Robust Image Watermarking Framework Powered by Convolutional Encoder-Decoder Network

Thien Huynh-The, Cam-Hao Hua, Nguyen Anh Tu, Dong-Seong Kim

Kumoh National Institute of Technology, S. Korea
Kyung Hee University, S. Korea
Nazarbayev University, Kazakhstan

Introduction

Fig. 1 Unprotected image is downloaded and used illegally by anyone with owner permission

- Share photos on social networks without any preliminary authorship authentication and protection.
- Several critical issues (e.g., illegal and malicious usage) harming copyright protection

☞ Digital image watermarking

Fig. 2 Digital image authenticated by a logo watermark

Background

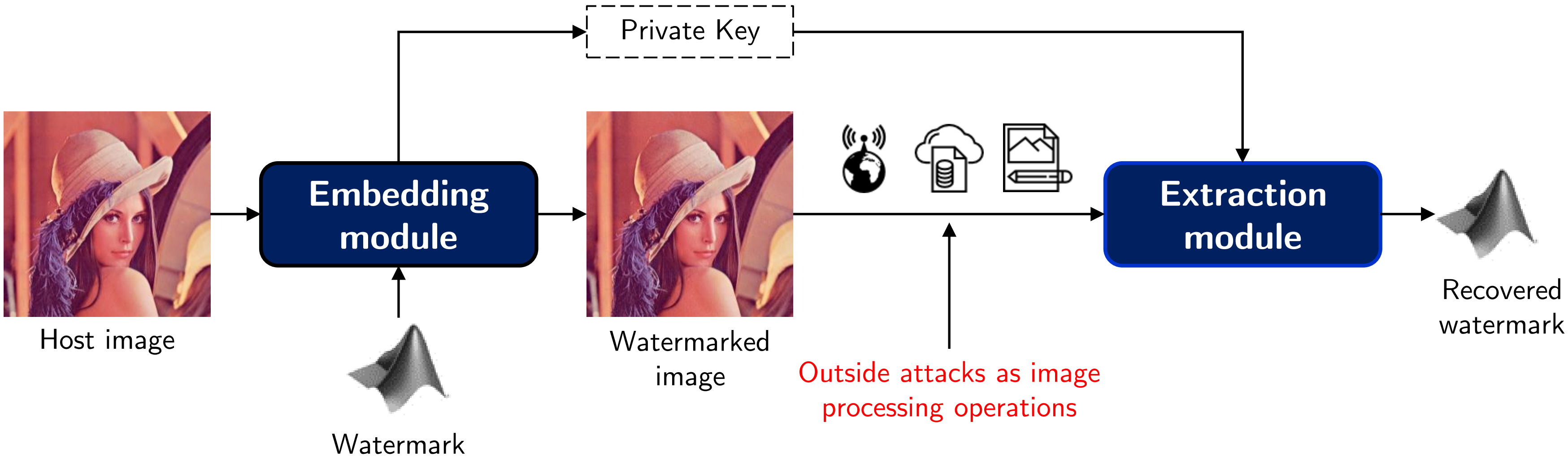


Fig. 3 General watermarking model includes embedding and extraction module

Definition: Digital watermarking is a process that allows the insertion of a watermark image into a host image for invisibility

- In an inverse process, the hidden information is recovered from the watermarked image to authenticate its originality
- During storage, transmission, and utilization, image may be suffered various intentional attacks

Taxonomy

- Many watermarking approaches have been introduced for enhancing **image imperceptibility** and **watermark robustness**.
- Generally, three categories of traditional watermark technique
  - Blind watermarking
  - Semi-blind watermarking
  - And non-blind watermarking
- Watermarking is performed on
  - Space-time domain
  - Transformed domain (e.g., Cosine, Fourier, and Wavelet)
- Host image and watermark information
  - Host image: gray-scale/color image
  - Watermark: bit stream, binary/gray-scale/color image

State-of-the-art

- Recently, **machine learning** (ML) is considered for supervised image watermarking
  - Genetic algorithm [Agarwal-2013]
  - Hidden Markov model [Amini-2017]
  - Neural network [Tsai-2017]
  - Extreme learning machine [Mishra-2018]
  - Support vector machine [Wang-2017]
- Compared with traditional watermarking, supervised approaches gains watermark robustness, but the efficiency is limited by the **high variety of digital attacks**
  - Geometric transformation
  - Non-geometric transformation
  - Lossy compression and etc.

