

# A Deep Learning–based Audio-in-Image Watermarking Scheme

Arjon Das

Department of Computer Science  
University of Nebraska Omaha  
arjondas@unomaha.edu

Xin Zhong

Department of Computer Science  
University of Nebraska Omaha  
xzhong@unomaha.edu

**Abstract**—This paper presents a deep learning–based audio-in-image watermarking scheme. Audio-in-image watermarking is the process of covertly embedding and extracting audio watermarks on a cover-image. Using audio watermarks can open up possibilities for different downstream applications. For the purpose of implementing an audio-in-image watermarking that adapts to the demands of increasingly diverse situations, a neural network architecture is designed to automatically learn the watermarking process in an unsupervised manner. In addition, a similarity network is developed to recognize the audio watermarks under distortions, therefore providing robustness to the proposed method. Experimental results have shown high fidelity and robustness of the proposed blind audio-in-image watermarking scheme.

**Index Terms**—Audio-in-image watermarking, deep learning, neural networks, robustness

## I. INTRODUCTION

Audio-in-image watermarking refers to the process of embedding and extracting audio watermark information covertly on a cover-image (see Fig. 1). An audio watermark is hidden into a cover-image to create a to-be-transmitted marked-image. The covert process suggests that the marked-image visually does not give out any watermark information, but processing the marked-image through an authorized extraction process can reveal the correct watermark.

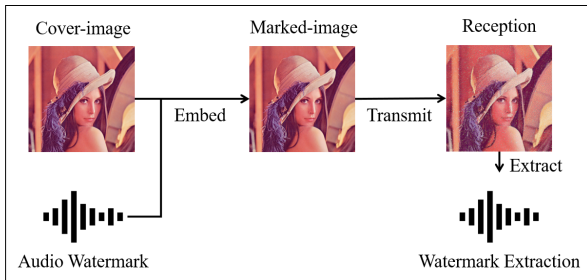


Fig. 1: Audio-in-Image Watermarking.

Based on different target application scenarios, there are many typical factors and properties a watermarking scheme should provide [1]–[3]. For example, the fidelity that secures a high similarity between the marked-image and the cover-image, and the blindness that the watermark extractions require no information about the original images. Compared

This work is supported by the National Science Foundation under Grant CIF/SaTC-2104267. This work is also partially supported by the Research Development Program of University of Nebraska Omaha.

to the highlighted properties in related areas, such as the undetectability in steganography [4] and the amount capacity in data hiding [5], an image watermarking scheme often considers the robustness that ensures correct watermark extraction and recognition when the marked-image is distorted or disturbed.

**Audio watermarks:** From the aspect of applications, an audio watermark facilitates different downstream scenarios by providing multiple information, for example, two-factor authentication [6], [7] by providing the speaker and content information. For research investigation, there is quite a number of work that separately focused on cover-image and cover-audio watermarking [8]–[10]. But audio-in-image watermarking remains a challenge [11] because audio signal spans between a limited range and inserting it into an image can cause severe tampering of its amplitudes. Therefore, it is of great interest to explore a watermarking scheme that combines audio and image.

**Deep learning–based image watermarking:** Manually designing watermarking schemes often requires domain knowledge, for example, understanding a signal’s frequency band to determine where to insert a specific watermark. As a result, almost every particular scenario requires a special design. Different schemes have emerged that incorporate deep neural networks to implement image watermarking, which adapts to the demands of increasingly diverse applications. For instance, Kandi *et al.* [12] incorporated two convolutional autoencoders to reconstruct one cover-image so that the zero or one bits in a binary watermark stream can be indicated. Li *et al.* [13] proposed to embed a watermark into a gray-scale cover-image with manual frequency-domain methods and extract it with a convolutional neural network. Although many methods have successfully applied deep neural networks to assist some aspects in watermarking, it has been pointed out [1] that designing deep learning architecture to fully learn the entire watermark embedding and extracting processes is more challenging. Moreover, because deep neural networks can be susceptible to noises [14], achieving robustness in deep learning–based watermarking is a major challenge because distorted marked-images can mislead the neural networks. Some proposals adopt adversarial training [15] to address the issue. Mun *et al.* [16] simulated attacks on marked-images during training, and applied reinforcement learning to learn

