*mathematics*

MDPI

*Article*

# DARI-Mark: Deep Learning and Attention Network for Robust Image Watermarking

**Yimeng Zhao [1], Chengyou Wang [1,2,*], Xiao Zhou [1,2] and Zhiliang Qin [1,3]**

[1] School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai 264209, China
[2] Shandong University–Weihai Research Institute of Industry Technology, Weihai 264209, China
[3] Weihai Beiyang Electric Group Co. Ltd., Weihai 264209, China
* Correspondence:wangchengyou@sdu.edu.cn; Tel.: +86-631-568-8338

**Abstract:** At present, deep learning has achieved excellent achievements in image processing and computer vision and is widely used in the field of watermarking. Attention mechanism, as the research hot spot of deep learning, has not yet been applied in the field of watermarking. In this paper, we propose a deep learning and attention network for robust image watermarking (DARI-Mark). The framework includes four parts: an attention network, a watermark embedding network, a watermark extraction network, and an attack layer. The attention network used in this paper is the channel and spatial attention network, which calculates attention weights along two dimensions, channel and spatial, respectively, assigns different weights to pixels in different channels at different positions and is applied in the watermark embedding and watermark extraction stages. Through end-to-end training, the attention network can locate nonsignificant areas that are insensitive to the human eye and assign greater weights during watermark embedding, and the watermark embedding network selects this region to embed the watermark and improve the imperceptibility. In watermark extraction, by setting the loss function, larger weights can be assigned to watermark-containing features and small weights to noisy signals, so that the watermark extraction network focuses on features about the watermark and suppresses noisy signals in the attacked image to improve robustness. To avoid the phenomenon of gradient disappearance or explosion when the network is deep, both the embedding network and the extraction network have added residual modules. Experiments show that DARI-Mark can embed the watermark without affecting human subjective perception and that it has good robustness. Compared with other state-of-the-art watermarking methods, the proposed framework is more robust to JPEG compression, sharpening, cropping, and noise attacks.

**Keywords:** image watermarking; deep learning; attention network; imperceptibility; robustness

**MSC:** 94A08, 94A62, 68T07, 68U10



**Citation:** Zhao, Y.; Wang, C.; Zhou, X.; Qin, Z. DARI-Mark: Deep Learning and Attention Network for Robust Image Watermarking. *Mathematics* **2023**, *11*, 209. https://doi.org/10.3390/math11010209

Academic Editor: Jakub Nalepa

Received: 27 November 2022
Revised: 19 December 2022
Accepted: 26 December 2022
Published: 31 December 2022



## 1. Introduction

In recent years, malicious image tampering has been a common phenomenon, and effective copyright protection methods are urgently needed. Digital watermarking [1] is used for ownership determination and intellectual property protection by hiding logos or specific information in images, audios, or videos, and has become a research hot spot in the field of information security [2]. Considering the practical application requirements, the main performance evaluation metrics of image watermarking are embedding capacity, imperceptibility, and robustness [3]. According to the different embedding domains, existing image watermarking methods are mainly divided into two categories: spatial-domain algorithms and transform-domain algorithms [4]. Spatial-domain watermarking algorithms embed the watermark by directly modifying the pixel values, such as the least significant bit (LSB) [5], patchwork [6], etc. The algorithm complexity is low, but the robustness is poor.

Transform-domain watermarking algorithms embed the watermark by modifying the image transform-domain coefficients. The commonly used frequency-domain transformation methods are discrete cosine transform (DCT) [7], discrete wavelet transform (DWT) [8], discrete Fourier transform (DFT) [9], etc. The DCT-based watermarking algorithms are simple, imperceptible, and robust, so they are the preferred frequency-domain transformation method. However, the performance of the conventional watermarking algorithms is limited by the method itself, and the performance needs to be improved.

Currently, machine learning [10] has made remarkable achievements in the fields of computer vision, pattern recognition, artificial intelligence, etc. Many machine learning methods have also been introduced in the watermarking field, such as support vector machines (SVM) [11], extreme learning machines (ELM) [12], Lagrangian support vector regression (LSVR) [13], K-nearest neighbors (K-NN) [14], etc. To achieve better performance, these machine learning methods are typically used with frequency-domain transformation methods. Deep learning has also achieved remarkable results in the field of watermarking [15]. The attention mechanism simulates the ability of human brain neurons to automatically capture important information, to rationally allocate computational resources. It can achieve a large performance improvement with a small computational overhead and has become a hot spot in the field of deep learning [16], such as the squeeze-and-excitation network (SENet) [17] and convolutional block attention module (CBAM) [18]. SENet can obtain the importance of different channels through a simple network structure and achieve the enhancement or suppression of channel features. CBAM extracts attention from both channel and spatial dimensions and is trained to focus on useful regions and suppress irrelevant information. However, the application of attention mechanism to image watermarking is still in its infancy and needs further research.

In this paper, a deep learning and attention network for robust image watermarking (DARI-Mark) is proposed. The main novelties and contributions are as follows:

(1) We introduce the attention network, which makes watermark embedding focus on the edge and other areas with complex texture, and the watermark extraction pays more attention to watermarked features and suppresses noise signals. This greatly improves the imperceptibility and robustness of DARI-Mark.

(2) We add the residual module to avoid gradient explosion or disappearance problems in deep neural networks.

(3) We establish an end-to-end network framework and add the attack layer between watermark embedding and extraction networks to improve the robustness against multiple attacks (JPEG compression, sharpening, cropping, Gaussian noise, salt-and-pepper noise, scaling) through iterative training.

The rest of this paper is organized as follows. Section 2 introduces related work. Section 3 presents the proposed watermarking framework and training details. Experimental results and the analysis of the proposed framework are presented in Section 4. Finally, Section 5 presents the conclusions and possible further work.

## 2. Related Work

Due to the rapid development of machine learning and deep learning, more and more watermarking algorithms based on machine learning and deep learning are being proposed. Many machine learning methods have been introduced to improve imperceptibility and robustness. For example, to improve the imperceptibility of the algorithm, Rai et al. [11] used an SVM to classify medical images into regions of interest and regions of noninterest. To improve robustness, Mehta et al. [13] trained an LSVR using low-frequency wavelet sub-band coefficients and used the trained LSVR to extract the watermark. Abdelhakim et al. [14] used a K-NN regression algorithm to predict the optimal embedding strength of the watermark, balancing imperceptibility and robustness.

Neural network frameworks for watermark embedding and extraction have also been widely proposed. For example, Kandi et al. [19] were the first to apply convolutional neural network technology to the field of watermarking. Zhu et al. [20] designed the first

watermarking framework HiDDeN that could be trained end to end, which could be used for steganography and watermarking. Mun et al. [21] used two watermark embedding methods, backpropagation and autoencoder, and the embedding strength was adjusted using a visual mask. Liu et al. [22] divided the whole training process into two training stages: training the encoder without considering the noise and adjusting the decoder parameters according to the noise. Luo et al. [23] introduced the channel coding module in the deep watermarking model to improve the robustness by increasing information redundancy. Lee et al. [24] proposed a deep neural network suitable for multiresolution images and a variety of practical application scenarios. Ahmadi et al. [25] added an attack layer between the encoder and decoder to simulate multiple image distortion attacks for end-to-end training. Jia et al. [26] randomly selected a real/simulated JPEG compression at the noise layer to achieve high robustness against JPEG compression attacks. Zhong et al. [27] trained the watermarking framework in an unsupervised way and achieved a better robustness without any prior knowledge. Sun et al. [28] simulated image distortion operations on online social platforms through deep neural networks and learned lossy channel information to achieve a high robustness for the platforms. These methods improved robustness but did not strike a good balance between imperceptibility and robustness.

To improve imperceptibility, the subjective feelings of people began to be considered instead of only using objective evaluation criteria, so the human visual system (HVS) and attention mechanisms began to be introduced into the image watermarking domain. For example, Yang et al. [29] selected CIEDE2000, which is more in line with the HVS, as an image quality evaluation index and proposed the B-R-G embedding principle. Yu [30] learned attention masks through convolutional neural networks to embed the watermark in locations that are insensitive to human vision. Jia et al. [26] introduced the channel attention network in the watermark embedding and extraction process. Bhowmik et al. [31] combined wavelet decomposition and the HVS for significance modeling and adaptively embedding watermarks based on saliency. However, the attention mechanism in deep learning has not really been applied to watermarking algorithms.