  ☞ Intentional attacks try to destroy/remove the hidden signature of author.

[**Agarwal-2013**] C. Agarwal, A. Mishra and A. Sharma, "Gray-scale image watermarking using GA-BPN hybrid network," *J. Vis. Commun. Image Represent.*, vol. 24, no. 7, pp. 1135-1146, Oct. 2013

[**Amini-2017**] M. Amini, M. O. Ahmad and M.N.S. Swamy, "A new locally optimum watermark detection using vector-based hidden Markov model in wavelet domain," *Signal Process.*, vol. 137, pp. 213-222, Aug. 2017.

[**Tsai-2011**] H.-H. Tsai and C.-C. Liu, "Wavelet-based image watermarking with visibility range estimation based on HVS and neural networks," *Pattern Recognit.*, vol. 44, no. 4, pp. 751-763, April 2011.

[**Mishra-2018**] A. Mishra, A. Rajpal and R. Bala, "Bi-directional extreme learning machine for semi-blind watermarking of compressed images," *J. Inf. Secur. Appl.*, vol. 38, pp. 71-84, Feb. 2018.

[**Wang-2017**] C. Wang, X. Wang, C. Zhang and Z. Xia, "Geometric correction based color image watermarking using fuzzy least squares support vector machine and Bessel K form distribution," *Signal Process.*, vol. 134, pp. 197-208, May 2017.

State-of-the-art

- Deep learning (DL) with convolutional neural networks (CNNs)
  - Fundamentally developed for computer vision tasks.
  - And lately exploited for image watermarking.
- Compared with ML-based, DL-based watermarking recovers watermark more robustly against diverse cyber-attacks based on the **capability of learning attack patterns**.

Method	Technique	Limitation
Kandi-2017	• Two CNNs for learning 1-bits and 0-bits embedding schemes. • Code books of feature maps are used for recovery	• High complexity • Non-attack pattern learning
Mun-2019	• Embedding watermark on the time-space domain • Simulate attacks for learning patterns	• More sensitive compared with frequency-domain embedding

- Applying DL techniques for an efficient image watermarking remains an open issue

[**Kandi-2017**] H. Kandi et al., "Exploring the learning capabilities of convolutional neural networks for robust image watermarking," *Comput. Secur.*, vol. 65, pp. 247-268, March 2017.
[**Mun-2019**] S.-M. Mun, S.-H. Nam, H. Jang, D. Kim and H.-K. Lee, "Finding robust domain from attacks: A learning framework for blind watermarking," *Neurocomputing*, vol. 337, pp. 191-202, April 2019.

Problem statement

- Current ML- and DL-based watermarking approaches
  - Inefficiency of handling various cyber-attacks.
  - Lack of a mechanism to learn simulated attacking patterns.

☞ Report inadequate performance under some common digital image transformations

Goal: Development of a high-performance image watermarking framework by exploiting deep learning technique

Objective:

- High image imperceptibility
- Strong watermark robustness
- Accurate watermark recovery

Solution

A deep convolutional encoder-decoder (CED) network that is capable of learning different realistic attacking patterns

Methodology

- A deep learning-based digital image watermarking framework for copyright protection and ownership authentication with key points
  - Embedding process performed on the **wavelet domain**.
  - Enhancing imperceptibility with optimal **block selection** and **bit encoding** schemes.
  - **Simulation of image transformations** for learning attacking patterns.
  - **Convolutional encoder-decoder network** for recognizing embedding map

