermarked image with corresponding parts of the original image, which has a significant impact on watermark quality but is difficult to detect visually.

Loss Functions

The overall loss function comprises four components: Encode Loss (L_{en}), Restore Image Loss (L_{re}), Message Decode Loss (L_{de}), and Contrast Learning Loss (L_{con}). These losses aim to ensure that I_W is close to I_P, M_R matches M_S, and I_R is distinct from I_P to prevent unauthorized reconstruction of the original image. The expression of the total loss function is presented in the following manner:

$$L_{total} = \lambda_1 L_{en} + \lambda_2 L_{re} + \lambda_3 L_{de} + \lambda_4 L_{con} \tag{12}$$

The L_{en} loss function is composed of two components: the L_2 loss, which assesses the pixel-level discrepancies within the image, and $Lpips$ (Zhang et al. 2018) , which gauges the image loss from the perspective of human visual perception. The L_{re} uses the L_2 loss between I_P and I_R to evaluate the image reconstruction quality. The L_{de} function calculates the discrepancy between M_S and M_R using Mean Squared Error (MSE) . The L_{con} loss function is the use the evaluation function in JCL during both encoding and decoding process.

$$L_{con} = \frac{1}{2}\mathbf{E}(I_P, I_W) + \frac{1}{2}\mathbf{E}(M_S, M_R) \tag{13}$$

where $\mathbf{E}(I_P, I_W)$ and $\mathbf{E}(M_S, M_R)$ represent the evaluation function of JCL encoding and JCL decoding process.

Experiments

Basic setup

Dataset 10000 images from COCO 2017, 5000 image from wikiart and 500 image from div2k are utilized for training. The testing datasets include 10500 images, in which 5000 is come from COCO 2017, 5000 from Wikiart and 500 from Div2k.

Implementation Details We adopt the image size of $W \times H \times C = 128 \times 128 \times 3$ and the message length of 32 bits under N_{All} for base model training. The Adam optimizer is used with a learning-rate = 0.0001 and a batch size of 16 to train IWRN. For the loss, the corresponding weight factors $\lambda_1 : \lambda_2 : \lambda_3 : \lambda_4 = 2 : -1 : 2 : 1$. And a NVIDIA RTX A6000 GPU is employed for acceleration under PyTorch 2.1.0.

Metrics To assess the performance of our watermarking IWRN impartially, we take both imperceptibility and robustness into consider. We choose the Peak Signal-to-Noise Ratio (PSNR) for visual similarity and Structural Similarity(SSIM) for pixel similarity. Additionally, the Mean Opinion Score (MOS), ranging from 1 (lowest) to 5 (highest), was used to gauge volunteers' satisfaction with image quality. To

Model	PSNR(DB)↑	ACC(%)↑						
		Cropout $P = 30\%$	Dropout $P = 30\%$	Resize $P = 70\%$	SaltPepper $P = 10\%$	Rotation $D = 15$	Affine $S = 10$	Average -
HiDDeN (Zhu et al. 2018)	30.45	73.96	75.78	80.43	84.31	92.39	89.44	82.71
IGA (Zhang et al. 2020a)	32.81	79.33	77.51	81.44	85.21	87.31	88.73	83.26
StegaStamp (Tancik, Mildenhall, and Ng 2020)	34.88	90.41	97.24	94.31	87.38	96.32	98.22	93.98
PIMoG (Fang et al. 2022)	36.28	92.73	96.34	98.76	87.23	96.32	97.12	94.75
CIN (Ma et al. 2022b)	39.74	99.99	99.99	97.44	99.02	94.24	94.94	97.60
DWSF (Guo et al. 2023)	38.07	98.51	99.99	98.54	94.85	98.20	98.14	98.03
MuST (Wang et al. 2024a)	38.41	96.53	94.33	96.67	85.24	92.30	93.85	93.15
Ours Model	**41.55**	**99.99**	**99.99**	**99.99**	**99.92**	**98.78**	**98.75**	**99.57**

Table 2: Comparison of different models under various attacks with noise strength indicated in the header.

measure the robustness of IRWN, we evaluate the accuracy (Acc $= 1 -$ BER) between the embedded M_S and the recovered M_R, where the BER(Bit Error Ratio) represents the percentage of incorrect bits.