In this paper, we propose a deep learning and attention network for robust image watermarking: DARI-Mark. The channel and spatial attention network is introduced into the existing end-to-end watermarking framework, and through correct loss function settings and multiple training iterations, the network automatically finds human-insensitive texturally complex regions of the cover image to embed the watermark during watermark embedding, suppress noise signals and improve the extraction of watermark-related features during watermark extraction, improving imperceptibility and robustness. The DCT is also applied to the image features before embedding the watermark, which adaptively embeds the watermark features in the frequency-domain features, which has less impact on the visual perception and a better imperceptibility.

## 3. Proposed Framework

This section presents the DARI-Mark in detail, and its flow chart is shown in Figure 1. The framework is mainly divided into four networks:

(1) An attention network, which finds complex regions of texture that are insensitive to the human eye to embed the watermark during watermark embedding, focuses on watermark-related features, and suppresses noise signals during watermark extraction.

(2) A watermark embedding network, which is responsible for embedding the watermark.

(3) An attack layer, which simulates attacks that are commonly encountered in the actual communication process and improves the robustness of the framework through end-to-end training.

(4) A watermark extraction network, which is responsible for the extraction of watermark information.

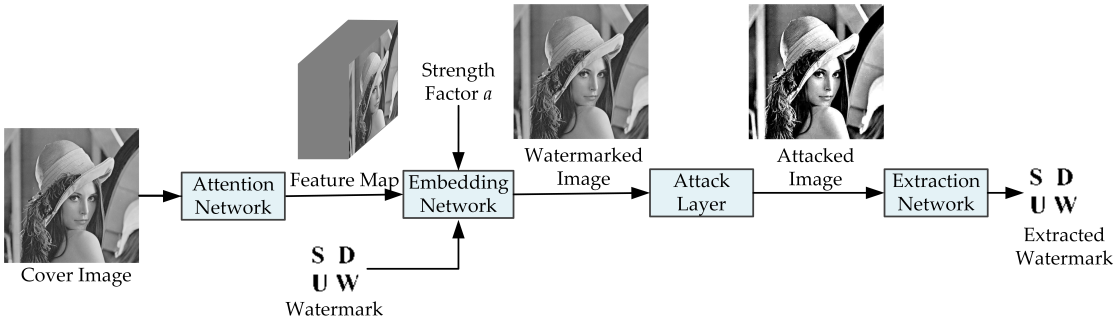


**Figure 1.** Flow chart of DARI-Mark.

The structure and working details of each subnetwork are as follows.

### *3.1. Attention Network*

In practical applications, watermarking algorithms are required to have good imperceptibility and robustness. To achieve this goal, we wanted to find regions for watermark embedding that had little impact on the HVS and focus on watermark features during watermark extraction, so an attention network was introduced. This intuition was derived from the classic attention mechanism CBAM [18], which sequentially infers attention values along both channel and spatial dimensions and generates an attention feature map.

Figure 2 shows the flow chart of the attention network. The input is the cover image with a size of $M \times N$; through a $3 \times 3$ convolution, the cover feature with 64 channels is generated. Then, features are extracted by max-pooling and average pooling operations on the cover feature map respectively, and then they are input into a fully connected (FC) layer to generate attention eigenvalue vectors. The two vectors are summed element by element to obtain the channel attention weight parameters, which are multiplied by the cover feature map to obtain the channel attention feature map. The spatial relationship of the channel attention feature map is extracted by max-pooling and average pooling layers. Then, the features are connected by a $3 \times 3$ standard convolution layers to generate spatial attention weights. Multiply the spatial attention weights with the channel attention feature map to generate an attention feature map with a size of $M \times N \times 64$.

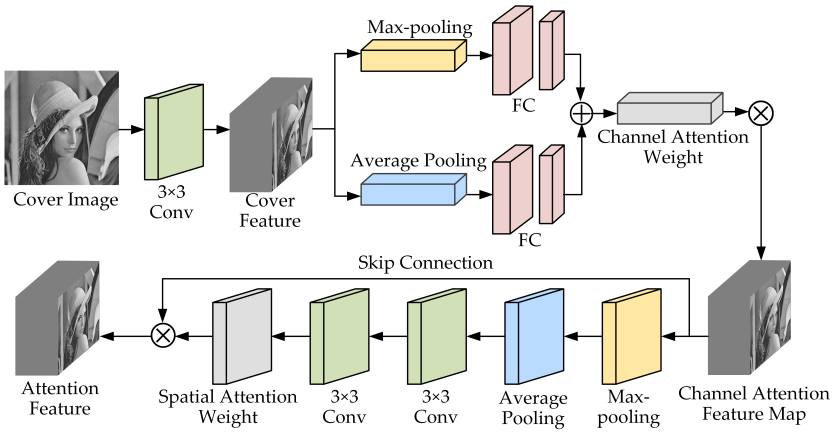


**Figure 2.** Attention network: preprocessing of the cover image.

Although the network structure is only slightly different from CBAM, the purpose of the implementation is very different. CBAM is used to locate salient regions of an image. In contrast, the proposed attention network is used to locate nonsignificant regions that are insensitive to the human eye during watermark embedding and to assign greater weight to them, selecting such regions to embed the watermark to improve imperceptibility. When extracting the watermark, this attention network is used to focus on the features of the watermark and suppress noise signals in the attacked image to improve robustness.

### *3.2. Embedding Network*

Figure 3 shows the flow chart of the watermark embedding network, the detailed process is as follows:

(1) DCT layer: A DCT is performed on the attention feature map generated by the attention network. This is because embedding the watermark in the DCT is consistent with the frequency-masking effect of the HVS, and the human eye has a large acceptable range of changes.

(2) Watermark embedding: The network structure is a $1 \times 1$ Conv and four $2 \times 2$ Conv. Each Conv includes convolution, batch normalization, and ReLU activation functions, and the number of convolution kernels is 64. The input is the frequency-domain information after the DCT transformation and watermark information, and the output is the information after the feature fusion. The $1 \times 1$ convolutional layer is to change the dimension of the tensor. The $2 \times 2$ convolutional layers embed the watermark data into bottom features. The $2 \times 2$ convolutional layers are designed as residual structures, with each residual structure corresponding to two convolutional layers, that is, a shortcut connection is added every two layers.

(3) Inverse DCT (IDCT): an inverse DCT is performed on the tensor generated by the watermark embedding network to transform it from the frequency domain back to the spatial domain.

(4) Embedding strength factor $\alpha$: The embedding strength factor $\alpha$ controls the embedding strength of the watermark, achieving the balance between imperceptibility and robustness. The tensor generated in step (3) is multiplied by $\alpha$ and then added to the attention feature map to generate the watermarked tensor. The experimental results showed that with the increase of $\alpha$, the robustness was enhanced. To enhance the robustness of the algorithm, $\alpha$ was taken as 1.0 in the training phase. In the actual application phase, $\alpha$ can be appropriately adjusted according to application requirements.

(5) A $3 \times 3$ Conv: The number of convolution kernels is 1. The input is the output of step (4), and the output is the watermarked image with a size of $M \times N$, the same as that of the cover image.

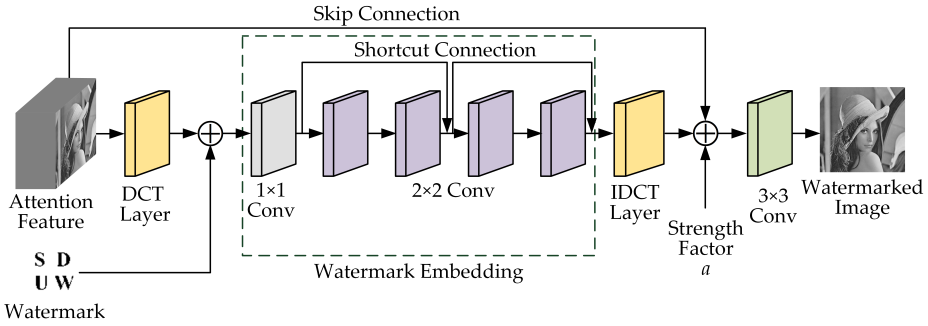


**Figure 3.** Embedding network: responsible for watermark embedding.

### *3.3. Attack Layer*

To enhance the algorithm's robustness to distortion attacks, DARI-Mark adds an attack layer between the watermark embedding network and the watermark extraction network. It simulates attacks that are commonly encountered in actual communications. Figure 4 shows the flow chart of the attack layer, which simulates four common attacks: JPEG compression (quality factor $Q = 50$), salt-and-pepper noise (noise density $s = 0.04$), Gaussian noise (standard deviation $\sigma = 0.1$), and sharpening (sharpening radius $R_S = 10$), forming a distinguishable attack layer.