a robust extraction. Zhong *et al.* [1] trained an end-to-end neural network to learn the entire embedding and extraction and proposed an invariance layer for the robustness of common image processing attacks that modify pixel values in a marked-image. Chen *et al.* [17] proposed a neural network, which included attack simulations during the training, to validate if the original and the extracted watermarks are identical.

In this paper, we propose a novel deep learning-based audio-in-image watermarking scheme that achieves blindness, fidelity, and robustness simultaneously. Our major contribution is threefold. First, a deep learning framework is proposed to fully learn the watermark embedding and extracting processes for audio-in-image watermarking. Second, to reduce the requirement of watermarking domain knowledge, the proposed watermarking model can be trained in an unsupervised manner that requires no manual labeling. Finally, we develop a Similarity Network to tolerate and recognize audio watermarks under distortions, therefore providing robustness to our combined watermarking and recognition process.

## II. THE PROPOSED SCHEME

This section describes details on (i) the architecture that fully learns the process of the audio-in-image watermarking (namely the WM Network), (ii) the Similarity Network that learns to recognize the audio watermarks under distortions, and (iii) the loss and training processes. The proposed scheme is summarized in Fig. 2.

### A. WM Network

Although the WM Network is trained as a single and deep network, we conceptually modularized it to the Encoder, Decoder, Embedder, and Extractor Nets to facilitate our description and illustration.

1) *The Encoder and Decoder Nets:* The Encoder Net learns to map the watermark  $W$  to its code  $W_c$ :  $W \rightarrow W_c$  from the training samples  $w_i \in W, i = 1, 2, 3, \dots$ . Inversely, the Decoder Net learns to map the extracted code  $W'_c$  to the watermark extraction  $W'$ :  $W'_c \rightarrow W'$ . The Encoder Net is a two layer LSTM. The audio input is firstly reshaped from  $8, 192 \times 1$  to  $128 \times 64$ . The first LSTM takes a 64-dimensional input vector per time-step, and both LSTM layers are configured to have 128-dimensional hidden state vectors. Hence the output from each LSTM cell is a 128-dimensional vector. After 128 time-steps, the output size will be  $128 \times 128$ . Symmetrically, the Decoder Net converts back a  $128 \times 128 \times 1$  feature map into a  $8, 192 \times 1$  audio data.

2) *The Embedder and the Extractor Nets:* The Embedder Net learns a function that includes the  $128 \times 128 \times 1$   $W_c$  into the  $128 \times 128 \times 3$  cover-image  $C$  while retaining visual fidelity. First,  $W_c$  and  $C$  are separately processed by two sequences of convolutional layers to produce  $128 \times 128 \times 64$  feature maps. Each sequence contains four convolutional layers of 8, 16, 32, and 64 filters respectively. Although identically structured, these sequences do not share any weights because they are processing image and audio separately. Then, the feature maps of both  $W_c$  and  $C$  are concatenated along the channel dimension to obtain a  $128 \times 128 \times 128$  feature map. To

promote the propagation of information throughout the feature map channels, a channel-wise fully connected layer [18] with 512 units is applied. The output ( $128 \times 128 \times 512$ ) is then fed into another sequence of convolutional layers with 128, 64, 32, 16, 8, 4, and 3 filters to obtain a marked-image  $M$ .

The Extractor Net aims at distilling out the watermark information embedded inside the marked-image. To achieve an  $M \rightarrow W'_c$  mapping, the network takes input  $M$  and passes it through sequences of convolutional layers with different filter numbers. A 512-unit channel-wise fully-connected layer is applied between two sequences of convolutional layers. The final convolutional layer of the extractor net outputs a  $128 \times 128 \times 1$   $W'_c$ . To ensure successful gradient propagation throughout the extractor net, we incorporated two skip connections [19].