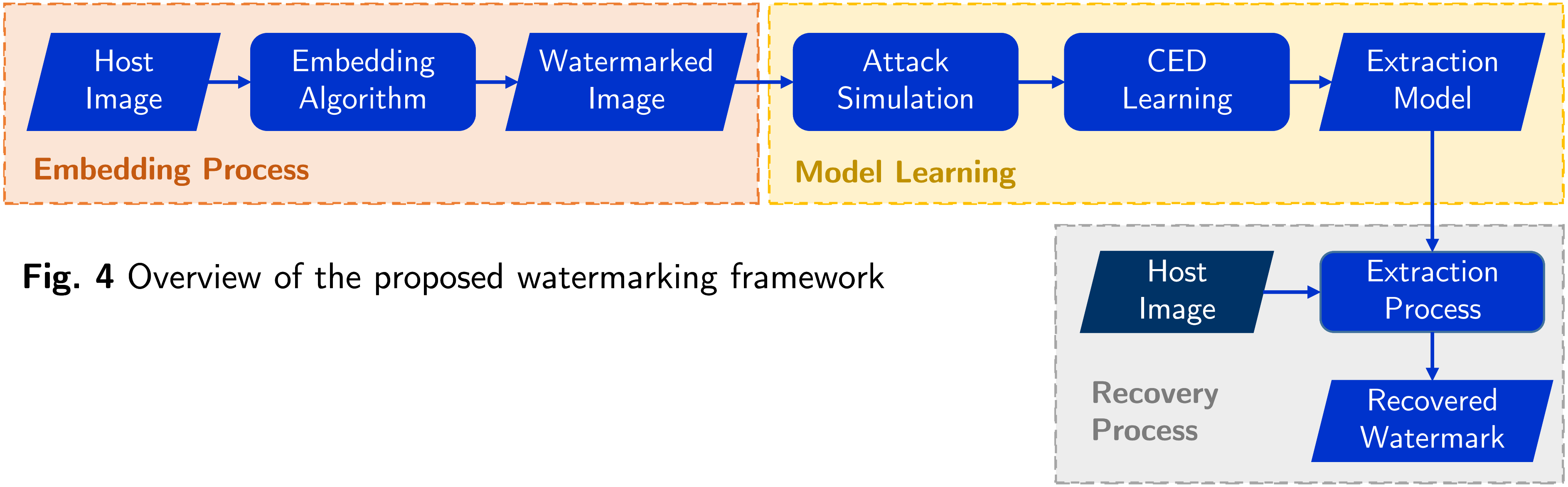


Fig. 4 Overview of the proposed watermarking framework

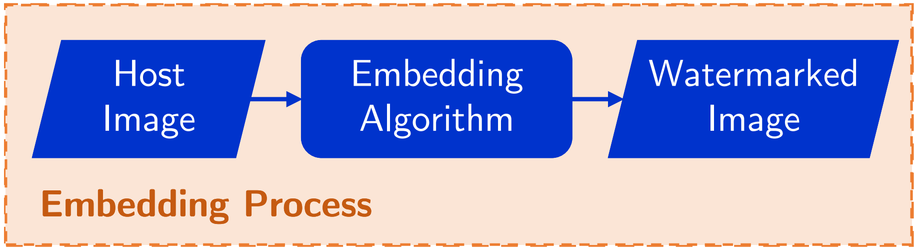
Embedding process

Leverage an encoding algorithm

- Establish wavelet blocks
- Quantize the coefficient difference of each block for 0-bits and 1-bits
- ☞ **According watermark bit, we adjust the values of wavelet coefficients.**

Optimize image imperceptibility

- Selective blocks
  - Larger difference blocks for 1-bits embedding
  - Smaller difference blocks for 0-bits embedding
- Half-quantity of coefficient adjustment
- Encoding thresholds are automatically determined based on minimizing MSE



Perform on the wavelet domain using DWT (the horizontal and vertical detail components ***cH*** and ***cV***)

After embedding watermark, the watermarked image is reconstructed by **IDWT**

Embedding process

For encoding 0-bits

Lower encoding threshold $\ell_0 = \upsilon - \lambda/2$

If $\delta_i^{\downarrow} > \ell_0$

Coefficient difference

Adjustment quantity $\chi^0 = \delta^{\downarrow} - \ell_0$

$$\begin{cases} cH_i = cH_i - \chi_i^0/2 \\ cV_i = cV_i + \chi_i^0/2 \end{cases} ;\forall cH_i \geq cV_i$$

$$\begin{cases} cH_i = cH_i + \chi_i^0/2 \\ cV_i = cV_i - \chi_i^0/2 \end{cases} ;\forall cH_i < cV_i.$$

Half-quantity adjustment

If $\delta_i^{\downarrow} \leq \ell_0$

$$\begin{cases} cH_i = cH_i \\ cV_i = cV_i. \end{cases}$$

For encoding 1-bits

Upper encoding threshold $\ell_1 = \upsilon + \lambda/2$

If $\delta_i^{\downarrow} < \ell_1$

Adjustment quantity $\chi^1 = \ell_1 - \delta^{\downarrow}$

$$\begin{cases} cH_i = cH_i + \chi_i^1/2 \\ cV_i = cV_i - \chi_i^1/2 \end{cases} ;\forall cH_i \geq cV_i$$

$$\begin{cases} cH_i = cH_i - \chi_i^1/2 \\ cV_i = cV_i + \chi_i^1/2 \end{cases} ;\forall cH_i < cV_i.$$