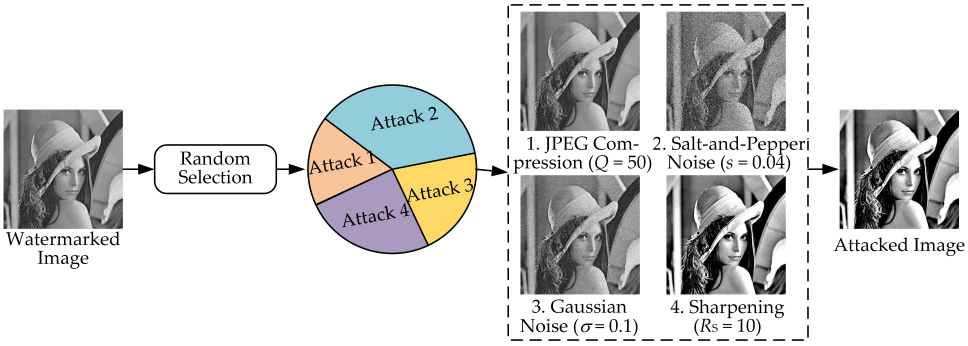


**Figure 4.** Attack layer: simulate four common attacks.

The probability of each attack is randomly assigned in each iteration to improve the generalization ability while satisfying the condition that the total probability must be 1.0. In each round of iterative training, attack methods are randomly selected. The network parameters are continuously updated by the loss function and optimization algorithm to generate a robust watermark extraction network.

### *3.4. Extraction Network*

The extraction network is the inverse process of the embedding network, as shown in Figure 5.

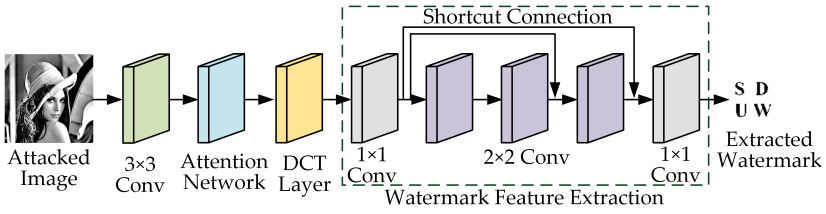


**Figure 5.** Extraction network: responsible for watermark extraction.

The detailed process is as follows:

(1) A $3 \times 3$ Conv: change the attacked image from $M \times N \times 1$ to $M \times N \times 64$.

(2) Attention network: The tensor with 64 channels is fed into the attention network to obtain the attention feature map. The attention feature map is used to provide the watermark feature extraction network with areas to focus on and to suppress.

(3) DCT: This is the same as step (1) of the embedding network. Because the watermark information is embedded in the frequency domain, the image features are first changed from the spatial domain to the frequency domain.

(4) Watermark feature extraction: The input is the DCT transformed tensor, and the output is the extracted watermark. The convolutional network parameters setting are basically consistent with the embedded network. Each Conv includes convolution, batch normalization, and ReLU activation functions. Since the final output is a watermark image with the channel number of 1, the number of convolution kernels in the last layer is set to 1. As with the watermark embedding, the $2 \times 2$ convolutional layers are also designed as residual structures, and since there are only three layers of $2 \times 2$ convolution, two residual structures are designed, corresponding to two and three convolutional layers, respectively.

### *3.5. Training Details*

The above four networks were encapsulated into an overall framework for end-to-end training. The watermark used in the training phase was randomly generated, and the optimizer used was a stochastic gradient descent (SGD). The specific network training parameters are shown in Table 1.

**Table 1.** The parameters setting in the training stage of the proposed framework.

| Parameters | Setting |
|---|---|
| Learning rate | 0.0001 |
| Momentum | 0.98 |
| Batch size | 32 |
| Iteration | 10,000 |
| Epoch | 100 |

The optimization goal of the framework was to minimize the difference between the cover image and the watermarked image and maximize the correct rate of watermarking extraction. Here, structural similarity (SSIM) and binary cross-entropy (BCE) were used as loss functions of the embedding network and the extraction network. Although mean squared error (MSE) and peak signal-to-noise ratio (PSNR) can also represent the difference between two images, they only represent the difference in pixel points at the same location, ignoring information such as the local structure of images. SSIM is a perceptual model that evaluates the difference between images from three aspects: brightness, contrast, and structure, which is more in line with the HVS. The purpose of introducing the attention network in this paper was to generate watermarked images that were more in line with human subjective perception, so SSIM was chosen as the loss function of the embedding network, denoted by $L_{\text{emb}}$:

$$L_{\text{emb}} = \text{SSIM}(\boldsymbol{I}, \boldsymbol{I}_{\text{W}}) = \frac{(2\mu_{\boldsymbol{I}}\mu_{\boldsymbol{I}_{\text{W}}} + c_1)(2\sigma_{\boldsymbol{I},\boldsymbol{I}_{\text{W}}} + c_2)}{(\mu_{\boldsymbol{I}}^2\mu_{\boldsymbol{I}_{\text{W}}}^2 + c_1)(\sigma_{\boldsymbol{I}}^2\sigma_{\boldsymbol{I}_{\text{W}}}^2 + c_2)} \tag{1}$$

where $\boldsymbol{I}$ is the cover image; $\boldsymbol{I}_{\text{W}}$ is the watermarked image; $\mu_{\boldsymbol{I}}$ is the average value of $\boldsymbol{I}$; $\sigma_{\boldsymbol{I}}$ is the standard deviation of $\boldsymbol{I}$; $\mu_{\boldsymbol{I}_{\text{W}}}$ is the average value of $\boldsymbol{I}_{\text{W}}$; $\sigma_{\boldsymbol{I}_{\text{W}}}$ is the standard deviation of $\boldsymbol{I}_{\text{W}}$; $\sigma_{\boldsymbol{I},\boldsymbol{I}_{\text{W}}}$ is the covariance of $(\boldsymbol{I}, \boldsymbol{I}_{\text{W}})$; and $c_1$ and $c_2$ are two metric constants, which were set to $10^{-4}$ and $9 \times 10^{-4}$ in this paper, respectively.

The loss function of the extraction network was represented by $L_{\text{ext}}$, which was defined as:

$$L_{\text{ext}} = -\sum_{i \in \boldsymbol{\Omega}} [y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \tag{2}$$

where $\boldsymbol{\Omega}$ is the image domain; $y_i$ is the pixel value of the real watermark at position $i$; and $\hat{y}_i$ is the pixel value of the extracted watermark at position $i$.

The loss function $L$ of the overall framework was defined as:

$$L = \lambda L_{\text{emb}} + (1 - \lambda) L_{\text{ext}} \tag{3}$$

where $\lambda$ is the weight between the two loss functions, which was set to 0.5 in this paper.

## 4. Experimental Results and Analysis

In this section, we conducted a series of experiments and verified and analyzed the experimental results. Section 4.1 introduces the dataset used in these experiments, Section 4.2 introduces the performance evaluation indexes, Section 4.3 shows the experimental results of the proposed framework, and Section 4.4 compares the proposed framework with the state-of-the-art methods.

The hardware configuration was a Window 10 PC with an Intel (R) Core (TM) i9-9920X CPU @ 3.50GHz and two NVIDIA GeForce RTX 2080 Ti GPUs.

### *4.1. Dataset*

We use the BOSSbase 1.01 dataset [32] and CIFAR 10 dataset [33] as training samples in the training stage. The BOSSbase 1.01 dataset consists of 1000 grayscale images with a size of 512 × 512, and the CIFAR 10 dataset consists of 60,000 32 × 32 color images in 10 categories. Since the image sizes of the two datasets are different, the BOSSbase 1.01 dataset

was preprocessed before training to generate 32 × 32 image blocks and then combined with the CIFAR 10 dataset as the training set.