Baseline Several SOTA methods have been selected as baselines, including: CIN (Ma et al. 2022b), a very effective method to combine the INN with digital watermarking; HiDDeN (Zhu et al. 2018), the first trainable framework in end-to-end method for watermarking; IGA (Zhang et al. 2020a), using attention to highlight pixels that hide data; Stegastamp (Tancik, Mildenhall, and Ng 2020), using GAN to improve the performance when facing Affine attacks; DWSF (Guo et al. 2023), a watermarking resolution against multi-step attacks; PIMOG (Fang et al. 2022), a watermarking method for screen shooting attack and MuST (Wang et al. 2024a), aiding to improve performance in multi-source image compositing scenarios.

Comparative Experiment

To further validate the effectiveness of our model, we compared IWRN with baseline methods using an image size of $W \times H \times C = 128 \times 128 \times 3$, and a message length of 32 bits for training. Due to the difficulty of implementing all noise pools in N_{ALL} across all baseline methods, we selected the noises listed in Table 2 as the comparison noise pool and trained the baseline methods. We conduct basic comparison between IWRN and other baseline methods and further compare the visual impact of watermarking

Basic Comparison In the basic comparison, we compare the accuracy rate and the impact of the watermark on the image under various noise conditions with other baseline methods. As we can see in Table 2, INN-based method-CIN performed well in PSNR and robustness to simple noise attacks, such as Dropout and Saltpepper, but under spatial transformation (Rotation, Affine), CIN is worse than some end-to-end method (StegaStamp, PIMoG and DWSF). Our IWRN effectively combines the strengths of INN with those of end-to-end network approaches. This integration facilitates SOTA performance under conditions of common noise as well as spatial transformation noise. The comparisons above indicate that IWRN not only withstands a greater variety of noise but also outperforms the baselines under specific noise conditions.

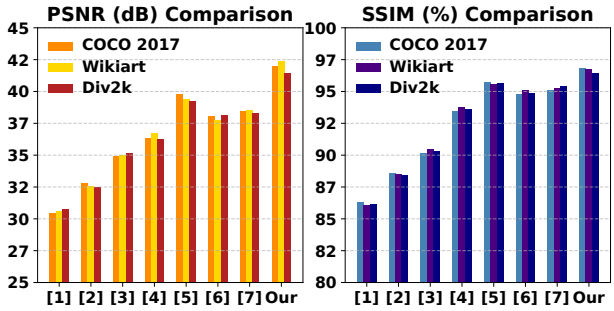

Figure 6: Visual qulity comparison. [1]HiDDeN, [2]IGA, [3]Stegastamp, [4]PIMoG, [5]CIN, [6]DWSF, [7]MuST

Methods	Hidden	IGA	SetgaStamp	PIMoG	DWFS	CIN	MuST	Ours
MOS	3.70	4.12	3.98	4.25	4.40	4.65	4.50	**4.75**

Table 3: MOS scores for different methods

Further Comparison in Visual Quality We further compare the impact of each method on visual quality by comparing the PSNR and SSIM after watermarking on three datasets and conducting MOS evaluations. As shown in Figure 6, watermarked images via IWRN have the best PSNR and SSIM to the original images, means IWRN can achieve the best imperceptibility. Compared to HiDDeN , IGA , Stegastamp, PIMoG, CIN, DWSF and MuST, our IRWN has 35.45%, 26.63%, 19.12%, 14.52%, 4.55%, 9.14%, 8.17% increasing in PSNR and 12.18%, 9.27%, 7.05%, 3.28%, 1.10%, 1.86%, 1.52% increasing in SSIM, respectively. Additionally, as shown in Table 3, MOS ratings from 30 volunteers rank our model the highest The experimental results show that IWRN impacts watermarking image less than other baselines and achieves the highest visual quality.

Diverse Experiments

Stability Experiment As the intensity factor of noise increases, the robustness of the model faces greater challenges due to the increased amount of distortion carried per unit area. We evaluate robustness under intensity factors of 1, 1.5, 2.0, and 2.5 noise parameters. As shown in Table 4, the results presented in the table indicate that the IWRN maintains high robustness even under a noise intensity more than twice.