If $\delta_i^{\downarrow} \geq \ell_1$

$$\begin{cases} cH_i = cH_i \\ cV_i = cV_i, \end{cases}$$

Objective: Reducing the total mean square error (MSE) of the watermarked image if compared with the original image

$$\upsilon = \underset{x \in [0, \max(\delta)]}{\arg\min} \left(\sum_i (\Delta_i - x)^2 \right)$$

Note: The embedding quality can be partly controlled by an embedding strength factor λ that is manually pre-defined.

Extraction model learning

Observation

- At each watermarked image, the coefficient difference δ_i
  - Smaller than the lower threshold ℓ_0 for 0-bits detection
  - Larger than the upper threshold ℓ_1 for 1-bits detection

☞ Conventional approaches try to estimate a threshold ℓ_e, where $\ell_0 \leq \ell_e \leq \ell_1$, to classify either 0-bit or 1-bit hidden in wavelet blocks [HuynhThe-2018].

Drawback

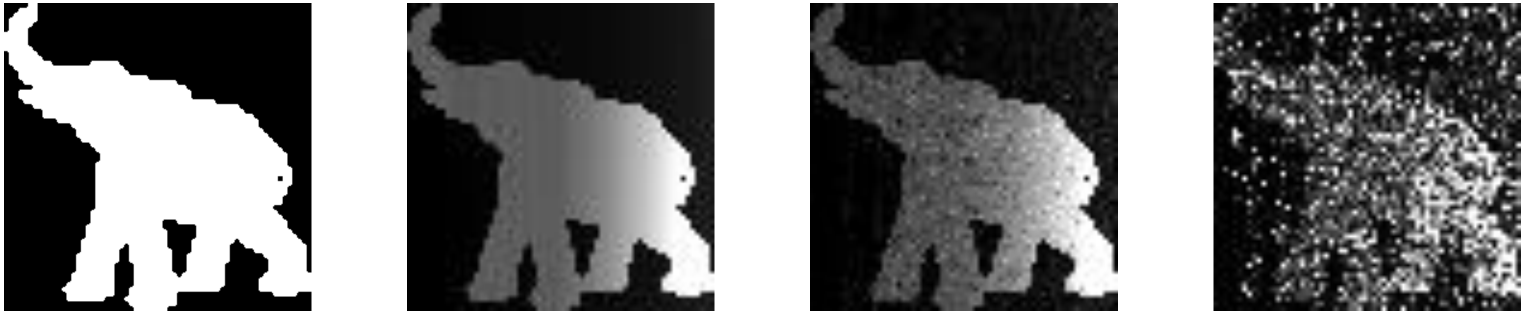
Under critical intentional attacks as digital image transformations, watermarked image is modified → coefficient difference value is changed → estimation of the threshold is incorrect → extraction accuracy of watermark is reduced.

[**HuynhThe-2018**] T. Huynh-The, C.-H Hua, N. A. Tu, T. Hur, J. Bang, D. Kim, M. B. Amin, B. H. Kang, H. Seung, S. Lee, "Selective bit embedding scheme for robust blind color image watermarking," *Inf. Sci.*, vol. 426, pp. 1-18, Feb. 2018.

Extraction model learning

Potential solution

- Supervised learning attack patterns over the coefficient differences of watermarked images regarding to
  - Simulation of watermarked image under various realistic attacks
  - Design of network for classification 0-bits and 1-bits



(a) (b) (c) (d)

Fig. 5 (a) Original watermark image, **(b)** the embedding map after visualizing coefficient difference value without attack (darker for values less than the lower threshold and brighter for values larger than the upper threshold), **(c)-(d)** under medium- and strong-level attacks

Extraction model learning

- By converting **embedding map** of coefficient difference to **image**, **watermark extraction** can be treated as **semantic image segmentation** (aka pixel-wise classification).