We use the Granada dataset [34] (standard grayscale images with a size of 512 × 512) as the test set in the testing stage. The Granada dataset covers a variety of image types, which can avoid the unreliability of the experimental data due to the proposed algorithm's certain bias for specific images.

### *4.2. Evaluation Metrics*

For the quantitative evaluation of the imperceptibility of DARI-Mark, we used the classic objective standard PSNR and SSIM to measure the degree of image distortion. PSNR is defined as:

$$V_{\mathrm{PSNR}}(\boldsymbol{I}, \boldsymbol{I}_{\mathrm{W}}) = 10\log_{10}\frac{W \times H \times P_{\mathrm{MAX}}^2}{\sum_{i=0}^{W-1}\sum_{j=0}^{H-1}[\boldsymbol{I}(i,j) - \boldsymbol{I}_{\mathrm{W}}(i,j)]^2} \tag{4}$$

where $W$ is the width of the image; $H$ is the height of the image; $P_{\mathrm{MAX}}$ is the maximum value of image pixels; and $i$, $j$ are the $x$, $y$ coordinates of each pixel, respectively. The higher the value of $V_{\mathrm{PSNR}}$, the better the imperceptibility of DARI-Mark.

For the quantitative evaluation of robustness, we used the bit error rate (BER) and normalized cross-correlation (NCC) as the evaluation index. The BER was used to evaluate the error rate of the original watermark and the extracted watermark, defined as:

$$V_{\mathrm{BER}}(w, w') = \frac{1}{L_{\mathrm{W}}}\sum_{i=1}^{L}|w_i - {w'}_i| \tag{5}$$

where $w$ is the original watermark; $w'$ is the extracted watermark; $w_i$ and ${w'}_i$, respectively, represent the information of the original watermark and the extracted watermark at position $i$; and $L_{\mathrm{W}}$ is the length of the watermark information. The NCC was used to evaluate the similarity between the original watermark and the extracted watermark, defined as:

$$V_{\mathrm{NCC}}(w, w') = \frac{\sum_{i=1}^{L}|w_i \times {w'}_i|}{\sqrt{\sum_{i=1}^{L} {w_i}^2} \times \sqrt{\sum_{i=1}^{L} {{w'}_i}^2}} \tag{6}$$

$V_{\mathrm{BER}} \in [0, 1]$, $V_{\mathrm{NCC}} \in [0, 1]$. The lower the BER, the larger the NCC value, the better the robustness of DARI-Mark.

### *4.3. Results*

We used all images in the Granada dataset as cover images to test the performance of DARI-Mark. To test imperceptibility, we embedded a watermark in each image, then computed the values of PSNR and SSIM, and averageed the results as the final result. To show the impact on the imperceptibility of $\alpha$, we used a variety of embedding intensities ($\alpha \in [0.2, 0.4, 0.6, 0.8, 1.0]$) in the test stage.

Figure 6 takes the gray-scale images Lena and Cameraman as examples, and visually shows watermarked images generated under different strength factors $\alpha$ and the difference from cover images. The PSNR of watermarked images generated by the Lena image under different strength factors $\alpha$ were: 48.24 dB, 44.40 dB, 42.43 dB, 40.14 dB, and 38.30 dB; the SSIM values were 0.9957, 0.9899, 0.9807, 0.9692, and 0.956. The PSNR of watermarked images generated by the cameraman image under different strength factors $\alpha$ were 48.29 dB, 44.10 dB, 42.89 dB, 40.61 dB, and 38.77 dB; the SSIM values were 0.9981, 0.9938, 0.9868, 0.9773, and 0.9656. Since the difference between the cover image and the watermarked image was not obvious, to facilitate observation, the pixel value of the local block of the difference image was multiplied by 30 and placed in the lower right corner. It can be seen

that DARI-Mark adaptively selected pixel points to embed the watermarking according to the attention network during watermark embedding.

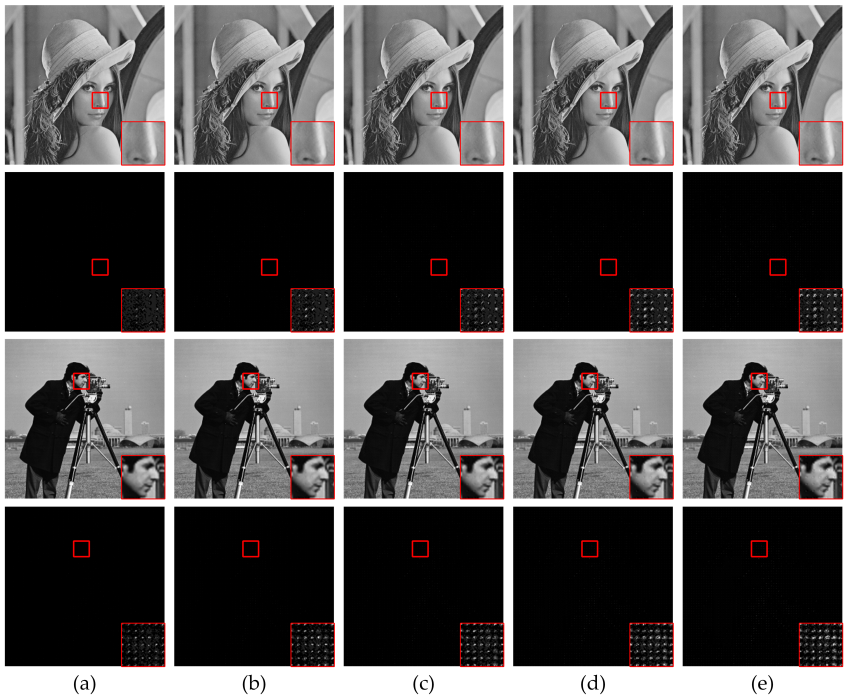


**Figure 6.** The watermarked images generated under different cases of strength factors $\alpha$ and the difference with the cover images: (**a**) $\alpha = 0.2$, (**b**) $\alpha = 0.4$, (**c**) $\alpha = 0.6$, (**d**) $\alpha = 0.8$, and (**e**) $\alpha = 1.0$.

Figure 7 intuitively shows the distribution of PSNR and SSIM values of the watermark images generated under different strength factors $\alpha$ in all images of the test set in the form of boxplots. The experimental results show that with the increase of $\alpha$, the PSNR and SSIM of watermarked images decreased, that is, the quality of the watermarked images decreased. However, even in the case of $\alpha = 1.0$, the PSNR was greater than 35 dB and the normal viewing was not affected. Therefore, in this paper, we selected $\alpha = 1.0$ when testing robustness.

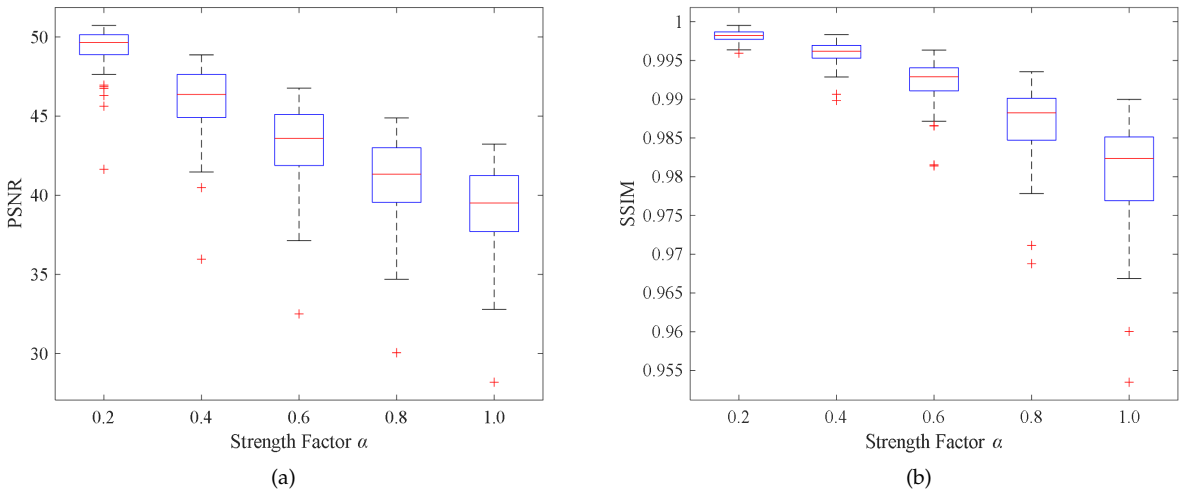


**Figure 7.** Boxplots for quality evaluation of watermarked images generated under different strength factors $\alpha$: (**a**) PSNR and (**b**) SSIM.

