An Efficient Robust Blind Watermarking Method Based on Convolution Neural Networks in Wavelet Transform Domain

Nguyen Chi Sy, Ha Hoang Kha, and Nguyen Minh Hoang

Abstract—**Digital watermarking is one of the most widely used techniques for the protection of ownership rights of digital audio, images, and videos. One of the desirable properties of a digital watermarking scheme is its robustness against attacks aiming at removing or destroying the watermark from the host data. Different from the common watermarking techniques based on the spatial domain or transform domain, in this paper, a novel scheme of digital image blind watermarking based on the combination of the discrete wavelet transform (DWT) and the convolutional neural network (CNN) is proposed. Firstly, the host images are decomposed by the DWT with 4 levels and, then, the low frequency sub-bands of the first level and the high frequency sub-bands of the fourth level are used as the input data and the output target data to train the CNN model for embedding and extracting the watermark. Experimental results show that the proposed scheme has superior performance against common attacks of JPEG compression, mean and median filtering, salt and pepper noise, Gaussian noise, speckle noise, brightness modification, scaling, cropping, rotation, and shearing operations.**

Index Terms—**Robust image watermarking, discrete wavelet transform, convolutional neural network, copyright protection.**

I. Introduction

The explosive growth of the Internet and the social networks has provided the increasing convenience for the transmission and sharing digital multimedia applications such as audio, images and videos. With the development of the advanced multimedia signal processing technologies, digital multimedia can be easily and simply acquired, copied and tampered. Thus, the issues related to multimedia information protection, copyright and content authentication have been of significant concerns [1]. With digital image data, there are extensive studies on how to prevent unauthorized users from illegally copying, and distributing, modifying the digital images [1], [2]. The digital watermarking techniques which embed hidden information (known as a watermark) to a host media to detect and trace copyright violations have attracted considerable interest from academia and industry [3]. Digital image watermarking can be found in various practical applications of copyright protection, image authentication, medical applications, tamper detection, digital fingerprinting [4]. The most important and desirable properties in applications of watermarking for protecting the owners' copyright are invisibility and robustness. Invisibility measures the changes in the quality of host images before and after watermarking. Robustness measures the ability that the embedded watermarks cannot be destroyed and removed by the signal processing operations. In general, there is a trade-off between invisibility and robustness [5]. Based on the specific applications, a watermarking technique can be appropriately chosen to obtain the desired properties.

Based on the domain in which the watermark is embedded, digital image watermarking techniques can typically be divided into different categories such as the spatial domain [6], [7], transform domain [8]-[10] or hybrid ones [11]. In spatial domain watermarking schemes, the watermarks are inserted in the host images in the spatial domain by modifying the gray level values of chosen pixels in images [7]. Although the spatial domain watermarking is simple to implement, it can be sensitive to common attacks such as JPEG compression, low-pass filtering, and the watermarks can be easily de-attached by using inverse operations [6], [12], [13]. Therefore, spatial domain watermarking techniques are not commonly used in many practical applications. Alternatively, to improve the robustness and imperceptibility, watermarking can be carried out in transform domains such as the fast Fourier transform (FFT) [14], discrete cosine transform (DCT) [15], discrete wavelet transform (DWT) [13], [16], DWT-DCT [17], [18]. The transform domain watermarking can offer better robustness against common attacks since watermark coefficients are spread over the host image.

More recently, to further enhance the imperceptibility of watermarked images and robustness of watermarks, artificial intelligence (AI) based methods in digital image watermarking have attracted great interests, see, for examples [19]-[24] and references therein. In [19], the authors introduced the blind watermarking scheme exploiting the back-propagation (BP) neural network (NN) in the DWT domain. The authors demonstrated that their algorithm offers imperceptibility and robustness to common attacks such as salt and pepper noise, median filtering, rotation, cropping and JPEG compression. Similarly, the authors in [20] studied on a blind watermarking algorithm

Manuscript received September 3, 2019; revised June 21, 2020.

Nguyen Chi Sy and Ha Hoang Kha are with the Faculty of Electrical & Electronics Engineering, Ho Chi Minh City University of Technology, VNU-HCM, Vietnam (e-mail: chisy.nguyen@gmail.com, hhkha@hcmut.edu.vn).

Nguyen Minh Hoang is with the Faculty of Management Information Systems, Banking University of HCM City, Vietnam (e-mail: nmhoang@gmail.com).

doi: 10.18178/ijmlc.2020.10.5.990

using a feed-forward NN in the DWT domain. Their simulation results revealed the good performance of imperceptibility and robustness against common attacks as in [19]. The authors of [22] developed a watermarking technique by using the combination of fractal dimension, BP NN, Arnold transform, and multiwavelet transform to improve the security, imperceptibility and robustness. Likewise, references in [23], [24] described the watermarking scheme using DWT and BP NN with good performance in terms of invisibility and imperceptibility. Alternatively, the study in [21] introduced a non-blind watermarking scheme based on a learning-based auto-encoder convolutional neural network (CNN). The experimental results in [21] indicated that their CNN based watermarking outperforms the previously existing methods in terms of peak signal-to-noise ratio (PSNR) and normalized correlation (NC).

Motivated from the above studies, in this paper, we develop a blind watermarking scheme by using the CNN in the wavelet transform domain. In our watermarking algorithm, we decompose the host images into sub-bands by using the DWT. Then, the selected low frequency sub-bands are used as the inputs for the CNN while the high frequency sub-bands are used as the target outputs. We carry out the training for the CNN by using the set of images in the standard database. The trained CNN is employed for both embedding and extracting the watermarks. Different from the method in [23] which used the DWT decomposition with 2 levels, and employed the BP NN for watermark embedding and extraction, our proposed method invokes the DWT decomposition with 4 levels and applied the CNN for watermark embedding and extraction. To evaluate the performance of our scheme, we carry out extensive numerical experiments on various images and different attacks of non-geometric and geometric transforms. The numerical results demonstrate that our watermarking scheme offers superior performance in terms of robustness and invisibility as compared with the other existing methods. The main contributions of the paper can be summarized as follows:

- We introduce the novel watermarking scheme by appropriately selecting the sub-bands in the DWT domain as the inputs and target outputs for training the CNN. Then, the trained CNN is used to embed and extract the watermarks.
- We conducted extensive experiments to demonstrate the superiority of the proposed scheme over other typical existing methods in terms of invisibility and robustness in the various attacks.

The rest of this paper is organized as follows. In Section II, we introduce the proposed watermarking scheme along with the details on the DWT, CNN. Then, the effectiveness of the proposed watermarking scheme which is evaluated by experimental results for various attacks is presented in Section III. Finally, the concluding remarks are given in Section IV.

II. Proposed Watermarking Scheme

In this work, we propose a blind digital watermarking scheme for copyright protection based on the DWT and CNN. The host image is transformed into the DWT domain and, then, the image in the transform domain is used for training the CNN to embed and extract the watermark efficiently.