### B. Similarity Network

Marked-images may distort during transmission. Consequently, the watermark extractions might get distorted. Although experimentations reveal that the proposed WM Network is robust to some noises, it is necessary to identify the authenticity of the distorted watermark extractions. Simple error (e.g., RMSE) analysis fails to indicate such a validation process. Hence, we introduce the Similarity Network to validate if a watermark  $W$  and its extraction  $W'$  are similar by structural analysis of audio signal pairs. The ideal training result is to be consistent with human judgment. Explicitly, if an audio clip is distorted but recognizable by humans, it should also be recognizable by the Similarity Network. As a result, from extraction to validation, the Similarity Network provides robustness from a semantic recognition perspective.

The Similarity Network takes a pair of  $8, 192 \times 1$  ( $W, W'$ ) as input and outputs a  $[0, 1]$  scalar where 0 means completely different and 1 means identical, and values in between indicating a similarity level. We formulate this similarity computation as metric learning and develop our Similarity Network based on the Siamese architecture [20]. Our Similarity Network firstly applies our trained Encoder Net so that the pre-trained LSTM layers can extract features from the audio inputs rapidly. After obtaining the  $128 \times 128 \times 1$  feature maps of  $W$  and  $W'$ , two identical weight-shared sequences of convolutional layers with max-pooling is applied to compress the feature maps to  $8 \times 8 \times 256$ . These feature maps are then flattened to 16,384 and passed to a 4,096-unit fully-connected layer. The absolute difference is computed between the two 4,096 feature vectors of  $W$  and  $W'$ , and a single-unit fully-connected layer with Sigmoid activation is used to output the similarity scalar.

### C. Training and Loss Functions

1) *Encoder-Decoder Net Pre-training:* While the marked-image is output at the middle of the WM Network, the watermark extraction has to propagate through the entire WM Network. As a result, the extraction may fail to pass through the rest of the network during the training. To address this issue, we pre-train the Encoder and Decoder Nets, reconstructing  $W$  to itself. Hence, the first step of pre-training the proposed scheme is to train an Encoder-Decoder net, thus inputting  $W$  to the Encoder, and the output of the Decoder should be  $W$  itself.