To assess DARI-Mark's robustness, we conducted six attacks on watermarked images: JPEG compression, sharpening, cropping (cropping ratio $c$), Gaussian noise, scaling (scaling ratio $r$), and salt-and-pepper noise. The BER and NCC of the extracted watermark are shown in Figures 8 and 9, respectively. It can be seen from the graph that as the value of $\alpha$ increased, the BER of the watermark extraction decreased and the NCC increased, that is, the robustness of DARI-Mark increased. When $\alpha = 1.0$, DARI-Mark was very robust against sharpening, salt-and-pepper noise, and image upscaling in the scaling attack. When $Q$ was 90, 70, 60, and 50, the BER of the watermark extraction was about 0.00, and the NCC was about 1.0000, which shows that the robustness was very good. When $Q$ = 30, that is, the image was overcompressed, the BER of the watermark was 0.11, the NCC of the watermark was 0.8632, and the watermark information was partially damaged. The robustness was good when the cropping ratio $c$ of the cropping attack was between 0.05 and 0.10. As $c$ increased, the BER increased, but it was still within the acceptable range for the human eye to be able to see the watermark information. When the standard deviation $\sigma$ was less than 0.10 in the Gaussian noise attack, the BER was approximately equal to 0.00 and the NCC is approximately equal to 1.0000, and as $\sigma$ increased, the BER increased, and the NCC decreased.

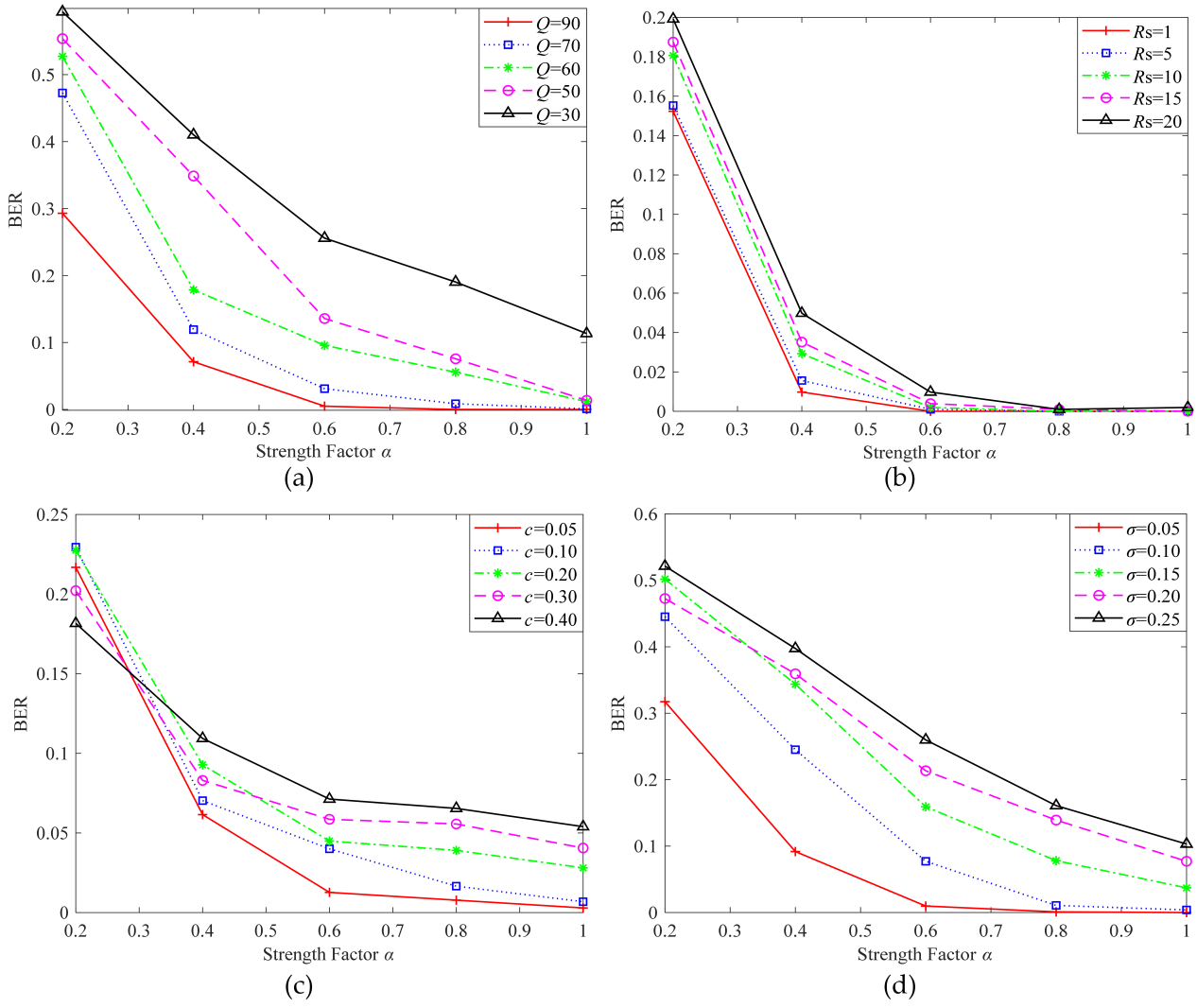


**Figure 8.** *Cont.*

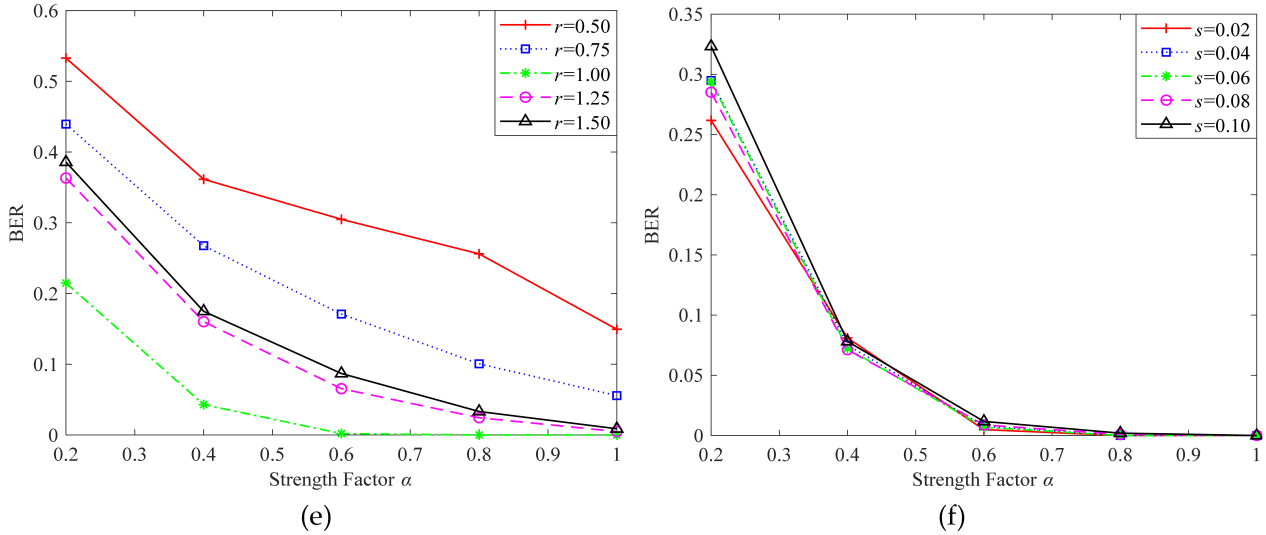


**Figure 8.** BER curves under different attacks: (**a**) JPEG compression, (**b**) sharpening, (**c**) cropping, (**d**) Gaussian noise, (**e**) scaling, and (**f**) salt-and-pepper noise.