Noise	Parameter	Intensity Factor(ACC↑)				
		1.0	1.5	2.0	2.5	Average
GausNoise	$\sigma = 0.2$	99.99	99.99	99.99	99.99	99.99
GausBlur	$k = 7$	99.99	99.99	99.99	99.99	99.99
SaltPepper	$p = 10\%$	99.99	99.96	99.57	98.97	99.62
Dropout	$p = 30\%$	99.99	99.99	99.99	99.96	99.98
Cropout	$p = 30\%$	99.58	99.89	98.82	95.36	98.41
Saturation	$f = 2$	99.96	99.79	99.70	99.26	99.68
Hue	$f = 0.1$	99.99	99.96	95.05	86.06	95.27
Brightness	$f = 1.5$	99.32	99.28	99.23	99.08	99.23
Contrast	$f = 1.5$	99.99	99.44	99.19	98.18	99.20
Resize	$p = 70\%$	99.99	99.99	99.10	94.47	98.39
Rotation	$d = 15$	98.78	97.33	95.83	94.69	96.66
Affine	$s = 10$	98.75	94.10	89.11	86.91	92.22

Table 4: Robustness to noise at various intensity factors of the noise parameter

Noise	Parameter	PSNR(dB)↑	SSIM(%)↑	ACC(%)↑
GausNoise	$\sigma = 0.2$	52.72	99.19	99.99
GausBlur	$k = 7$	50.50	98.22	99.99
SaltPepper	$p = 10\%$	45.83	96.79	99.99
Dropout	$p = 30\%$	47.51	97.89	99.99
Cropout	$p = 30\%$	43.76	95.64	99.99
Saturation	$f = 2$	45.19	96.38	99.99
Hue	$f = 0.1$	47.41	97.15	99.99
Brightness	$f = 1.5$	44.55	96.60	99.88
Contrast	$f = 1.5$	45.20	95.64	99.99
Resize	$p = 70\%$	49.90	98.88	99.99
Rotation	$d = 15$	42.97	94.87	99.99
Affine	$s = 10$	45.34	96.03	99.13
Average	-	46.74	96.94	99.91

Table 5: Performance in various kinds of attacks

Individual-Noise Performance In order to further test the performance of IWRN in the face of single noise, We further train the base model used in previous experiments against individual types of noise to obtain 12 Individual-Noise models. As shown in Table 5, the average PSNR, SSIM and ACC of Individual-Noise models make a significant improvement reaching 46.74dB, 96.94% and 99.91% respectively, which proves that IWRN can achieve superior imperceptibility and robustness against individual noise attack.

Ablation Study In our ablation study, we trained multiple models using the noise pool N_{All} with different architectures, evaluating their performance by measuring PSNR and average ACC. Four modules in the framework are discussed: DA-INN, LWN, DOM and JCL. In the ablation experiments, when the DA-INN is not utilized, a conventional dense block replaces the DA-dense block. In the absence of LWN, standard convolution operations are employed for upsampling and downsampling to maintain tensor consistency. When the DOM is omitted, the watermark decoding is conducted using a fully connected layer to decode message from I_{RS}. Through the ablation experiments

Module				PSNR(DB)↑	ACC(%)↑
DAI	LWN	DOM	JCL		
	✓	✓	✓	37.19	96.02
✓		✓	✓	38.39	98.59
✓	✓		✓	35.85	95.10
✓	✓	✓		40.16	98.92
✓	✓	✓	✓	41.55	99.57

Table 6: Ablation study results

Stage	#Params (M)	#VRAM (GB)	Time	
			H/100 Epoch	M/1 Epoch
Train	127.1	32.56	9.11	–
Test	127.1	1.59	–	3.40

Table 7: Computational Costs

shown in Table 6, four modules prove their effects in the framework. DA-INN, LWN, DOM and JCL makes 11.72%, 8.23%, 15.89%, 3.46% increasing in PSNR and 3.69%, 0.99%, 4.70%, 0.65% increasing in ACC, respectively.

Computational Costs As a highly practical technology, the computing resources and decoding speed required by digital watermarking technology are also important indicators. As shown in Table 7, IWRN, as an INN-based architecture, has 127.1M parameters, which is comparable to the classic VGG-16 model. The training process requires 32.56GB of GPU memory, so it is recommended to use an RTX A6000 or reduce the batch size below 16. The amount of GPU memory used during the testing phase is 1.59GB, which can be run on most GPU for testing and practical use. When the batch size is set to 16, training 15,500 images over 100 epochs takes 9.11 hours, and testing 10,500 images requires 3.40 minutes. Therefore, IWRN can be trained and tested in most laboratories, and it has excellent encoding and decoding speeds.