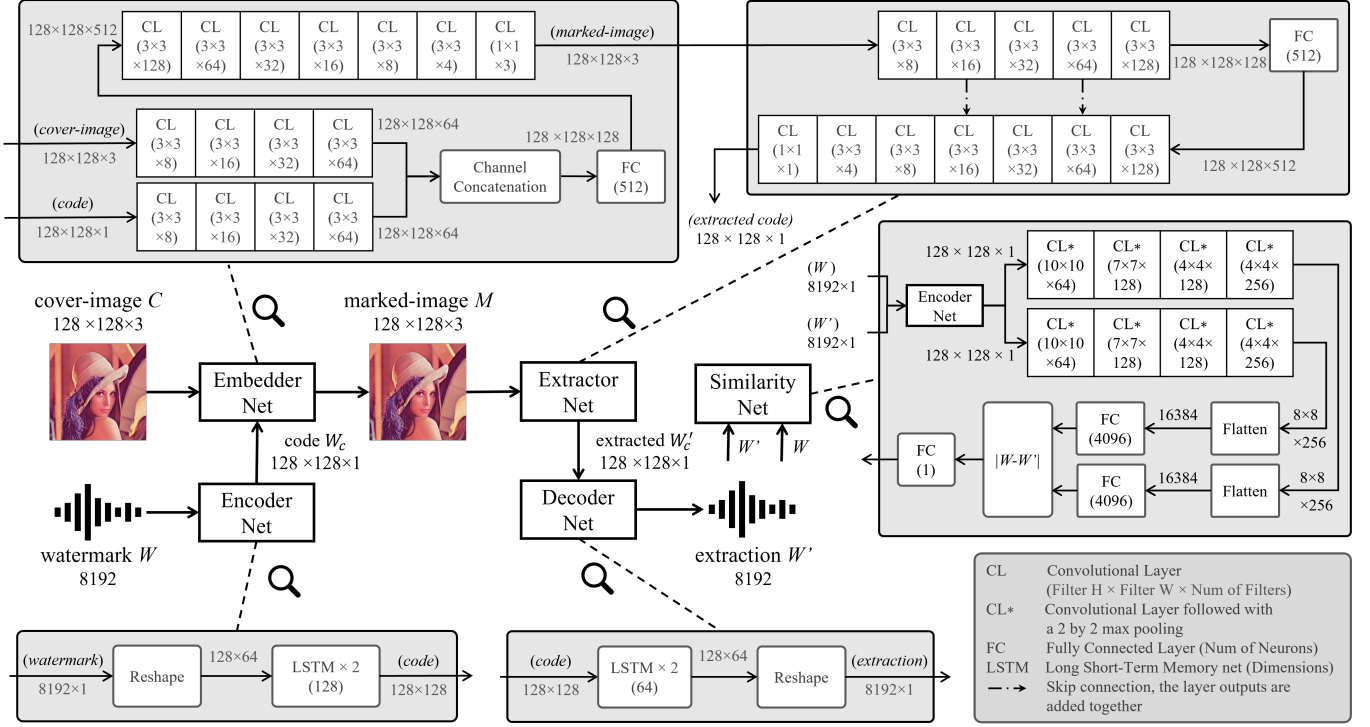


Fig. 2: Proposed scheme: The WM Network (Encoder, Embedder, Extractor, and Decoder Nets), and the Similarity Network.

Given training sample watermarks  $w_i \in W, i = 1, 2, 3, \dots$ , the loss function for training this Encoder-Decoder net is a simple mean squared error between  $w_i$  and its reconstruction  $\bar{w}_i$ :

$$\mathcal{L}(w_i, \bar{w}_i) = (w_i - \bar{w}_i)^2 \quad (1)$$

2) *Training the WM Network:* The second step is to train the entire WM Network as a single network where the (pre-trained) Encoder and Decoder Nets are fine-tuned, and the Embedder and Extractor Nets are trained from scratch. Given the samples  $w_i \in W, i = 1, 2, 3, \dots$  and  $c_i \in C, i = 1, 2, 3, \dots$ , an extraction loss that minimizes the difference between the extraction  $w'_i$  and  $w_i$  is applied to ensure a successful extraction of the watermark. In addition, a fidelity loss that minimizes the difference between a marked-image  $m_i$  and  $c_i$  is applied to enable a high fidelity. Mean squared errors are applied to compute the differences, and the loss is given as:

$$\mathcal{L}(w_i, w'_i, c_i, m_i) = \lambda_1(w_i - w'_i)^2 + \lambda_2(c_i - m_i)^2 \quad (2)$$

where  $\lambda_1$  and  $\lambda_2$  are the weighing factors.

3) *Training the Similarity Network:* Given the input pair  $w_i^1$  and  $w_i^2$ , Similarity Network is trained to recognize if they are the same. Let  $y(w_i^1, w_i^2)$  be the label, then  $y(w_i^1, w_i^2) = 1$  if  $w_i^1$  and  $w_i^2$  are the same, and  $y(w_i^1, w_i^2) = 0$  otherwise. The network prediction  $p$  is a decimal between 0 to 1, and we applied the binary cross-entropy to compute the loss:

$$\mathcal{L}(w_i^1, w_i^2) = y(w_i^1, w_i^2) \log(p) + (1 - y(w_i^1, w_i^2)) \log(1 - p) \quad (3)$$

### III. EXPERIMENTS AND ANALYSIS

To the best of our knowledge, our proposal is the first deep learning-based audio-in-image watermarking scheme, so it is difficult to find peer methods for a fair analogous study. We

only found a related handcrafted audio-in-image method [11] that has reported a Root Mean Square Error (RMSE) of 0.022325 averaged over four audio watermarks, whereas our method has an RMSE of 0.009452 averaged over 5,800 audio watermark samples. To further illustrate the effectiveness of the proposed scheme, we have conducted typical experiments and analyses in deep learning and image watermarking.