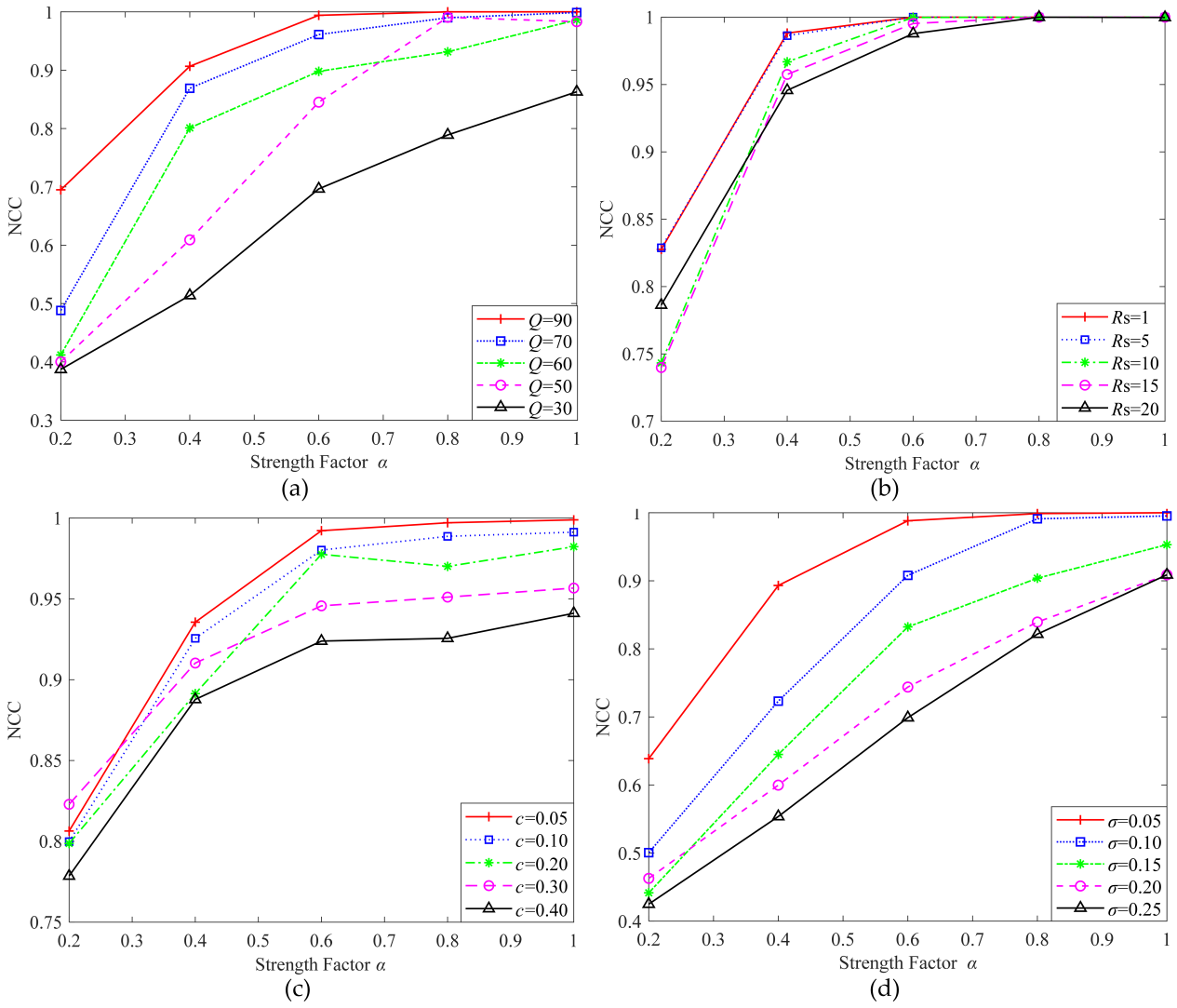


**Figure 9.** *Cont.*

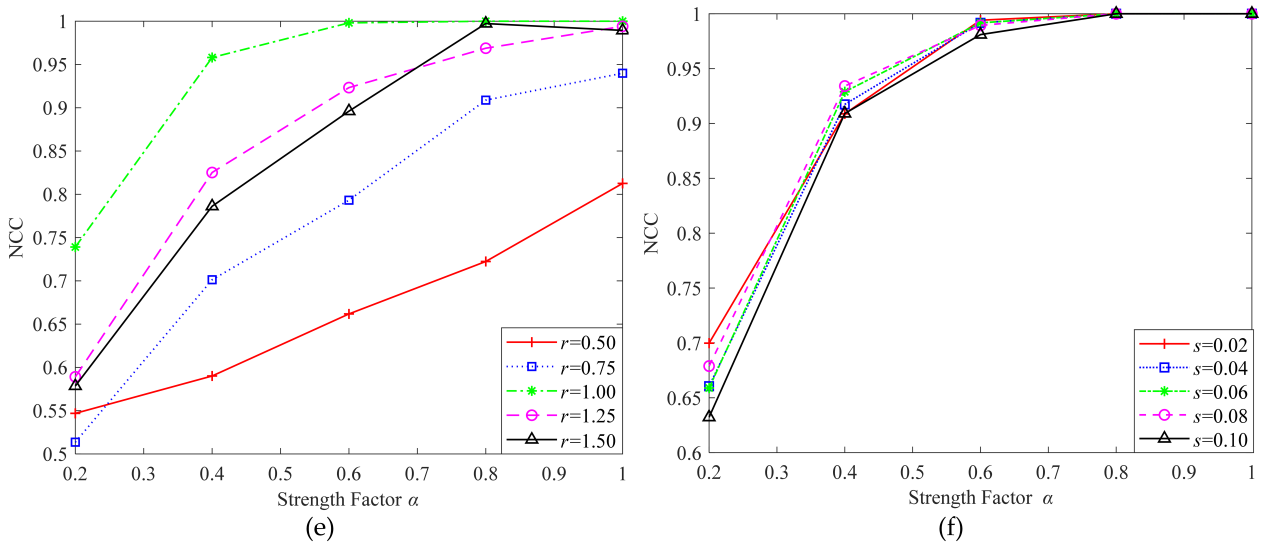


**Figure 9.** NCC curves under different attacks: (**a**) JPEG compression, (**b**) sharpening, (**c**) cropping, (**d**) Gaussian noise, (**e**) scaling, and (**f**) salt-and-pepper noise.

To observe the subjective effect, we took the Lena image as an example to visually show the generated attacked images and watermark extraction results under different attacks when $\alpha = 1.0$, as shown in Figure 10. It can be seen that even if the watermarked image was attacked and the image quality was damaged, the BER of the extracted watermark was less than 8%, indicating that the effect of extracting the watermark could still meet the demand of copyright protection. For example, the attack network did not include a cropping attack, that is, it was not trained on cropping attacks. However, in the testing phase, when any 40% area of the watermarked image was cropped, the BER value was 0.0586 and the NCC was 0.9457. The subjective effect of the attacked image and extracted watermark are shown in the third row and the fourth column of Figure 10, and the watermark information can still be obtained. This shows that DARI-Mark had good robustness and generalization ability.

### *4.4. Comparison with the State-of-the-Art Methods*

Lee et al. [24] and ReDMark [25] used standard grayscale images and binary watermarks in training and testing, and to improve robustness, the same method was used to simulate an attack method in the training phase, which was consistent with DARI-Mark. Therefore, to evaluate and discuss the performance of DARI-Mark, the proposed method was compared with the state-of-the-art methods: Lee et al.'s method [24] and ReDMark [25]. In the original paper, Lee et al. [24] embedded the watermark in the Y channel (luminance) of the color image. When replicating, we changed it to embed the watermark directly in the grayscale image. To ensure the reliability of the comparison results, the same training set, validation set, test set, and randomly generated binary watermarks were used when replicating [24,25]. During training, the attack layers of the three methods were the same, including JPEG compression ($Q$ = 50), Gaussian noise ($\sigma$= 0.1), salt-and-pepper noise ($s$ = 0.04), and sharpening ($R_S$ = 10). The rest is consistent with the network structure of the original paper.