Conclusion

In this paper, we propose a blind watermarking method for artwork image copyright protection, IWRN, which can ensure both the imperceptibility of the watermark and its robustness against 12 kinds of noise attacks. We first design LWN, where the watermark is less easily perceived by the human eye, thereby improving the imperceptibility of the watermark. Then, we establish DA-INN with a DOM, which offers the advantage of computational reversion, and combines the deform-attention mechanism and decoding optimization to enhance the model's robustness against noises. Additionally, we design JCL mechanism for encoder and decoder to improve imperceptibility and robustness simultaneously. Experiments show that our IWRN performs well in terms of reversibility and robustness, and compared with other state-of-the-art methods, it improves the PSNR, SSIM and accuracy by an average of 16.79%, 5.76% and 7.67% respectively and gets the best MOS score. In future work, we will consider extending our ideas to the task of copyright protection of artwork videos.

Acknowledgments

The work is supported by Fundamental Research Funds for the Central Universities (No. 2023RC08), by the National Science and Technology Major Project (2023ZD0121503), by scientific research program of Beijing Municipal Education Commission KZ202110011017, by National Natural Science Foundation of China (U22B2038, 62272058, U23A20319, 62277001, 62402055), by the Open Research Subject of State Key Laboratory of Intelligent Game (No. ZBKF-24-12). by the 8th Young Elite Scientists Sponsorship Program by CAST (2022QNRC001), by Beijing Natural Science Foundation (No.L233034).

References

Arab, M. A.; Ghorbanpour, A.; and Hefeeda, M. 2024. FlexMark: Adaptive Watermarking Method for Images. In *Proceedings of the 15th ACM Multimedia Systems Conference*, 56–66.

Barni, M.; Bartolini, F.; and Piva, A. 2001. Improved wavelet-based watermarking through pixel-wise masking. *IEEE transactions on image processing*, 10(5): 783–791.

Bassia, P.; Pitas, I.; and Nikolaidis, N. 2001. Robust audio watermarking in the time domain. *IEEE Transactions on multimedia*, 3(2): 232–241.

Begum, M.; and Uddin, M. S. 2020. Digital image watermarking techniques: a review. *Information*, 11(2): 110.

Chen, X.; and He, K. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15750–15758.

Darvish Rouhani, B.; Chen, H.; and Koushanfar, F. 2019. Deepsigns: An end-to-end watermarking framework for ownership protection of deep neural networks. In *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems*, 485–497.

Das, U. K.; Samaddar, S. G.; and Keserwani, P. K. 2018. Digital forensic enabled image authentication using least significant bit (lsb) with tamper localization based hash function. In *Intelligent Communication and Computational Technologies: Proceedings of Internet of Things for Technological Development, IoT4TD 2017*, 141–155. Springer.

Dinh, L.; Krueger, D.; and Bengio, Y. 2014. Nice: Nonlinear independent components estimation. *arXiv preprint arXiv:1410.8516*.

Fang, H.; Jia, Z.; Ma, Z.; Chang, E.-C.; and Zhang, W. 2022. PIMoG: An effective screen-shooting noise-layer simulation for deep-learning-based watermarking network. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2267–2275.

Fang, H.; Qiu, Y.; Chen, K.; Zhang, J.; Zhang, W.; and Chang, E.-C. 2023. Flow-based robust watermarking with invertible noise layer for black-box distortions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 5054–5061.

Guan, Z.; Jing, J.; Deng, X.; Xu, M.; Jiang, L.; Zhang, Z.; and Li, Y. 2022. DeepMIH: Deep invertible network for multiple image hiding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1): 372–390.

Guo, H.; Zhang, Q.; Luo, J.; Guo, F.; Zhang, W.; Su, X.; and Li, M. 2023. Practical Deep Dispersed Watermarking with Synchronization and Fusion. In *Proceedings of the 31st ACM International Conference on Multimedia*, 7922–7932.

Hsu, C.-T.; and Wu, J.-L. 1999. Hidden digital watermarks in images. *IEEE Transactions on image processing*, 8(1): 58–68.

Huang, J.; Luo, T.; Li, L.; Yang, G.; Xu, H.; and Chang, C.-C. 2023. ARWGAN: Attention-Guided Robust Image Watermarking Model Based on GAN. *IEEE Transactions on Instrumentation and Measurement*, 72: 1–17.