A. Discrete Wavelet Transform (DWT)

The DWT is a frequency domain method which is commonly used in image processing. The DWT is a means to effectively present the image into a multi-scale analysis with the lower computational cost. This method in image processing includes the decomposition of images into frequency channels of constant bandwidth [25]. The DWT has been applied in various image processing applications, for examples, noise reduction and image compression. In the DWT, a two-dimensional image is expressed in the DWT domain by applying the low-pass and high-pass filters to the rows and columns of the image and, then, the transformed image at level 1 is partitioned into four sub-bands, namely LL, LH, HL and HH where the first letter denotes the low (L) or high (H) pass filtering operations to the rows while the second letter refers to the filtering operations to the columns [25], [26]. The division of each sub-band can be repeatedly carried out until the required number of levels is obtained. The forward DWT is defined by [27]

$$\mathrm{W}_{\psi}[j,k]=\frac{1}{\sqrt{M}}\sum_{n=0}^{M-1}f[n]\psi_{j,k}[n] \tag{1}$$

$$\mathrm{W}_{\varphi}[j_o,k]=\frac{1}{\sqrt{M}}\sum_{n=0}^{M-1}f[\mathrm{n}]\varphi_{j_o,k}[n] \tag{2}$$

where $f[n]$ is the input signal (i.e., the pixel row or column of the image), $j_o < j < J$, $M = 2^J$, $\psi_{j,k}[n]$ is the wavelet function and $\varphi_{j_o,k}[n]$ is the scale function; $j = 0,1,2,...J-1$, $k = 0,1,2,...2^J-1$. The constant J is the maximum scale. On the other hand, we commonly fix $j_o = 0$ for implementation. The scaling function and the wavelet function are respectively given by

$$\psi_{j,k}[n]=2^{\frac{-j}{2}}\psi_{j,k}[2^{-j}n-k] \tag{3}$$

$$\varphi_{j,k}[n]=2^{\frac{-j}{2}}\varphi_{j,k}[2^{-j}n-k]. \tag{4}$$

The inverse DWT can be computed by

$$f[n]=\frac{1}{\sqrt{M}}\sum_{j=j_o}^{J-1}\sum_{k=0}^{M-1}\mathrm{W}_{\psi}[j,k]\psi_{j,k}[\mathrm{n}]+ \frac{1}{\sqrt{M}}\sum_{k=0}^{J-1}\mathrm{W}_{\varphi}[j_o,k]\varphi_{j_o,k}[\mathrm{n}]. \tag{5}$$

A two-dimensional image after four levels of the DWT decomposition is demonstrated in Fig. 1. Then, we select the high frequency sub-bands to embed the watermark since the human visual system is more sensitive to the LL sub-band which represents the low frequency component [28].

B. Convolutional Neural Network (CNN)

CNNs consist of one or more convolutional layers with pooling/subsampling steps followed by one or more fully connected layers (a multilayer neural network) as demonstrated in Fig. 2 [29]. An image is feed to the network as an input which goes through multiple convolutions, pooling layers and finally a fully connected layer to produce the outputs. CNNs have proven to be efficient in various applications such as image classification, object detection, and recognition since they are able to efficiently learn and represent the image by the limited number of parameters (kernels, biases, weights). In CNNs, the back propagation (BP) algorithm using the standard gradient descent method is commonly invoked for updating kernels, weights and biases in the layers [30], [31]. The following will present the BP algorithm in fully connected networks and, then, BP updated for convolutional and sup-sampling layers in a 2D CNN which will be applied for image watermarking.

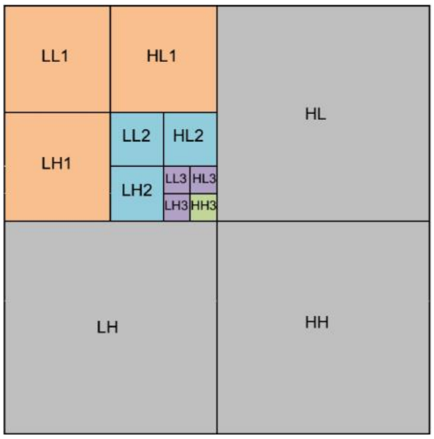


Fig. 1. The resultant image after four levels of the DWT decomposition.

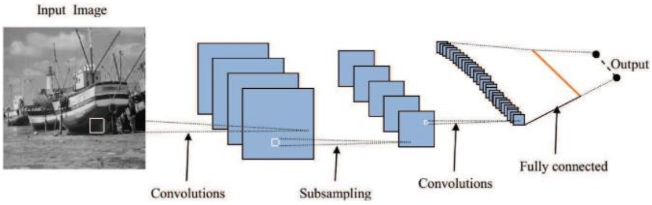


Fig. 2. Convolutional neural network architecture.

Feedforward pass: Let W^t and b^t be a weight matrix and vector associated to layer t. Then, the output of layer t can be written as

$$o^t = f(v^t), v^t = W^t v^{t-1} + b^t \qquad (6)$$

where $f(.)$ is an activation function. In our experiments, we use tanh, mean and sigmoid functions. Denoting the value of output k for training sample n by y_k^n and the corresponding target output (desired output) by d_k^n, the error for the individual sample n is computed by

$$E^n = \frac{1}{2}\sum_{k=1}^{c}(d_k^n - y_k^n)^2 = \frac{1}{2}\left\|d^n - y^n\right\|_2^2 \qquad (7)$$

where c is the number of the outputs.

Back-propagation: Denoting the bias sensitivity at layer t as δ^t, the sensitivity of layer t in the BP process is defined by

$$\delta^t = (W^{t+1})^T \delta^{t+1} \circ f'(v^t) \qquad (8)$$

where $\circ$ stands for element-wise multiplication and the sensitivities for the output layer $t = L$ is

$$\delta^L = f'(v^L) \circ (y^n - d^n). \qquad (9)$$

Accordingly, the partial derivative of each weight at layer t which is an outer product between the input and the sensitivity is given by

$$\frac{\partial E}{\partial W^t} = o^{t-1}(\partial^t)^T. \qquad (10)$$

Then, the neuron weights in layer t are updated by

$$\Delta W^t = -\eta \frac{\partial E}{\partial W^t} \qquad (11)$$

where η is a learning rate parameter [30].

Convolutional layer: One of the major building blocks of a CNN is the convolutional layer which performs a convolution operation on the input image with filters (kernels) to produce a feature map. By using numerous filters for convolutions, one can obtain various feature maps at the output of the convolution layer. Since the convolutional layer is followed by a subsampling layer, to calculate the sensitivities at layer t, we need up-sample on the sensitivity map of the sub-sampling layer. In sub-sampling layer, for each map j in the convolutional layer, gradient formula is given by

$$\partial_j^t = \beta_j^{t+1}(f'(v_j^t) \circ up(\partial_j^{t+1})) \qquad (12)$$

where $up(.)$ is an up-sampling operation and β is a multiplicative bias parameter.

Subsampling layer: A subsampling or pooling layer often follows a convolutional layer to down-sample the output of the convolutional layer. The major function of a subsampling layer is to reduce the number of parameters to be learned by CNNs. The outputs of the subsampling layer are the down-sampled version of the input maps. By denoting $down(.)$ as a subsampling operation, we have

$$o_j^t = f(\beta_j^t down(o_j^{t-1}) + b_j^t) \qquad (13)$$

where b is an additive bias. Assume that the sub-sampling is followed by a full connection network, the BP algorithm can be used to obtain the sensitivity of the sub-sampling layer. First, we can calculate the gradient of convolution kernel by

$$\partial_j^t = f'(v_j^t) \circ conv2(\partial_j^{t+1}, rot180^o(K_j^{t+1}),' full') \qquad (14)$$

Accordingly, the gradient for the additive bias is given by

$$\frac{\partial E}{\partial b_j} = \sum_{uv}(\partial_j^t)_{uv} \qquad (15)$$

while the gradient for multiplicative bias β is computed by

$$\frac{\partial E}{\partial \beta_j} = \sum_{uv}(\partial_j^t \circ p_j^t)_{uv} \qquad (16)$$

where $p_j^l = down(o_j^{t-1})$.