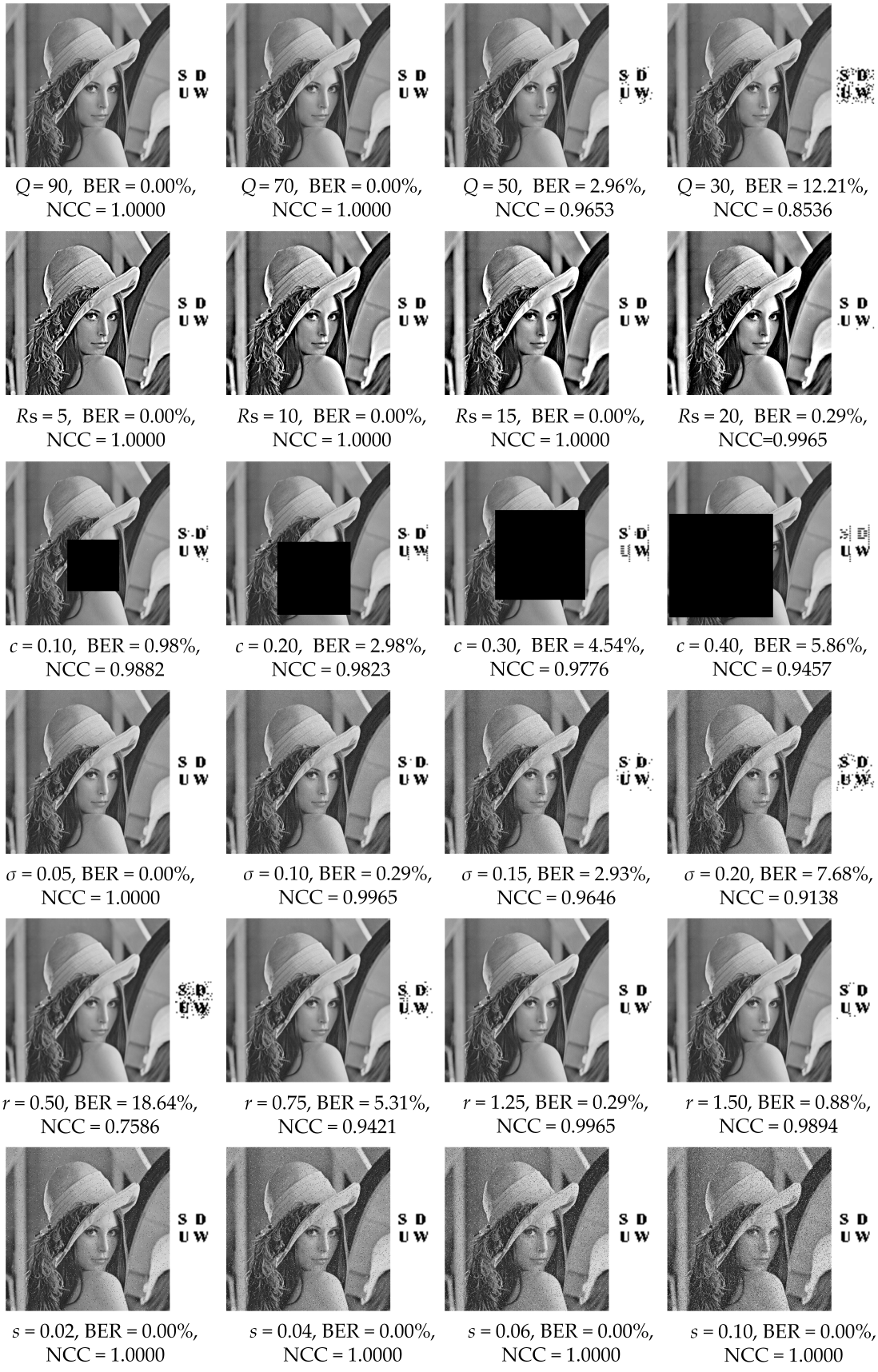


**Figure 10.** Visualization of attacked images and extracted watermarks.

It is known from Section 4.3 that as the $\alpha$ increases, DARI-Mark's imperceptibility decreases and its robustness increases. Since the focus of this paper was on robust watermarking, we selected $\alpha = 1.0$ for comparison. The PSNR of the three methods are shown in Table 2. It can be seen that the PSNR of the watermarked image generated by DARI-Mark was improved, and it had better imperceptibility. To compare the robustness, the imperceptibility of the three methods was the same by modifying the watermark embedding strength $\alpha$, that is, the PSNR of the generated watermarked image was the same. The robust performance comparison results are shown in Table 3 (the BER values are presented as percentages). The data with the lowest BER and highest NCC of the three methods under each attack are bolded to indicate which method was more robust under that attack. It can be seen that DARI-Mark was more robust to JPEG compression, sharpening, cropping, Gaussian noise, and salt-and-pepper noise. However, the robustness to geometric attacks such as scaling attacks needs to be improved, which is a limitation of DARI-Mark.

**Table 2.** PSNR values of three methods.

| Methods | PSNR (dB) |
|---|---|
| Lee et al. [24] | 35.46 |
| ReDMark [25] | 35.93 |
| DARI-Mark | 37.38 |

**Table 3.** BER and NCC comparison results of three methods under multiple attacks.

| Attack | | Lee et al. [24] | | ReDMark [25] | | DARI-Mark | |
|---|---|---|---|---|---|---|---|
| | | BER (%) | NCC | BER (%) | NCC | BER (%) | NCC |
| JPEG compression | $Q = 50$ | 8.06 | 0.9055 | 3.03 | 0.9698 | **1.36** | **0.9832** |
| | $Q = 70$ | 4.24 | 0.9601 | 1.60 | 0.9799 | **0.00** | **1.0000** |
| | $Q = 90$ | 0.96 | 0.9988 | 0.00 | 1.0000 | **0.00** | **1.0000** |
| Sharpening | $R_S = 1$ | 0.98 | 0.9894 | 0.95 | 0.9892 | **0.00** | **1.0000** |
| | $R_S = 5$ | 1.72 | 0.9799 | 1.47 | 0.9823 | **0.00** | **1.0000** |
| | $R_S = 10$ | 2.01 | 0.9763 | 3.42 | 0.9599 | **0.00** | **1.0000** |
| | $R_S = 20$ | 3.51 | 0.9575 | 4.98 | 0.9457 | **0.19** | **0.9998** |
| Cropping | $c = 0.10$ | 4 | 0.954 | 2.14 | 0.974 | **0.68** | **0.9917** |
| | $c = 0.20$ | 5.46 | 0.9386 | 5.33 | 0.9321 | **3.81** | **0.9499** |
| | $c = 0.30$ | 17.10 | 0.8216 | 8.67 | 0.9008 | **4.05** | **0.9567** |
| Gaussian noise | $\sigma = 0.05$ | 5.99 | 0.9706 | 0.30 | 0.9943 | **0.09** | **0.9988** |
| | $\sigma = 0.15$ | 27.00 | 0.6802 | 5.08 | 0.9399 | **3.71** | **0.9563** |
| | $\sigma = 0.25$ | 38.17 | 0.5702 | 12.73 | 0.8571 | **10.30** | **0.9088** |
| Scaling | $r = 0.50$ | 10.98 | 0.8796 | **5.92** | **0.9711** | 14.47 | 0.8343 |
| | $r = 0.75$ | 3.36 | 0.9599 | **0.00** | **1.0000** | 5.56 | 0.9399 |
| | $r = 1.50$ | 1.53 | 0.9832 | **0.00** | **1.0000** | 0.48 | 0.9953 |
| Salt-and-pepper noise | $s = 0.04$ | 0.29 | 0.9965 | 1.45 | 0.9802 | **0.00** | **1.0000** |
| | $s = 0.06$ | 1.53 | 0.9832 | 1.17 | 0.9858 | **0.00** | **1.0000** |
| | $s = 0.10$ | 3.19 | 0.9621 | 3.51 | 0.9600 | **0.00** | **1.0000** |

## 5. Conclusions

In this paper, we proposed a robust image watermarking framework called DARI-Mark. Considering the masking effect of the HVS, an attention network was introduced in the end-to-end watermarking framework. Through the setting of the loss function, the attention network in the watermark embedding network helped the network automatically select texturally complex regions to embed the watermark for improving imperceptibility. During watermark extraction, the attention module assigns larger attention weights to features containing watermarks and smaller weights to noisy signals, increasing the robustness of watermark extraction. In addition, to avoid the phenomenon of gradient disappearance

or explosion when the network is deep, DARI-Mark added residual modules. Experiments showed that DARI-Mark had a good imperceptibility and robustness to multiple attacks. Compared with other state-of-the-art watermarking methods, DARI-Mark was more robust to JPEG compression, sharpening, cropping, Gaussian noise, and salt-and-pepper noise attacks.