Jia, Z.; Fang, H.; and Zhang, W. 2021. Mbrs: Enhancing robustness of dnn-based watermarking by mini-batch of real and simulated jpeg compression. In *Proceedings of the 29th ACM international conference on multimedia*, 41–49.

Jing, J.; Deng, X.; Xu, M.; Wang, J.; and Guan, Z. 2021. HiNet: deep image hiding by invertible network. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4733–4742.

Khan, A.; Wong, K.; and Baskaran, V. M. 2024. Trade-off independent image watermarking using enhanced structured matrix decomposition. *Multimedia Tools and Applications*, 1–29.

Kundur, D.; and Hatzinakos, D. 1997. A robust digital image watermarking method using wavelet-based fusion. In *Proceedings of International Conference on Image Processing*, volume 1, 544–547. IEEE.

Liu, Y.; Guo, M.; Zhang, J.; Zhu, Y.; and Xie, X. 2019. A novel two-stage separable deep learning framework for practical blind watermarking. In *Proceedings of the 27th ACM International conference on multimedia*, 1509–1517.

Luo, Y.; Zhou, T.; Liu, F.; and Cai, Z. 2023. IRWArt: Levering Watermarking Performance for Protecting High-quality Artwork Images. In *Proceedings of the ACM Web Conference 2023*, 2340–2348.

Ma, R.; Guo, M.; Hou, Y.; Yang, F.; Li, Y.; Jia, H.; and Xie, X. 2022a. Towards Blind Watermarking: Combining Invertible and Non-invertible Mechanisms. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1532–1542.

Ma, R.; Guo, M.; Hou, Y.; Yang, F.; Li, Y.; Jia, H.; and Xie, X. 2022b. Towards Blind Watermarking: Combining Invertible and Non-invertible Mechanisms. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1532–1542.

Qu, X.; Yin, X.; Wei, P.; Lu, L.; and Ma, Z. 2023. AudioQR: deep neural audio watermarks for QR code. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 6192–6200.

Tancik, M.; Mildenhall, B.; and Ng, R. 2020. Stegastamp: Invisible hyperlinks in physical photographs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2117–2126.

Van Schyndel, R. G.; Tirkel, A. Z.; and Osborne, C. F. 1994. A digital watermark. In *Proceedings of 1st international conference on image processing*, volume 2, 86–90. IEEE.

Wan, W.; Wang, J.; Zhang, Y.; Li, J.; Yu, H.; and Sun, J. 2022. A comprehensive survey on robust image watermarking. *Neurocomputing*, 488: 226–247.

Wang, C.; Li, S.; Liu, Y.; Meng, L.; Zhang, K.; and Wan, W. 2023. Cross-scale feature fusion-based JND estimation for robust image watermarking in quaternion DWT domain. *Optik*, 272: 170371.

Wang, G.; Ma, Z.; Liu, C.; Yang, X.; Fang, H.; Zhang, W.; and Yu, N. 2024a. MuST: Robust Image Watermarking for Multi-Source Tracing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5364–5371.

Wang, H.; Wang, H.; Xia, J.; and Zhang, F. 2024b. Robust Watermarking against Camera Shooting for PowerPoint Presentation. *IEEE Signal Processing Letters*.

Xiao, D.; Zhao, A.; and Li, F. 2022. Robust watermarking scheme for encrypted images based on scrambling and Kronecker compressed sensing. *IEEE Signal Processing Letters*, 29: 484–488.

Xu, Y.; Mou, C.; Hu, Y.; Xie, J.; and Zhang, J. 2022. Robust invertible image steganography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7875–7884.

Yang, K.; Wang, R.; and Wang, L. 2022. Metafinger: Fingerprinting the deep neural networks with meta-training. In *31st International Joint Conference on Artificial Intelligence (IJCAI-22)*.

Yu, C. 2020. Attention based data hiding with generative adversarial networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 1120–1128.

Zhang, H.; Wang, H.; Cao, Y.; Shen, C.; and Li, Y. 2020a. Robust data hiding using inverse gradient attention. *arXiv preprint arXiv:2011.10850*.

Zhang, J.; Chen, D.; Liao, J.; Fang, H.; Zhang, W.; Zhou, W.; Cui, H.; and Yu, N. 2020b. Model watermarking for image processing networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 12805–12812.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhu, J.; Kaplan, R.; Johnson, J.; and Fei-Fei, L. 2018. Hidden: Hiding data with deep networks. In *Proceedings of the European conference on computer vision (ECCV)*, 657–672.