C. Watermark Embedding Algorithm

Our proposed scheme based on the DWT and CNN is described as follows. The host gray images with size 512x 512 pixels are decomposed by using DWT2 to obtain 4 sub-bands LL, HL, LH, HH; then the sub-band LL is taken to analyze by using DWT2 to create 4 other sub-bands LL1, HL1, LH1, HH1; sub-band HH1 are decomposed by the DWT2 to obtain LL2, HL2, LH2, HH2; and then sub-band HH2 are decomposed again to obtain 4 sub-bands LL3, HL3, LH3, HH3 at the end. This resultant image after DWT is demonstrated in Fig. 1. Accordingly, the sub-band (256 × 256) at the low frequency LL is divided into non-overlapped blocks, namely $I(x, y)$ $(1 \le x, y \le 8)$ with the size of 8 × 8 pixels (i.e., there are 32 × 32 blocks of 8 × 8 pixels). These blocks are used as data inputs of the proposed CNN model. Each block is assigned a corresponding pixel namely $O(i, j)$ $(1 \le i, j \le 32)$ of sub-band HH3 as detailed in Fig. 3. This pixel is an expected output data of the CNN model. The sets of $I(x, y)$ and $O(i, j)$ are used to train the CNN model including an input layer with 64 nodes; a convolutional layer with the size of feature maps 15, size of kernels [3 3], activation function 'tanh'; a pool layer with sub sampling factor 3, sub sampling type 'mean'; a fully connected layer with 150 nodes, activation function 'tanh'; a fully connected layer with 1 node, activation function 'sigm'. After training, we can estimate the output values denoted as $O'(i, j)$ according to the new input data. The pixel in sub-band HH3 $O(i, j)$ $(1 \le i, j \le 32)$ is embedded information of binary image watermark $W(i, j)$ of 32 × 32. If $W(i, j)$=0, the watermarked pixel is $O(i, j)$=min($O(i, j)$, $O'(i, j)$- α), otherwise if $W(i, j)$==1, the watermarked pixel $O(i, j)$=max($O(i, j)$, $O'(i, j)$+ α) where α is a system parameter and can be determined by the requirements of the users [22]. The step by step procedure for embedding a binary watermark is described in Algorithm 1.

Algorithm 1: The proposed scheme for embedding a watermark

```
Input: trained CNN, host image, watermark W.
Host image is decomposed by using DWT2 to obtain the set of $I(x, y)$ and the set of $O(i, j)$. The set of I(x, y) is put into the trained CNN to obtain the set of $O'(i, j)$.
The watermark is embedded by using the following loops:
for $i$=1 to 32 do
    for $j$=1 to 32 do
        if $W(i, j)$ == 0 then
            $O(i, j)$=min($O(i, j)$,$O'(i, j)$- $\alpha$ )
        else
            $O(i, j)$=max($O(i, j)$,$O'(i, j)$+ $\alpha$ )
        end if
    end for
end for
Use the inverse DWT2 to obtain the watermarked image.
Output: watermarked image.
```

D. Watermark Extracting Algorithm

Watermark recovery is almost the inverse process of watermark embedding except that the training procedure and inverse transform are not required. The watermarked image is decomposed by using DWT2 to produce 4 sub-bands LL, HL, LH, HH; then sub-band LL is decomposed by using DWT2 to create 4 sub-band LL1, HL1, LH1, HH1; next sub-band HH1 is decomposed to obtain sub-bands LL2, HL2, LH2, HH2; and finally sub-band HH2 are decomposed by DWT2 to have 4 sub-bands LL3, HL3, LH3, HH3. Sub-band HH3 is selected as $O(i, j)$. The sub-band (256 × 256) at the low frequency LL is divided into blocks namely $I(x, y)$ $(1 \le x, y \le 8)$ with size 8 × 8 non-overlap blocks (i.e., 32 × 32 blocks). These blocks are put into the trained proposed CNN model to create outputs namely $O'(i, j)$; Then, the extracted watermark is obtained by [22]

$$W'(i, j) = \begin{cases} 1 & O(i, j) \ge O'(i, j) \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

The detailed procedure for extracting the watermark is given in Algorithm 2.

Algorithm 2: The proposed scheme for extracting the watermark

```
Input: trained CNN, watermarked image.
Watermarked image is decomposed by using DWT2 to obtain the set of $I(x, y)$ and the set of O($i, j$). The set of $I(x, y)$ is put into the trained CNN to obtain the set of O'($i, j$).
The watermark is extracted by using the following loops:
for $i$=1 to 32 do
    for $j$=1 to 32 do
        if $O(i, j) \ge O'(i, j)$ then
            $W'(i, j) = 1$
        else
            $W'(i, j) = 0$
        end if
    end for
end for
Output: extracted watermark
```

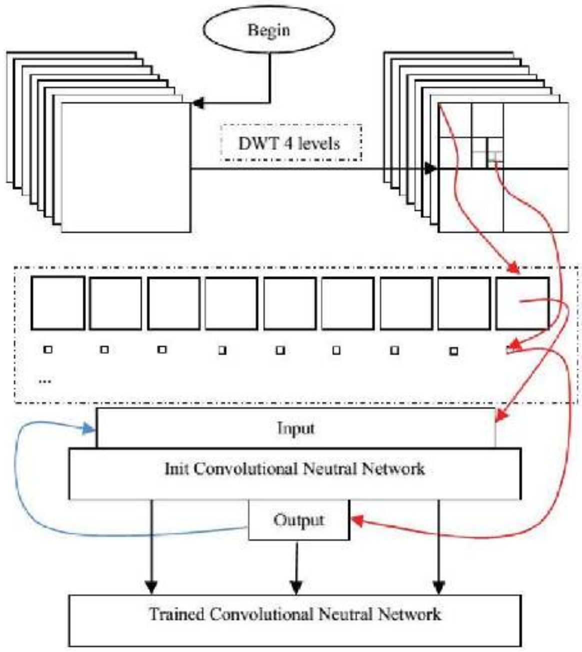

Begin
DWT 4 levels
...
Input
Init Convolutional Neutral Network
Output
Trained Convolutional Neutral Network

Fig. 3. Flowchart of preparing the data for training the proposed CNN.

E. Performance Metrics for Evaluating Watermarking Schemes

To evaluate the performance of our watermarking scheme, we use performance metrics of the peak signal-to-noise ratio (PSNR), normalized correlation (NC) and structural similarity (SSIM). Let A and B be the original and modified images (e.g., host image and watermarked image, watermark image and extracted watermark image), respectively. Assume that the size of the image is M × N and (*i*, *j*) is the pixel at row i and column j of the image. To measure the quality of the watermarked image, we use the PSNR defined as

$$PSNR = 10\log_{10}(\frac{255^2}{MSE})(dB) \tag{18}$$

where

$$MSE = \frac{1}{MxN}\sum_{i=1}^{M}\sum_{j=1}^{N}(A(i,j)-B(i,j))^2. \tag{19}$$

It is obvious that the higher PSNR, the higher quality of the watermarked image. Alternatively, we can use other distortion measures such as NC and SSIM to measure the similarity of two images. The NC is defined as [3]

$$NC = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}A(i,j).B(i,j)}{\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}A(i,j)^2}\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}B(i,j)^2}}. \tag{20}$$

To compute the SSIM metric for two images of the same size, three components, namely luminance, contrast and structure are used. At each step, the local windows *x* and *y* of two images are used to find the local statistics and SSIM index. The range of SSIM is from 0 to 1, where 1 indicates that two images are identical. The local SSIM measure is the product of three components: luminance, contrast and structure given by [32]

$$SSIM = \frac{(2.\gamma_x.\gamma_y + C_1)(2.\rho_{xy} + C_2)}{(\gamma_x{}^2 + \gamma_y{}^2 + C_1)(\rho_x{}^2 + \rho_y{}^2 + C_2)} \tag{21}$$

where γ_x, ρ_x and γ_y, ρ_y are weighted means and weighted variances of *x* and *y*, respectively; ρ_{xy} is weighted crosscovariance between *x* and *y*; *C1* and *C2* are constants [32].