However, the robustness of DARI-Mark against geometric attacks (such as scaling, rotation, etc.) was not ideal. In future research work, we will improve the framework to make it robust against more types of image distortion attacks.

**Author Contributions:** Y.Z. and C.W. conceived the algorithm and designed the experiments; Y.Z. performed the experiments; Y.Z., X.Z. and Z.Q. analyzed the results; Y.Z. drafted the manuscript; C.W., X.Z. and Z.Q. revised the manuscript. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported in part by the Shandong Provincial Natural Science Foundation (Nos. ZR2021MF060, ZR2017MF020), in part by the Joint Fund of Shandong Provincial Natural Science Foundation (No. ZR2021LZH003), in part by the National Natural Science Foundation of China (No. 61702303), in part by the Science and Technology Development Plan Project of Weihai Municipality (No. 2022DXGJ13), in part by the Scientific Research Project of Shandong University–Weihai Research Institute of Industry Technology (No. 0006202210020011), in part by the Education and Teaching Reform Research Project of Shandong University, Weihai (No. Y2021054), and in part by the 17th Student Research Training Program (SRTP) at Shandong University, Weihai (Nos. A22086, A22299).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Zheng, Q.; Liu, N.; Wang, F. An adaptive embedding strength watermarking algorithm based on shearlets' capture directional features. *Mathematics* **2020**, *8*, 1377.
2. Wu, H.Z.; Shi, Y.Q.; Wang, H.X.; Zhou, L.N. Separable reversible data hiding for encrypted palette images with color partitioning and flipping verification. *IEEE Trans. Circuits Syst. Video Technol.* **2017**, *27*, 1620–1631.
3. Mahto, D.K.; Singh, A. A survey of color image watermarking: State-of-the-art and research directions. *Comput. Electr. Eng.* **2021**, *93*, 107255.
4. Verma, V.S.; Jha, R.K. An overview of robust digital image watermarking. *IETE Tech. Rev.* **2015**, *32*, 479–496.
5. Faheem, Z.B.; Ali, M.; Raza, M.A.; Arslan, F.; Ali, J.; Masud, M.; Shorfuzzaman, M. Image watermarking scheme using LSB and image gradient. *Appl. Sci.* **2022**, *12*, 4202.
6. Li, Y.; Li, J.; Shao, C.; Bhatti, U.A.; Ma, J. Robust multi-watermarking algorithm for medical images using patchwork-DCT. In Proceedings of the 8th International Conference on Artificial Intelligence and Security, Xining, China, 15–20 July 2022; Lecture Notes in Computer Science; Springer: Berlin, Germany, 2018; Volume 13340, pp. 386–399.
7. Ernawan, F.; Ariatmanto, D.; Firdaus, A. An improved image watermarking by modifying selected DWT-DCT coefficients. *IEEE Access* **2021**, *9*, 45474–45485.
8. Wang, B.; Zhao, P. An adaptive image watermarking method combining SVD and Wang-Landau sampling in DWT domain. *Mathematics* **2020**, *8*, 691.
9. Cedillo-Hernandez, M.; Cedillo-Hernandez, A.; Garcia-Ugalde, F.J. Improving DFT-based image watermarking using particle swarm optimization algorithm. *Mathematics* **2021**, *9*, 1795.
10. Shekhar, H.; Seal, S.; Kedia, S.; Guha, A. Survey on applications of machine learning in the field of computer vision. In Proceedings of the 1st International Conference on Emerging Technology in Modelling and Graphics, Kolkata, India, 6–7 September 2020; Volume 937, pp. 667–678.
11. Rai, A.; Singh, H.V. SVM based robust watermarking for enhanced medical image security. *Multimed. Tools Appl.* **2017**, *76*, 18605–18618.
12. Singh, R.P.; Dabas, N.; Chaudhary, V.; Nagendra. Online sequential extreme learning machine for watermarking in DWT domain. *Neurocomputing* **2016**, *174*, 238–249.
13. Mehta, R.; Rajpal, N.; Vishwakarma, V.P. A robust and efficient image watermarking scheme based on Lagrangian SVR and lifting wavelet transform. *Int. J. Mach. Learn. Cybern.* **2017**, *8*, 379–395.

14. Abdelhakim, A.M.; Abdelhakim, M. A time-efficient optimization for robust image watermarking using machine learning. *Expert Syst. Appl.* **2018**, *100*, 197–210.
15. Kannan, D.; Gobi, M. An extensive research on robust digital image watermarking techniques: A review. *Int. J. Signal Imaging Syst. Eng.* **2015**, *8*, 89–104.
16. Hassanin, M.; Anwar, S.; Radwan, I.; Khan, F.S.; Mian, A. Visual attention methods in deep learning: An in-depth survey. *arXiv* **2022**, arXiv:2204.07756.
17. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
18. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the 15th European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; Lecture Notes in Computer Science; Springer: Berlin, Germany, 2018; Volume 11211, pp. 3–19.
19. Kandi, H.; Mishra, D.; Gorthi, S.R.S. Exploring the learning capabilities of convolutional neural networks for robust image watermarking. *Comput. Secur.* **2017**, *65*, 247–268.
20. Zhu, J.; Kaplan, R.; Johnson, J.; Fei, L.F. HiDDeN: Hiding data with deep networks. In Proceedings of the 15th European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; Lecture Notes in Computer Science; Springer: Berlin, Germany, 2018; Volume 11219, pp. 682–697.
21. Mun, S.M.; Nam, S.H.; Jang, H.; Kim, D.; Lee, H.K. Finding robust domain from attacks: A learning framework for blind watermarking. *Neurocomputing* **2019**, *337*, 191–202.
22. Liu, Y.; Guo, M.; Zhang, J.; Zhu, Y.; Xie, X. A novel two-stage separable deep learning framework for practical blind watermarking. In Proceedings of the ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 1509–1517.
23. Luo, X.; Zhan, R.; Chang, H.; Yang, F.; Milanfar, P. Distortion agnostic deep watermarking. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13545–13554.
24. Lee, J.E.; Seo, Y.H.; Kim, D.W. Convolutional neural network-based digital image watermarking adaptive to the resolution of image and watermark. *Appl. Sci.* **2020**, *10*, 6854.
25. Ahmadi, M.; Norouzi, A.; Karimi, N.; Samavi, S.; Emami, A. ReDMark: Framework for residual diffusion watermarking based on deep networks. *Expert Syst. Appl.* **2020**, *146*, 113157.
26. Jia, Z.; Fang, H.; Zhang, W. 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, Virtual Online, China, 20–24 October 2021; pp. 41–49.
27. Zhong, X.; Huang, P.C.; Mastorakis, S.; Shih, F.Y. An automated and robust image watermarking scheme based on deep neural networks. *IEEE Trans. Multimed.* **2021**, *23*, 1951–1961.
28. Sun, W.; Zhou, J.; Li, Y.; Cheung, M.; She, J. Robust high-capacity watermarking over online social network shared images. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 1208–1221.
29. Yang, Y.; Zou, T.; Huang, G.; Zhang, W. A high visual quality color image reversible data hiding scheme based on B-R-G embedding principle and CIEDE2000 assessment metric. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 1860–1874.
30. Yu, C. Attention based data hiding with generative adversarial networks. In Proceedings of the 34th AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 1120–1128.
31. Bhowmik, D.; Oakes, M.; Abhayaratne, C. Visual attention-based image watermarking. *IEEE Access* **2016**, *4*, 8002–8018.
32. Bas, P.; Filler, T.; Pevny, T. "Break our steganographic system": The ins and outs of organizing BOSS. In Proceedings of the 13th International Conference on Information Hiding, Prague, Czech Republic, 18–20 May 2011; Lecture Notes in Computer Science; Springer: Berlin, Germany, 2011; Volume 6958, pp. 59–70.
33. Krizhevsky, A. Learning Multiple Layers of Features from Tiny Images. Ph.D. Thesis, University of Toronto: Toronto, ON, Canada, 2009.
34. Fdez-Vidal, X.R. Dataset of Standard 512 × 512 Grayscale Test Images. 2019. Available online: https://decsai.ugr.es/cvg/CG/base.htm (accessed on 18 November 2020)