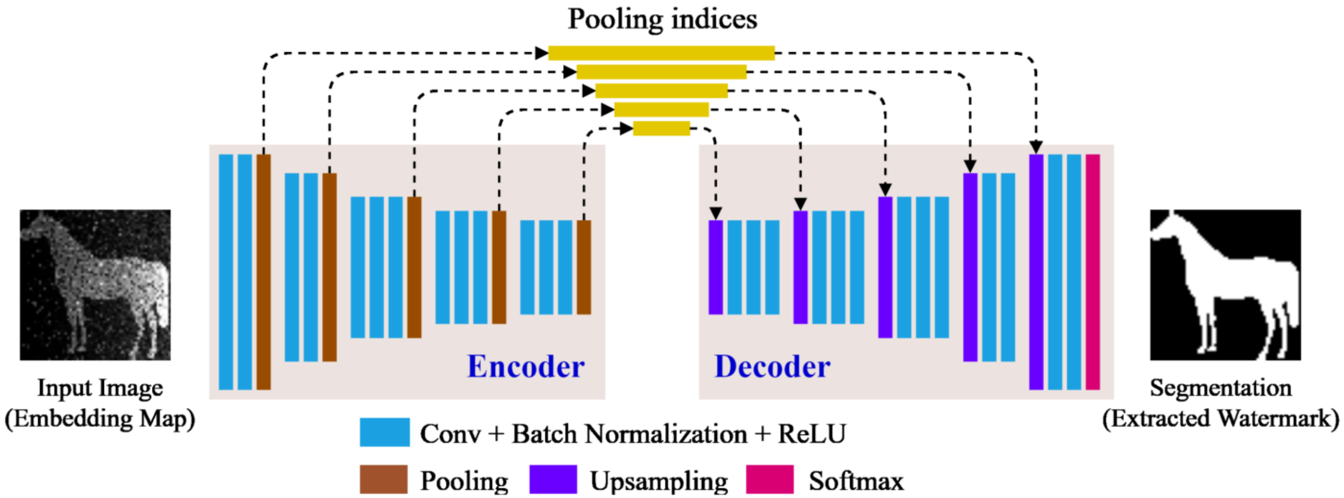


Fig. 6 General architecture of convolutional encoder-decoder network deployed for watermark extraction with an embedding map as the input and the recovered watermark as output

- The architecture is configured with **initial weights of VGG-16**

Extraction model learning

- Learning different attack patterns over learning embedding maps by **simulating various kinds of image transformation** as cyber-attacks.
- The below is how to generate a dataset for learning attack patterns.
- During training, the label image (aka the ground truth watermark) is supplied.
- Once the training is done, the CED-based extraction model is ready for watermark recovery

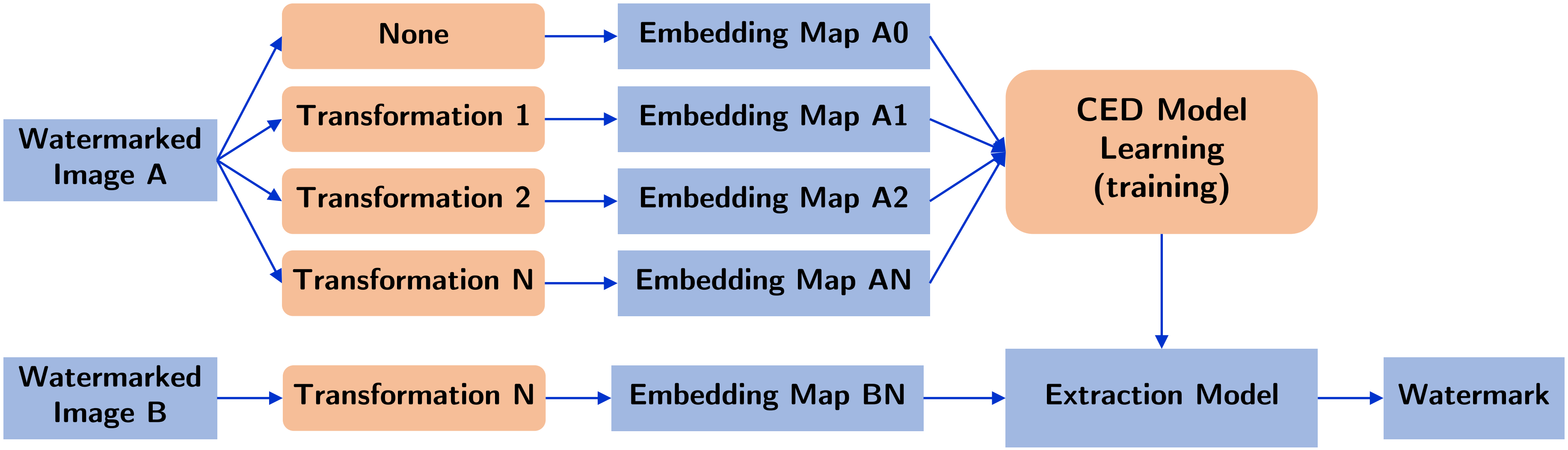
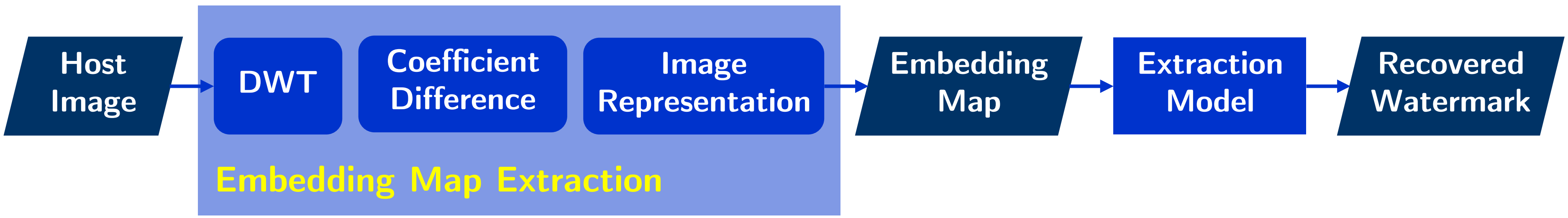