III. Experiment Results

Our experiments are carried out to test the imperceptibility of the hidden watermark as well as the robustness of the proposed scheme against attacks. Additionally, the performance of our method is compared with those of the ones in [18] and [23]. Both experiments are conducted on the host grey images with the size of 512 × 512 pixels, 8 bits/pixel and the binary watermark with the size of 32 × 32 pixels. The images in the experiments are taken from standard image database at http://sipi.usc.edu/database of the University of Southern California. The same initial structure of CNN is used in two following experiments including an input layer with 64 nodes; a convolutional layer with the size of feature maps 15, size of kernels 3x3 and activation function 'tanh'; an average pooling layer with sub sampling factor 3, a fully connected layer with 150 nodes with the activation function 'tanh'; a fully connected layer with 1 node, activation function 'sigmoid'. The 8 × 8 pixel blocks of the images in the transform domain are used for training the CNN. The trained CNN is used to embed and extract the watermark. All of the experiments are conducted in MATLAB R2013a on an Intel Core i5-2450M CPU @ 2.50GHz personal computer with 4 GB (RAM).

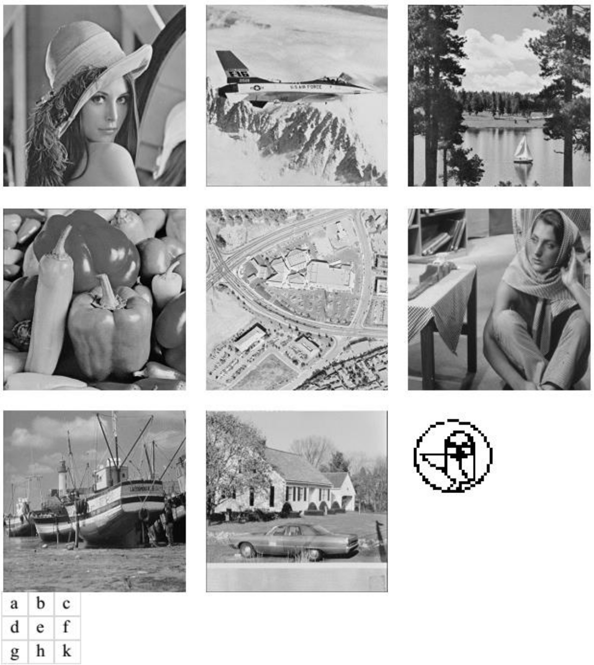


Fig. 4. (a) Host image Lena, (b) Host image Airplane, (c) Host image Sailboat on lake, (d) Host image Peppers, (e) Host image Aerial, (f) Host image Barbara, (g) Host image Boat, (h) Host image House, (k) Original watermark.

A. Experiment 1: The Invisibility and Robustness Performance of the Proposed Scheme

The aim of this experiment is to evaluate the performance in terms of invisibility and the ability to extract exactly the watermark by training the CNN model for 8 different grayscale images in the conditions with and without attacks. 8 grayscale images of 512 × 512 pixels are shown in Fig. 4(a)–Fig. 4(h) while the 32 × 32 binary watermark is presented in Fig. 4(k). We decompose each host image into sub-bands by using DWT. Using the analysis scheme, for each sub-band LL and sub-band HH3 of an image we divide sub-band LL (size 256 × 256 pixels) into 8 × 8 non-overlapped blocks and assign each block with the corresponding pixel in sub-band HH3. The pairs of 8 × 8 blocks in sub-band LL and the pixel in sub-band HH3 are used as the inputs and the desired output to train the CNN.

Results of the performance in terms of invisibility and robustness for embedding watermark of the proposed scheme based on the CNN in DWT domain in conditions of no attacks are shown in Fig. 5 and Table I. Fig. 5 illustrates the watermarked images and their corresponding PSNRs between the original and watermarked versions. As can been seen, the PSNR values of all 8 images are greater than 42 dB. Note that the PSNR value greater than 30 dB for a processed image is acceptable to human eyes [33]. Thus, the proposed watermarking scheme offers high invisibility. Table I provides the performance metrics to measure the invisibility of the watermarked images and the robustness of the extracted watermarks. As observed from the results in Table I, our algorithm achieves good performance in terms of PSNR, SSIM. With such high PSNR values, no visual artifacts can be noticed in the watermarked images. Additionally, the watermarks can be extracted from these 8 watermarked images with NC = 1.

TABLE I: INVISIBILITY AND ROBUSTNESS OF THE PROPOSED SCHEME IN THE CONDITION OF NO ATTACKS

Host images	Invisibility		Robustness	
	PSNR(dB)	SSIM	NC	SSIM
Lena	43.2954	0.9847	1	1
Airplane	42.9516	0.9839	1	1
Sailboat on lake	42.3986	0.9876	1	1
Peppers	42.9849	0.9846	1	1
Aerial	41.5314	0.9912	1	1
Barbara	42.1113	0.9883	1	1
Boat	43.2814	0.9874	1	1
House	42.8983	0.9885	1	1

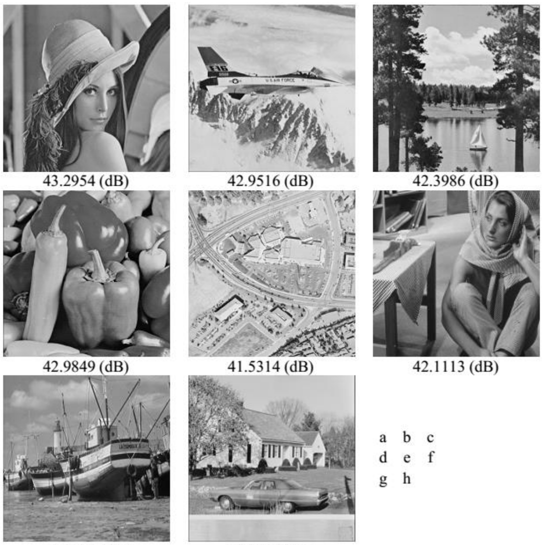


Fig. 5. The PNSR and visual performance of the watermarked images in conditions of no attack of the proposed algorithm.

Next, we investigate the performance of our watermarking scheme under various kinds of attacks including both geometric and non-geometric attacks. For non-geometric attacks, we consider the following operations: mean filtering; median; noise: salt & pepper, Gaussian, speckle; JPEG compression; brightness increase/decrease. First, we investigate the impacts of noise on the invisibility of the watermarked images and the robustness of the watermark. Fig. 6 shows the watermarked images attacked by Gaussian noise with zero mean and variance 0.01, salt & pepper noise with density 0.01, and speckle noise with variance 0.01. The PSNR and SSIM of the watermarked image after being attacked by noise are given in Table II. The reduction of PSNR and SSIM of the watermarked image of Lena as compared to the results without attack in Table I reveals the impacts of noise attacks. However, the NC and SSIM values of the extracted watermark in Table II are acceptable, which demonstrates the robustness of the proposed watermarking scheme against noise attacks.