#### A. Data Preparation

To train the proposed WM Network, we rescaled (128 × 128) COCO Dataset [21] for the cover-image and resampled (8192 sampling rate) Speech Commands Dataset [22] audio clips for watermark audio. Combining these two datasets, we sampled 42,600, 9,700, and 5,800 audio-image pairs for training, validation, and testing. None of the training audio commands match with either validation or testing audio commands.

Speech Command Dataset samples are paired with extracted audio samples derived from the pre-trained WM Network with varying marked-image distortions to prepare the Similarity Network training dataset. This allowed us to generate 340,000, 64,000, and 46,000 audio-audio pairs for training, validation, and testing. The audio pairs are labeled as 0 or 1, where 0 means an audio pair is different and 1 means the same.

#### B. Training, Validation, and Testing Results

After preparing the datasets, we train the proposed scheme according to the strategies described in Section II-C. Fig. 3a shows the losses in training and validation of the WM Network, indicates a proper fit for hundreds of epochs. Fig. 3b shows the training and validation loss of the Similarity Network for 100 epochs, post training which achieved a validation accuracy of 99.44% on identifying whether a watermark, watermark extraction pair ( $W, W'$ ) is the same or not.

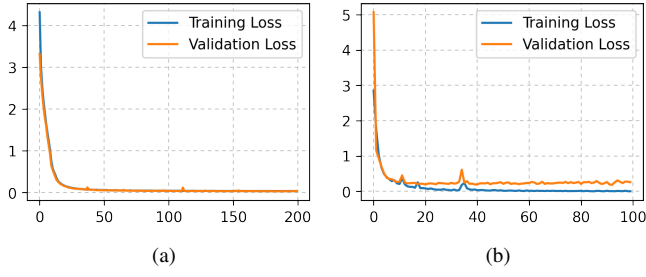


Fig. 3: (a) Learning Curves of the WM Network for 200 epochs, and (b) Learning Curve of the Similarity Network for 100 epochs.

At the testing phase of the WM Network, we calculated the extracted watermark's RMSE and the marked-image's Structural Similarity Index Measure (SSIM) [23] score with the original watermark and cover-image, respectively. At validation and testing, the watermark's RMSEs are 0.010664 and 0.009452, and the marked-image's SSIMs are 0.988365 and 0.988230, where we can observe successful extractions and high fidelity. Fig. 4 shows examples of cover-images, watermarks, marked-images, and extractions. To evaluate the Similarity Network, we used binary accuracy and the cross-entropy loss value. At validation and testing, the cross-entropy values are 0.29 and 0.42, and the accuracies are 99.33% and 98.98%. Thus, the Similarity Network has successfully learned to recognize the watermarks even under noisy situations.

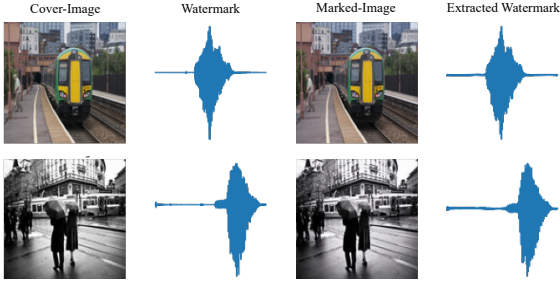


Fig. 4: WM Network sample inputs and outputs: (from left to right) cover-image, watermark, marked-image, and watermark extraction.