Fig. 7 Scheme of generate training set of embedding maps extracted from attacked image.

Recovery process



The watermark recovery flows

- Input: an embedded image suffering arbitrary digital image transformations
- Output: a recovered watermark
- Transform to wavelet domain
- Calculate coefficient differences in wavelet blocks
- Structure and represent embedding map of difference values into image
- Predict watermark bit over the pixel-wise classification using trained CED model

$$\overset{\text{Recovered bit}}{w'_{(x,y)}} = \arg\max_{k} p_{(x,y)}(k)$$

Experimental result

- Dataset of host images
  - Training: BOSSbased-1.01 - 10,000 512x512 gray-scale images
  - Testing: 60 images as follows
- Watermark: a 64x64 binary image
- For attack simulation
  - 49 digital image transformations + 1 non-attack
  - 500,000 embedding maps are generated for training CED model
- Training setup: 30 epochs, mini-batch size of 64, and initial learning rate of 0.01
- Evaluation metrics of host image imperceptibility: PSNR and SSIM, and watermark robustness: NC

Fig. 8 Host gray-scale image set for evaluating CED-based extraction model.

Experimental result

TABLE I
RESULTS OF IMAGE IMPERCEPTIBILITY BENCHMARK

Image	PNSR (dB)	SSIM
Avion	50.31	0.9980
Baboon	47.56	0.9982
House	49.49	0.9979
Lena	47.46	0.9942
Malight	48.01	0.9951
Peppers	48.15	0.9954
Sailboat	50.73	0.9988
Toucan	48.04	0.9934
Average (60 images)	**47.85**	**0.9948**

- The quality of host image with the embedding strength $\lambda = 40$
- Average of 60 test images
  - PSNR: **47.85** dB
  - SSIM: **0.9948**
- As a performance trade-off
  - Smaller embedding strength $\rightarrow$ better imperceptibility
  - Greater embedding strength $\rightarrow$ more breakable watermark under critical attacks

Experimental result

- Achieve **high accuracy** of watermark recovery
- Robustly **against many common digital transformations**, except geometric rotation

TABLE II
RESULTS OF WATERMARK ROBUSTNESS BENCHMARK

Median Filtering		**Average Filtering**		**Gaussian Filtering**		**Blurring**		**Scaling**		**Cropping**	
Size	NC	Size	NC	Size	NC	No. Pixels	NC	Ratio	NC	Size	NC
3×3	1.0000	3×3	1.0000	3×3	1.0000	3	1.0000	up 200%	1.0000	32×32	1.0000
5×5	1.0000	5×5	1.0000	5×5	1.0000	5	1.0000	up 400%	1.0000	64×32	1.0000
7×7	1.0000	7×7	1.0000	7×7	1.0000	7	1.0000	down 50%	1.0000	32×64	1.0000
9×9	0.9984	9×9	0.9999	9×9	1.0000	9	0.9997	down 25%	1.0000	64×64	1.0000