TABLE II: SSIM & PSNR VALUES OF THE WATERMARKED IMAGE FROM DIFFERENT ATTACKS ON WATERMARKED IMAGE LENA

Noise	Invisibility		Robustness	
attacks	PSNR(dB)	SSIM	NC	SSIM
Gaussian	20.0449	0.5888	0.8746	0.9950
Salt & Peppers	25.2962	0.8216	0.9598	0.9985
Speckle	25.5902	0.7993	0.9643	0.9989

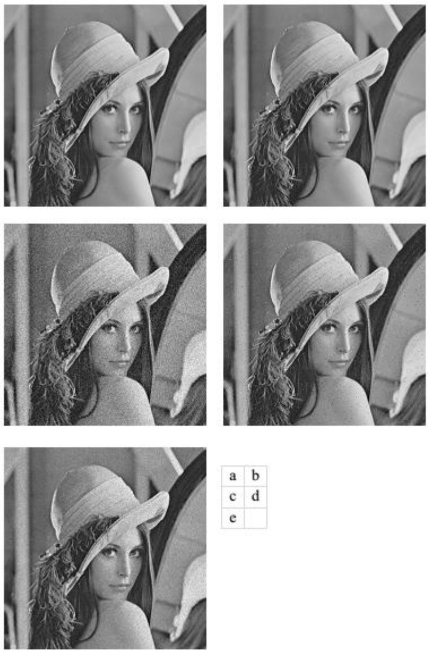


Fig. 6. (a) Host image Lena, (b) Watermarked image Lena, (c) Gaussian noise, (d) Salt& Pepper noise, (e) Speckle noise.

TABLE III: THE ROBUSTNESS OF THE PROPOSED SCHEME AGAINST JPEG ATTACKS

Host images	*Q*=80		*Q*=70		*Q*=60		*Q*=40		*Q*=30		*Q*=20	
	NC	SSIM	NC	SSIM	NC	SSIM	NC	SSIM	NC	SSIM	NC	SSIM
Lena	1	1	0.9976	0.9999	0.9957	0.9998	0.9289	0.9982	0.7953	0.9920	0.5585	0.9732
Airplane	0.9994	1	0.9976	0.9999	0.9945	0.9997	0.8929	0.9971	0.8022	0.9916	0.6096	0.9715
Sailboat on lake	1	1	0.9957	0.9999	0.9853	0.9995	0.9489	0.9975	0.8884	0.9941	0.7067	0.9764
Peppers	1	1	0.9976	0.9999	0.9976	0.9999	0.9534	0.9985	0.8471	0.9909	0.5723	0.9612
Aerial	1	1	0.9951	0.9998	0.9841	0.9993	0.9680	0.9988	0.9413	0.9982	0.8504	0.9939
Barbara	1	1	0.9982	0.9999	0.9957	0.9999	0.9541	0.9979	0.8786	0.9924	0.6999	0.9778
Boat	1	1	0.9970	0.9999	0.9976	0.9999	0.9617	0.9985	0.8815	0.9952	0.6997	0.9793
House	1	1	0.9988	0.9999	0.9970	0.9998	0.8923	0.9975	0.8239	0.9950	0.6653	0.9837

Next, we study the effects of the JPEG compression to the watermarked images and extracted watermark. JPEG is an image compression standard which is widely used. The compression ratio of the JPEG is associated with the quality factor between 0 and 100. When the quality factor is decreased, the image compression is improved, but the quality of the resulting image is significantly reduced. To evaluate the robustness of the watermark, the results of NC and SSIM for watermarks are shown in Table III for the different quality factors of JPEG compression on the watermarked images. As can be seen from Table III, the proposed method can perfectly extract the watermarks (with NC and SSIM about 1) while the quality factors are greater than 80. An average NC of 8 extracted watermarks is still equal to 0.6703 for the quality factor of 20. Furthermore, Fig. 7 indicates the NC of the extracted watermark on the images attacked by JPEG compression. When the quality factors are greater than 30, the NCs are greater than 0.8. The results in Fig. 7 show that the robustness of the proposed scheme is assured. For visual observation on the extracted watermarks, Fig. 8 shows the extracted watermarks in the conditions of being attacked by JPEG with different quality factors.

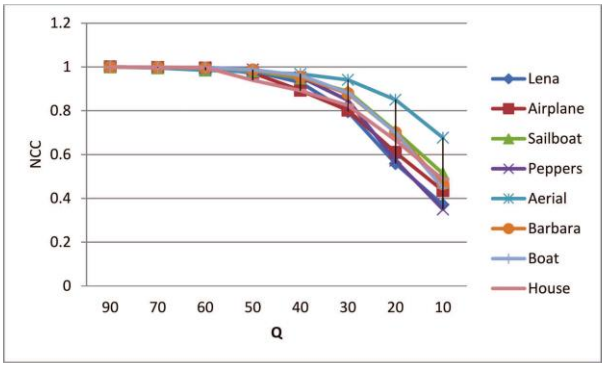
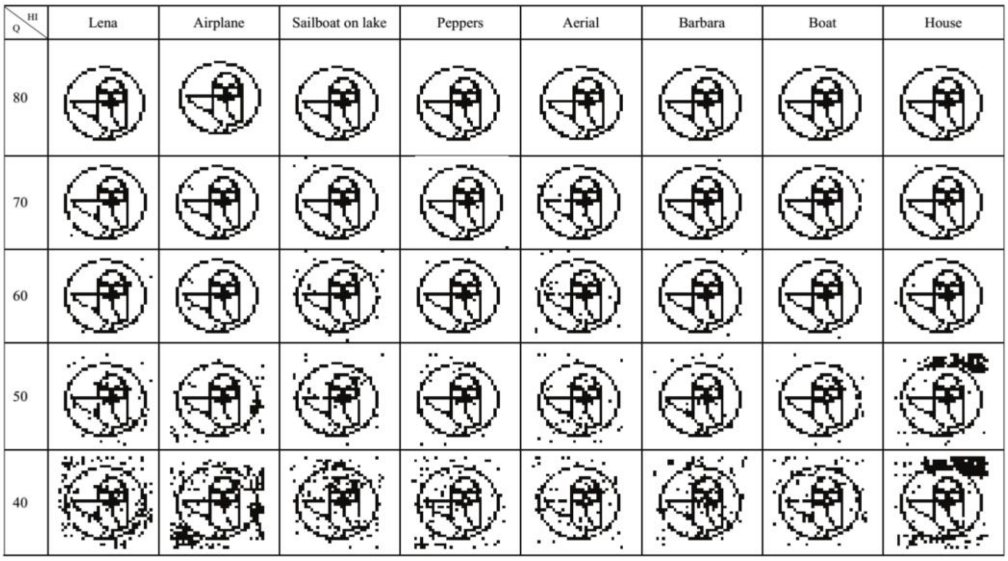


Fig. 7. The NC values of JPEG attacks with different quality factors on 8 watermarked images.



Q: Quality
HI: Host Image

Fig. 8. The extracted watermarks of the proposed scheme against JPEG attacks.