### C. Robustness

The encoder introduces some decomposition and redundancy by enlarging the 8,192 lengths  $128 \times 64$  watermark to a  $128 \times 128$  code. As a result, the WM Network can resist noise to an extent without presenting any adversarial examples while training. Moreover, because our watermark extraction is verified by the Similarity Network, which can tolerate some distortions, we analyzed the overall scheme's (WM Network + Similarity Network) robustness by plotting the accuracy of the Similarity Network versus swept-over noise parameters. To train the Similarity Network, we generate noisy audio samples derived from WM Network with marked images where 0 to 50% cutout (randomly removing regions) or  $-5^\circ$  to  $+5^\circ$  rotation is applied. We then tested the Similarity Network with watermark pairs within and beyond the range of its training noise parameters. E.g., we tested  $(W, W')$  pairs for up to 90% cutout noise exposure on the marked-images. Fig. 6a shows the accuracy of Similarity Network versus up to 90% cutout, and Fig. 6b shows the accuracy versus  $-6^\circ$  to  $+6^\circ$  rotation on

marked-images. On extracted watermark pair validation task, the model has a high and decent tolerance range against image processing and geometric rotation attacks, respectively.

### D. Ablation Study of the Similarity Network

This section demonstrates the indispensability of the Similarity Network by ablation study. If there are some noises on the marked-image, the watermark extraction can be noisy while retaining some resemblance with the original. However, simple error values cannot accurately indicate this similarity. Fig. 5 shows two such instances with  $(W, W')$  pairs, and each pair is audibly the same for human hearing despite having high RMSE values. Analytically, these values incorrectly indicate low similarity compared to the test set's low average RMSE of 0.009452 (considering as baseline). In contrast, the Similarity Network correctly outputs high similarity values.

For a large-scale experiment on the testing set, Fig. 6a and 6b show the RMSE and the Similarity Network accuracy versus the distortion parameters, where we find that under noise, the Similarity Network can capture the semantic similarity that is consistent with human judgments while the RMSE values are simply computing the differences.

Original Watermark	Extracted Watermark	RMS Error	Similarity Network
		0.0443456	0.99982
		0.0626943	1.0

Fig. 5: Examples of watermark and noisy extractions with RMSE and Similarity Net prediction.

## IV. CONCLUSION

This paper introduces a robust and blind audio-in-image watermarking scheme with deep learning. We designed a WM Network to fully learn the audio-in-image watermarking process without any supervision and a Similarity Network to enhance robustness from the semantic recognition perspective. In the future, our primary goal is to improve the robustness of the WM Network itself so that our proposed scheme can be resistant to different attacks, especially geometric distortions.

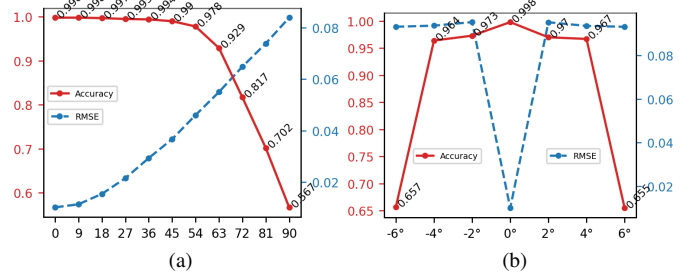


Fig. 6: Distortion parameters versus the Similarity Network accuracy and watermark RMSE, (a): % of cutout and (b): rotation degree.