Rotation		**Gaussian Noise**		**Salt&Pepper Noise**		**JPEG Compression**				**Others**	
Degree	NC	σ^2	NC	*den*	NC	QF (%)	NC	QF (%)	NC	Name	NC
1	0.9532	0.001	1.0000	0.1	1.0000	10	0.9999	60	1.0000	Attack-free	1.0000
2	0.9428	0.002	1.0000	0.5	1.0000	20	1.0000	70	1.0000	Hist. Equal.	1.0000
3	0.9463	0.003	1.0000	1	1.0000	30	1.0000	80	1.0000		
4	0.9392	0.005	1.0000	2	1.0000	40	1.0000	90	1.0000		
5	0.9194	0.010	1.0000	5	0.9999	50	1.0000				

Method Comparison

TABLE III
METHOD PERFORMANCE COMPARISON ON LENA

Attack	Tsai	Tsougenis	Huynh-The	Kandi	Proposed
PSNR (dB)	41.53	40.38	48.17	58.91	47.46
Attack-free	0.9924	1.0000	1.0000	1.0000	1.0000
Median filtering 3×3	0.6940	0.9726	0.9804	0.9100	1.0000
Average filtering 3×3	N/A	0.9760	0.9900	0.8600	1.0000
Gaussian filtering 3×3	0.7904	1.0000	0.9968	0.8800	1.0000
Blurring 6 pixels	N/A	0.9356	0.9372	N/A	1.0000
Scaling down 50%	N/A	0.8126	0.9984	N/A	1.0000
Scaling up 200%	-0.0196	0.9934	1.0000	N/A	1.0000
Cropping 1%	0.8666	0.9792	0.9984	1.0000	1.0000
Cropping 4%	0.8614	0.8444	0.9882	1.0000	0.9961
Rotation 5%	-0.0142	0.9928	0.1760	0.3700	0.9194
Gaussian noise $(\sigma^2 = 0.006)$	0.7792	N/A	0.8164	0.3500	1.0000
Gaussian noise $(\sigma^2 = 0.025)$	N/A	0.8542	0.9418	0.9450	1.0000
Pepper noise $(den = 0.3\%)$	0.8892	N/A	0.9628	1.0000	1.0000
Pepper noise $(den = 1\%)$	N/A	0.9994	0.8846	1.0000	1.0000
Lossy JPEG $(QF = 30\%)$	0.2216	0.8470	0.8116	0.6100	1.0000
Lossy JPEG $(QF = 40\%)$	N/A	0.8666	0.8562	0.6650	1.0000
Lossy JPEG $(QF = 50\%)$	0.3666	0.8762	0.8834	0.6800	1.0000
Lossy JPEG $(QF = 70\%)$	0.5482	0.9524	0.9412	0.7100	1.0000
Average NC	0.5813	0.9314	0.8980	0.7987	0.9953

H.-H. Tsai et al., "Color image watermark extraction based on support vector machines," *Inf. Sci.*, vol. 177, no. 2, pp. 550-569, Jan. 2007.
E.D.Tsougenis et al., "Adaptive color image watermarking by the use of quaternion image moments," *Expert Syst. Appl.*, vol. 41, no. 14, pp. 6408-6418, Oct. 2014.
H. Kandi et al., "Exploring the learning capabilities of convolutional neural networks for robust image watermarking," *Comput. Secur.*, vol. 65, pp. 247-268, March 2017.
T. Huynh-The et al., "Improving digital image watermarking by means of optimal channel selection," *Expert Syst. Appl.*, vol. 62, pp. 177-189, Nov. 2016.

Conclusion

- A deep learning-based digital image watermark framework
  - **An encoding algorithm on wavelet domain**
    - Wavelet block selection
    - Half-quantity wavelet coefficient difference quantization
    - MSE-minimization based encoding thresholds
  - ☞ **Improve image imperceptibility**
  - Deep learning-based extraction model
    - Simulation of digital image transformations
    - Learning of attack patterns
    - Convolutional encoder-decoder network
  - ☞ **Enhance watermark robustness**
- Further investigate with various watermark images and more complex attack scenarios
  - Embedding a new watermark on a watermarked image
  - Combined attacks such as filtering + compression

Thank you

Dr. Thien Huynh-The
Email: thienht@kumoh.ac.kr