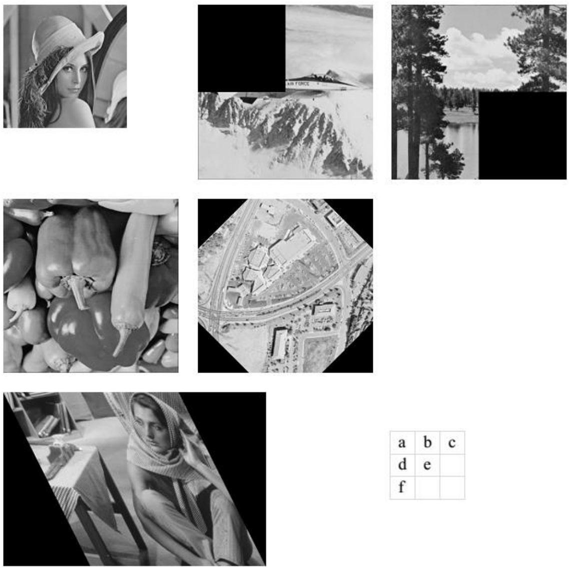


Fig. 9. The watermarked image after being attacked, a) scale 70%, b) Crop ¼ Top Left, c) Crop ¼ Bottom Right, d) Rotation 180°, e) Rotation 45°, f) Shear.

Next, we investigate the impacts of the median filtering, mean filtering and brightness varying operations on the watermarked images. Note that the filtering operations are often used in image processing and, thus, they can make the variations of the watermarks. In our experiments, a 3 × 3-pixel median filter, a 3 × 3 pixel mean filter, and increase/decrease 20% brightness are applied on the watermarked images. The NC and SSIM performance for each case of attacks is given in Table IV. As can be seen from Table IV, the watermark can be extracted perfectly under the changes of brightness 20%. On the other hand, almost all the NC values in Table IV are greater than 0.9. This result proves that the robustness of the proposed scheme is assured against the mean filtering, median filtering, brightness variation attacks.

TABLE IV: ROBUSTNESS OF PROPOSED SCHEMES IN CONDITIONS OF BEING ATTACKED BY MEAN ATTACK [3 3], MEDIAN ATTACK [3 3] AND INCREASE/ DECREASE BRIGHTNESS 20%

Host images	Mean attack		Median attack		Inc/Decr	
	NC	SSIM	NC	SSIM	NC	SSIM
Lena	0.9485	0.9969	0.9674	0.9981	1	1
Airplane	0.9510	0.9968	0.9687	0.9980	1	1
Sailboat on lake	0.9009	0.9959	0.9193	0.9969	1	1
Peppers	0.9453	0.9983	0.9579	0.9988	1	1
Aerial	0.8634	0.9937	0.9043	0.9957	1	1
Barbara	0.9318	0.9979	0.9306	0.9969	1	1
Boat	0.9397	0.9979	0.9491	0.9981	1	1
House	0.9269	0.9959	0.9504	0.9973	1	1

Now, we consider the geometric attacks as such as rotation, scaling, cropping and shearing operations. In the experiments we conduct the geometric attacks as follows: for scaling attacks, the watermarked images are scaled down by 70%, for cropping attacks, ¼ sizes of watermark images are cropped at the top left or the bottom right, for rotation attacks, the watermarked images are rotated by 180^0 or 45^0. The watermarked images which are affected by these geometric attacks are illustrated in Fig. 9. We use the proposed CNN watermarking scheme for the attacked images to extract the watermark. The extracted watermarks are shown in Fig. 10 and the NC and SSIM performance metrics are given in Table V. From Fig. 10, we can see that watermarks can be fully recovered for all attacks in the experiments however the watermarks are transformed by the same geometric operations.

B. Experiment: Performance Comparison between the Proposed Scheme with the Previous Methods

To evaluate the performance of our proposed watermarking scheme, we compare the PSNR and NC performance of our method with the ones in [18] and [23]. It should be noted that the study in [18] is a typical method of watermarking in the transform domain of the DWT-DCT combination while the approach in [23] is to use DWT and BP NNs.

To compare the performance of the proposed scheme with one in [18], we use 8 gray scale images shown in Fig. 4 as the inputs and the same watermark in [18] as shown Fig. 11(a). First, the invisibility and the robustness of the proposed scheme in the condition of no attacks are determined. Without attacks, the PSNR in our method is measured to be 43.2999 dB as compared to 42.6950 dB in [18]. The NC of the extracted watermark is equal to 1 for both schemes. To compare the robustness of our scheme with the one in [18], the JPEG compression with quality=50, salt & pepper noise with density=0.001 and, Gaussian noise (mean=0, variance=0,002) are used. Fig. 11 shows the original watermark and extracted watermarks with and without attacks. All of the extracted watermarks are clearly observed. Table VI shows that the robustness under salt& pepper noise, Gaussian noise of our algorithm are superior to those in [18].

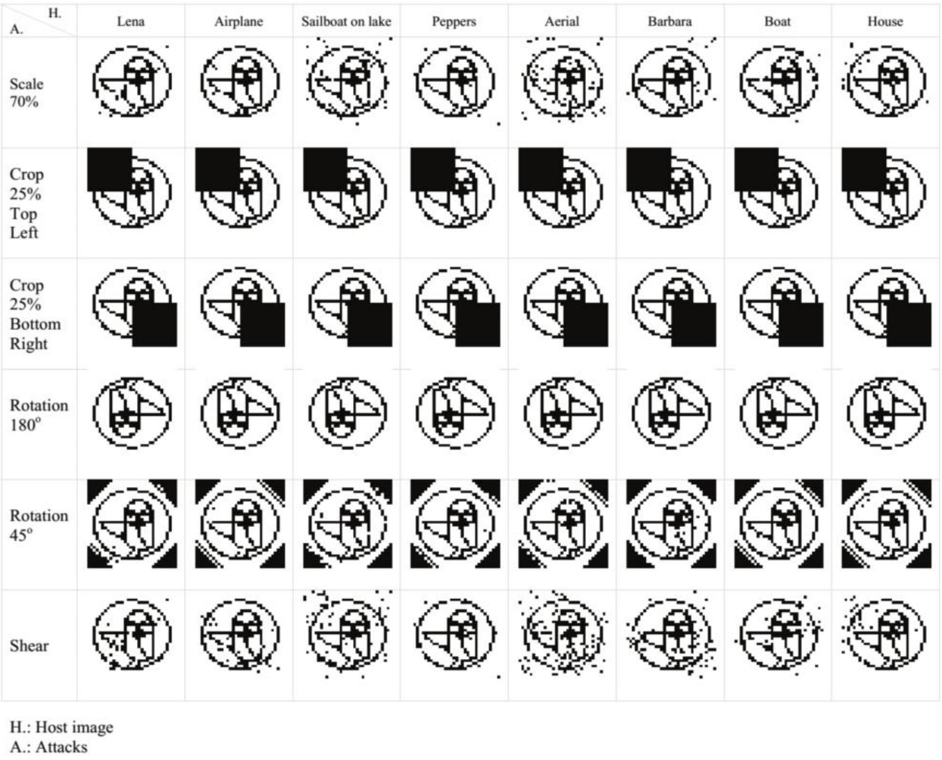


Fig. 10. The extracted watermarks in the condition of geometric attacks.