## REFERENCES

- [1] X. Zhong, P. Huang, S. Mastorakis, and F. Y. Shih, "An automated and robust image watermarking scheme based on deep neural networks," *IEEE Transactions on Multimedia*, 2020.
- [2] H. Berghel and L. O'Gorman, "Protecting ownership rights through digital watermarking," *Computer*, vol. 29, no. 7, pp. 101–103, 1996.
- [3] M. Asikuzzaman, M. J. Alam, A. J. Lambert, and M. R. Pickering, "A blind high definition videowatermarking scheme robust to geometric and temporal synchronization attacks," in *2013 Visual Communications and Image Processing (VCIP)*, 2013, pp. 1–6.
- [4] M. Sharifzadeh, C. Agarwal, M. Aloraini, and D. Schonfeld, "Convolutional neural network steganalysis's application to steganography," in *2017 IEEE Visual Communications and Image Processing (VCIP)*, 2017, pp. 1–4.
- [5] D. Hou, W. Zhang, J. Liu, S. Zhou, D. Chen, and N. Yu, "Emerging applications of reversible data hiding," in *Proceedings of the 2nd International Conference on Image and Graphics Processing*, ser. ICIGP '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 105–109. [Online]. Available: <https://doi.org/10.1145/3313950.3313952>
- [6] L. Dovydaitis, T. Rasmay, and V. Rudzionis, "Speaker authentication system based on voice biometrics and speech recognition," in *Business Information Systems Workshops*, W. Abramowicz, R. Alt, and B. Franczyk, Eds. Cham: Springer International Publishing, 2017, pp. 79–84.
- [7] S. Mastorakis, X. Zhong, P. Huang, and R. Tourani, "DLWIoT: Deep learning-based watermarking for authorized iot onboarding," in *2021 IEEE 18th Annual Consumer Communications Networking Conference (CCNC)*, 2021, pp. 1–7.
- [8] V. Potdar, S. Han, and E. Chang, "A survey of digital image watermarking techniques," in *INDIN '05. 2005 3rd IEEE International Conference on Industrial Informatics*, 2005., 2005, pp. 709–716.
- [9] Z. Chen, L. Li, H. Peng, Y. Liu, and Y. Yang, "A novel digital watermarking based on general non-negative matrix factorization," *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 1973–1986, 2018.
- [10] A. Kanhe and A. Gnanasekaran, "Robust image-in-audio watermarking technique based on DCT-SVD transform," *EURASIP J. Audio Speech Music Process.*, vol. 2018, no. 1, Dec. 2018.
- [11] M. A. Bhat, P. G. Arfaat, and S. M. Hussain, "Audio watermarking in images using wavelet transform," *IJCST*, vol. 2, pp. 405–411, 01 2011.
- [12] H. Kandi, D. Mishra, and S. R. S. Gorthi, "Exploring the learning capabilities of convolutional neural networks for robust image watermarking," *Computers & Security*, vol. 65, pp. 247–268, 2017. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0167404816301699>
- [13] D. Li, L. Deng, B. Bhooshan Gupta, H. Wang, and C. Choi, "A novel CNN based security guaranteed image watermarking generation scenario for smart city applications," *Information Sciences*, vol. 479, pp. 432–447, 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0020025518301452>
- [14] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *2016 IEEE European Symposium on Security and Privacy (EuroSP)*, 2016, pp. 372–387.
- [15] C. Yi, H. Li, R. Wan, and A. C. Kot, "Improving robustness of DNNs against common corruptions via gaussian adversarial training," in *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, 2020, pp. 17–20.
- [16] S. Mun, S. Nam, H. Jang, D. Kim, and H. Lee, "Finding robust domain from attacks: A learning framework for blind watermarking," *Neurocomputing*, vol. 337, pp. 191–202, 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0925231219300955>
- [17] Y. Chen, T. Fan, and H. Chao, "WMNet: A lossless watermarking technique using deep learning for medical image authentication," *Electronics*, vol. 10, no. 8, 2021. [Online]. Available: <https://www.mdpi.com/2079-9292/10/8/932>
- [18] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. Efros, "Context encoders: Feature learning by inpainting," in *Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [20] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, vol. 2. Lille, 2015.
- [21] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 740–755.
- [22] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *ArXiv e-prints*, Apr. 2018. [Online]. Available: <https://arxiv.org/abs/1804.03209>
- [23] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.