TABLE V: INVISIBILITY AND ROBUSTNESS OF THE PROPOSED SCHEME IN THE CONDITION OF GEOMETRIC ATTACKS

Host images	Scale 70%		Crop ¼ Top Left		Crop ¼ Bottom Right		Rotation 180		Rotation 45		Shear	
	NC	SSIM	NC	SSIM	NC	SSIM	NC	SSIM	NC	SSIM	NC	SSIM
Lena	0.9915	0.9996	0.8604	0.9796	0.8688	0.9842	0.8811	0.9887	0.8751	0.9996	0.9872	0.9992
Airplane	0.9933	0.9996	0.8604	0.9796	0.8688	0.9842	0.8811	0.9887	0.8806	0.9996	0.9823	0.9990
Sailboat on lake	0.9754	0.9990	0.8604	0.9796	0.8688	0.9842	0.8811	0.9887	0.8744	0.9995	0.9667	0.9988
Peppers	0.9945	0.9998	0.8604	0.9796	0.8688	0.9842	0.8811	0.9887	0.8779	0.9997	0.9915	0.9997
Aerial	0.9637	0.9986	0.8604	0.9796	0.8688	0.9842	0.8811	0.9887	0.8744	0.9995	0.9453	0.9977
Barbara	0.9896	0.9997	0.8604	0.9796	0.8688	0.9842	0.8811	0.9887	0.8695	0.9994	0.9661	0.9988
Boat	0.9896	0.9996	0.8604	0.9796	0.8688	0.9842	0.8811	0.9887	0.8792	0.9997	0.9817	0.9993
House	0.9915	0.9997	0.8604	0.9796	0.8688	0.9842	0.8811	0.9887	0.8799	0.9996	0.9804	0.9992

In order to compare the performance of the proposed scheme with those of the algorithm in [23], we use the test image “Mandrill” (512 × 512 pixels), and 1024 bits of a binary pattern watermark 32 × 32 as shown in Fig. 12(a) and (c). Fig. 12 shows the host image ”Mandrill”, watermarked image and extracted watermarks by our scheme for different attacks. The PSNR of the watermarked images (or attacked images) and NC of the extracted watermark of our proposed schemes and approach in [23] are listed in TABLE VII. With the comparable PSNR, our scheme offers the higher NC for all considered attacks. This means that our scheme outperforms the approach in [23] in terms of robustness against the attacks.

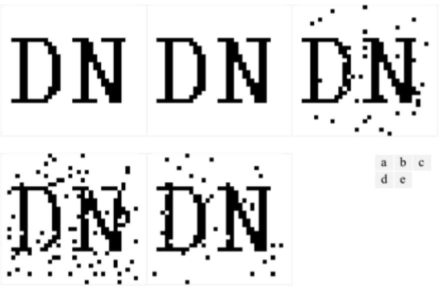


Fig. 11: The extracted watermarks of the proposed scheme with the same watermark as in [18], a) original watermark, b) no attack, c) JPEG Q=50, d) Salt & Pepper (density=0.001), e) Gaussian Noise (0,002).

TABLE VI: THE PERFORMANCE COMPARISON BETWEEN [18] AND THE PROPOSED SCHEME ON LENA IMAGE WITH ATTACKS

Attacks	NC	
	Ref. [36]	Proposed scheme
JPEG(Q=50)	**1**	0.9780
Salt & Pepper (density=0.001)	0.8825	**0.9476**
Gaussian Noise (0,002)	0.7844	**0.9786**

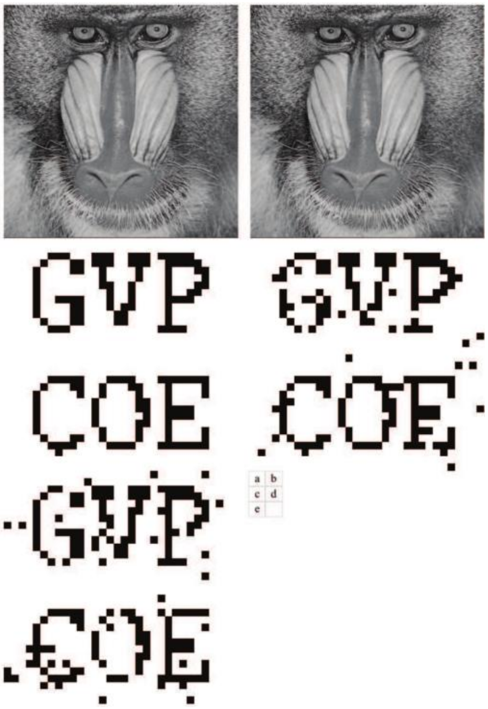


Fig. 12. a. Host image "Mandrill", b. Watermarked image "Mandrill", c. Original watermark, d. Extracted watermark in condition of being attacked by JPEG compression(Q=95), d. Extracted watermark in condition of being attacked by Salt & Pepper Noise (0.001).

Comparison results are listed in Table VII. According to this table, the NC values of the extracted watermarks by our method are always higher than those values of the extracted watermarks by the method in [23] under common image processing operators. Thus, it is obvious that our approach has superior performance.

TABLE VII: THE PERFORMANCE COMPARISON BETWEEN [23] AND THE PROPOSED SCHEME

Attacks	Ref. [23]		Proposed scheme	
	PSNR(dB)	NC	PSNR(dB)	NC
No Attack	43.36	1	42.9536	1
Resizing	29.54	0.6387	22.7426	0.8186
JPEG(Q=95)	42.18	0.8588	39.1722	0.9816
Salt & Pepper (density=0.001)	39.42	0.8797	35.0553	0.9745
Median	29.66	0.5629	22.8568	0.7656

IV. CONCLUDING REMARKS

In this paper, we have proposed a new watermarking scheme based on the combination of two powerful signal processing tools: CNN and DWT. The DWT has been used to decompose the host images into different sub-bands. Then, the pixels in the selected low frequency and high frequency sub-bands have been used as the inputs and the desired outputs to train the CNN. The watermark embedding and extracting processes are performed by the trained CNN. The extensive experimental results have been conducted to measure the invisibility and robustness of the proposed watermarking scheme. By the numerical results, the proposed method has demonstrated its superior performance in terms of PSNR, NC, and SSIM in comparison with the other methods.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

N. C. Sy conducted the research and experimental results; H. H Kha and N. M. Hoang analyzed the data; N. C. Sy wrote the paper; H. H. Kha and N. M. Hoang edited the paper; all authors had approved the final version.

REFERENCES

[1] S. I. Hasan, J. Boaddh, and A. Shrivastava, "A survey on visual properties and techniques of digital image data hiding," *International Journal of Scientific Research & Engineering Trends*, vol. 5, pp. 154–158, 2019.

[2] T. Wang and H. Li, "A novel digital image watermarking algorithm based on curvelet transform," *International Journal of Digital Content Technology and its Applications*, Jan. 2013.

[3] F. Y. Shih, *Digital Watermarking and Steganography: Fundamentals and Techniques*, Taylor & Francis, CRC Press, 2 edition, 2017.

[4] U. Yadav, J. P. Sharma, D. Sharma, and P. K. Sharma, "Different watermarking techniques and its applications: A review," *International Journal of Scientific & Engineering Research*, vol. 5, no. 4, April 2014.

[5] H. B. Kekre, S. Natu, and T. Sarode, "Performance evaluation of digital color image watermarking using column walsh wavelet transform," in *Proc. the International Conference on Signal, Networks, Computing, and Systems*, vol. 395, pp. 149–159, 2016.

[6] N. Nikolaidis and I. Pitas, "Robust image watermarking in the spatial domain," *Signal Processing*, vol. 66, no. 3, pp. 385-403, 1998.

[7] M. El-Gayyar and J. von zur Gathen, "Watermarking techniques spatial domain," Technical Report, University of Bonn Germany, 2006.

[8] S. Ranjbar, F. Zargari, and M. Ghanbari, "A highly robust two stage contourlet-based digital image watermarking method," *Signal Processing: Image Communication*, vol. 28, no. 10, pp. 1526-1536, 2013.

[9] H. Y. Leung, L. M. Cheng, and L. L. Cheng, "A robust watermarking scheme using selective curvelet coefficients," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 7, no. 2, pp. 163–181, 2009.

[10] S. C. Nguyen, K. H. Ha, and H. M. Nguyen, "A new image watermarking scheme using contourlet transforms," in *Proc. The 3nd International*

Conference on Information Technology, Computer, and Electrical Engineering, 2016.

[11] Y. R. Rao, E. Nagabhooshanam *et al.*, "Image watermarking using hybrid wavelets and directional filter banks," *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, 2012.

[12] Y. R. Rao, E. Nagabhooshanam, and P. Nikhil, "Directional based watermarking scheme using a novel data embedding approach," *Advanced Computing: An International Journal*, 2012.

[13] S. C. Nguyen, K. H. Ha, and H. M. Nguyen, "Digital image watermarking in spatial and transform domain: A survey and improved ppproach," in *Proc. 2015 International Symposium on Electrical and Electronic Engineering (ISEE2015)*, 2015.

[14] X. Zhang and Y. Yang, "A geometric distortion resilient image watermark algorithm based on dft-svd," *Computer Engineering*, 2006.

[15] I. J. Cox, J. Kilian, F. T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia," *IEEE Transactions on Image Processing*, vol. 6, no. 12, pp. 1673–1687, Dec. 1997.

[16] S. Kaur and M. Lal, "An invisible watermarking scheme based on modified fast haar wavelet transform and rsgwpt," in *Proc. 2015 2nd International Conference on Recent Advances in Engineering Computational Sciences (RAECS)*, Dec. 2015, pp. 1-5.

[17] R. K. Arya, S. Singh, and R. Saharan, "A secure non-blind block based digital image watermarking technique using dwt and dct," in *Proc. 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Aug. 2015, pages 2042–2048.

[18] A. Winarno, D. R. I. M. Setiadi, A. A. Arrasyid, C. A. Sari, and E. H. Rachmawanto, "Image watermarking using low wavelet subband based on 8x8 sub-block DCT," in *Proc. 2017 International Seminar on Application for Technology of Information and Communication (iSemantic)*, Oct. 2017, pp. 11–15.

[19] N. Ramamurthy and S. Varadarajan, "Robust digital image watermarking scheme with neural network and fuzzy logic approach," *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, no. 9, pp. 555–562, 2012.

[20] M. Vafaei, H. Mahdavi-nasab, and H. Pourghassem, "A new robust blind watermarking method based on neural networks in wavelet transform domain," *World Applied Sciences Journal*, vol. 22, no. 11, pp. 1572–1580, 2013.

[21] H. Kandi, D. Mishra, and S. R. K. S. Gorthi, "Exploring the learning capabilities of convolutional neural networks for robust image watermarking," *Computers & Security*, vol. 65, pp. 247–268, 2017.

[22] S. Huang and W. Zhang, "Digital watermarking based on neural network and image features," in *Proc. 2009 Second International Conference on Information and Computing Science*, May 2009, vol. 2, pp. 238–240.

[23] B. Jagadeesh, P. R. Kumar, and P. C. Reddy, "Robust digital image watermarking in discrete wavelet transform domain using back propagation neural networks," *GJMS Special Issue for Recent Advances in Mathematical Sciences and Applications*, vol. 2, pp. 19–24, 2014.

[24] Q. Baoming, Z. Pulin, and K. Qiao, "A digital watermarking algorithm based on wavelet packet transform and bp neural network," in *Proc. 2011 Seventh International Conference on Computational Intelligence and Security*, Dec. 2011, pp. 503–506.

[25] A. Benoraira, K. Benmahammed, and N. Boucenna, "Blind image watermarking technique based on differential embedding in dwt and dct domains," *EURASIP Journal on Advances in Signal Processing*, 2015, vol. 1, no. 55, Jul. 2015.

[26] A. Akter, Nur-E-Tajnina, and M. A. Ullah, "Digital image watermarking based on DWT-DCT: Evaluate for a new embedding algorithm," in *Proc. 2014 International Conference on Informatics, Electronics Vision (ICIEV)*, May 2014, pp. 1–6.

[27] H. Adeli, Z. Zhou, and N. Dadmehr, "Analysis of EEG records in an epileptic patient using wavelet transform," *Journal of Neuroscience Methods*, vol. 123, no. 1, pp. 69-87, 2003.

[28] A. Zear, A. K. Singh, and P. Kumar, "A proposed secure multiple watermarking technique based on DWT, DCT and SVD for application in medicine," *Multimedia Tools and Applications*, vol. 77, p. 48634882, February 2018.

[29] P. Kim, *Convolutional Neural Network*, Apress, Berkeley, CA, Berkeley, CA, pp. 121–147, 2017.

[30] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, September 1998.

[31] J. Bouvrie, *Introduction Notes on Convolutional Neural Networks*, 2014.

[32] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 14, pp. 600–612, 2004.

[33] T.-S. Chen, C.-C. Chang, and M.-S. Hwang, "A virtual image cryptosystem based upon vector quantization," *IEEE Transactions on Image Processing*, vol. 7, no. 10, pp. 1485–1488, Oct 1998.

[34] A. Joseph and K. Anusudha, "Robust watermarking based on DWT SVD," *arXiv e-prints*, September 2013.

[35] C. Song, S. Sudirman, M. Merabti, and D. Llewellyn-Jones, "Analysis of digital image watermark attacks," in *Proc. 2010 7th IEEE Consumer Communications and Networking Conference*, Jan. 2010, pp. 1–5.

[36] J.-J. Pan, Y.-Y. Tang, and B.-C. Pan, "The algorithm of fast mean filtering," in *Proc. 2007 International Conference on Wavelet Analysis and Pattern Recognition*, Nov. 2007, vol. 1, pp. 244–248.

Nguyen Chi Sy received the B.Eng. degree from Danang University of Technology in 1997, M. Eng. degree from IFI (L'Institut de la Francophonie pour l'Informatique), Vietnam and France in 2000, all in information and technology. From 2001 to now, he is with the Centre of Information and Technology and with the Department of Engineering and Technology, Phuyen University. He is now a PhD student at the Faculty of Electrical and Electronics Engineering, Ho Chi Minh City University of Technology. His current research interests are of digital watermarking, image processing, fractal imaging and free and open source software.

Ha Hoang Kha received the B.Eng. and M.Eng. degrees from Ho Chi Minh City University of Technology, in 2000 and 2003, respectively, and the Ph.D. degree from the University of New South Wales, Sydney, Australia, in 2009, all in electrical engineering and telecommunications. From 2000 to 2004, he was a research and teaching assistant with the Department of Electrical and Electronics Engineering, Ho Chi Minh City University of Technology. He was a visiting research fellow at the School of Electrical Engineering and Telecommunications, the University of New South Wales, Australia, from 2009 to 2011. He was a postdoctoral research fellow at the Faculty of Engineering and Information Technology, University of Technology Sydney, Australia from 2011 to 2013. He is currently a lecturer at the Faculty of Electrical and Electronics Engineering, Ho Chi Minh City University of Technology, Vietnam. His research interests are in digital signal processing and wireless communications, with a recent emphasis on convex optimization techniques in signal processing for wireless communications.

Nguyen Minh Hoang received the B.Eng. degree from Ho Chi Minh City University of Technology in 1986, M. Eng. degree from Asian Institute of Technology (AIT), Thailand in 1994, and the Ph.D. degree from Institute of Communication Network, Vienna University of Technology, Austria in 2001. He is currently with Saigon Institute of ICT (SaigonICT) and with the Department of Electrical and Electronics Engineering, Ho Chi Minh City University of technology. His research interests are of computer and communication networks, wireless and mobile communication networks, mobile ad-hoc and wireless sensor networks.