Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.055150

REVIEW

Robust Deep Image Watermarking: A Survey

Yuanjing Luo, Xichen Tan and Zhiping Cai*

College of Computer, National University of Defense Technology, Changsha, 410005, China

*Corresponding Author: Zhiping Cai. Email: zpcai@nudt.edu.cn

Received: 19 June 2024 Accepted: 03 September 2024 Published: 15 October 2024

ABSTRACT

In the era of internet proliferation, safeguarding digital media copyright and integrity, especially for images, is imperative. Digital watermarking stands out as a pivotal solution for image security. With the advent of deep learning, watermarking has seen significant advancements. Our review focuses on the innovative deep watermarking approaches that employ neural networks to identify robust embedding spaces, resilient to various attacks. These methods, characterized by a streamlined encoder-decoder architecture, have shown enhanced performance through the incorporation of novel training modules. This article offers an in-depth analysis of deep watermarking's core technologies, current status, and prospective trajectories, evaluating recent scholarly contributions across diverse frameworks. It concludes with an overview of the technical hurdles and prospects, providing essential insights for ongoing and future research endeavors in digital image watermarking.

KEYWORDS

Deep image watermarking; multimedia security; data protection deep neural network

1 Introduction

In the digital age, with the rapid and widespread development of the internet and network technologies, multimedia content has proliferated extensively [1]. Protecting the copyright and integrity of digital media, especially widely disseminated images, has become crucial [2,3]. To address this issue, digital watermarking technology was proposed to address this issue. Unlike steganography, which also belongs to the field of information hiding, the core of digital watermarking technology lies in providing technical support for copyright protection, content authentication, and integrity verification. By cleverly embedding text, images, or logos into digital media such as pictures, audio, or video, this technology ensures that the information can be detected and verified when necessary. The watermark can be visible or invisible, but it is designed to ensure its detectability during the retrieval process. Steganography, on the other hand, focuses on hiding the existence of information itself, mainly used for secure communication and privacy protection. It achieves this by embedding information into the redundant or less noticeable parts of media files, minimizing the perceptual impact on the original media [4].

Over the past two decades, numerous traditional watermarking methods have been developed to safeguard image information [5,6]. These methods embed signatures into spatial or frequency domain feature regions, achieving effective imperceptibility and robustness against attacks, thereby providing basic copyright protection [7]. However, these traditional techniques face challenges in handling modern complex attacks and maintaining content quality. They are vulnerable to operations like compression, cropping, and noise addition, often resulting in significant image quality degradation during watermark embedding and extraction [8]. Traditional watermarking methods may require expertise not only in the watermarking domain but also in cryptographic algorithms to ensure the security and robustness of the watermark. This additional layer of complexity can make the application of traditional methods more challenging and less accessible to those without a background in cryptography [9].

In recent years, the development of deep learning technology has brought new breakthroughs to watermarking techniques. Deep learning-based watermarking schemes, such as CNN-based models, often simplify the process by automating feature extraction and embedding decisions, reducing the need for extensive cryptographic knowledge and making the application more straightforward for a wider range of researchers. Recent studies have shown that machines can replace manual methods for finding embedding spaces [10]. By using neural networks to locate stable embedding spaces and resist various attacks, robust watermark encoders and decoders can be constructed [11]. These deep learning-based watermarking algorithms, collectively known as deep watermarking, have garnered significant attention since their inception [11–13]. Compared to traditional algorithms, deep watermarking frameworks are simpler, utilizing encoder-decoder structures and enhancing performance by integrating innovative modules during model training. The process begins with the input of the watermark and the original image, where the encoder generates an encoded image that is visually indistinguishable from the original but hides the watermark, which the decoder can accurately retrieve. For example, HiDDeN [12] pioneered the autoencoder architecture by introducing joint training of the encoder and decoder along with a noise layer. Building on the HiDDeN framework, many advanced autoencoder watermarking schemes have been developed [14–16], enhancing imperceptibility and robustness [17–19]. UDH [20] further simplified the structure, making the encoder input only watermark-related, fostering exploration and innovation in watermarking research [21]. Recently, Invertible Neural Networks (INN) have gained attention for their ability to facilitate reversible image transformations by learning stable, invertible mappings between data and latent distributions [22,23]. INN has shown excellent performance in many image-centric applications, including digital watermarking. The process of embedding and recovering watermarks can be conceptualized as reversible operations performed by the forward and backward functions of INN. This method achieves more covert digital watermark embedding [24–28]. With its good performance, deep watermarking technology excels in applications focused on digital rights management (DRM). In DRM, deep watermarking effectively tracks and verifies the legitimacy of digital content. This technology ensures that digital media is protected by embedding watermarks that are both imperceptible and resilient against various attacks, allowing for accurate identification and validation of ownership and authenticity.

In summary, deep watermarking technology leverages deep learning models, particularly the encoder-decoder structure, to overcome many limitations of traditional watermarking techniques, offering greater flexibility and performance. This paper provides a comprehensive review of deep watermarking technology (primarily deep image watermarking), covering its core technologies, current developments, and future directions, thereby offering valuable insights for further research and applications in this field. The main contributions of this paper are as follows:

- **Overview of Existing Technologies:** A brief introduction and summary of existing deep watermarking technologies to give readers a basic understanding.

- **Detailed Review and Comparison:** An in-depth review of the strengths and weaknesses of current deep watermarking schemes based on CNN and INN paradigms, facilitating a comprehensive understanding of the characteristics of different frameworks and enabling readers to select suitable schemes based on their needs.
- **Technical Challenges:** A summary of the main technical challenges faced by current deep watermarking technologies, providing readers with a clear understanding of the practical difficulties associated with deep watermarking.
- **Future Directions:** Proposing potential future directions for deep watermarking, offering guidance for further research and application in this field.

The remainder of this paper is organized as follows: Section 2 introduces the general concept of watermarking, providing background knowledge; Section 3 reviews traditional image watermarking techniques, analyzing their limitations and shortcomings; Section 4 presents the basic framework and processes of deep watermarking, explaining its working principles; Section 5 reviews current deep watermarking schemes based on different paradigms, discussing their respective strengths and weaknesses; Section 6 summarizes the technical challenges and obstacles of existing deep watermarking technologies, analyzing their causes and impacts; Section 7 proposes future directions for deep watermarking, exploring potential research and application areas; and finally, Section 8 concludes the paper.

2 Theoretical Basis of Image Watermarking

2.1 Definition

Originally, the concept of "digital watermarking" emerged in the early 1990s as a technique to embed information into digital media such as images, audio, or video for copyright protection, content integrity verification, and data security [29]. Watermarks can be visible or invisible but are usually designed to be detectable during retrieval, serving as a deterrent against unauthorized use. This distinguishes watermarking from steganography, where the primary objective is to conceal the very existence of the message. When the embedded carrier is an image, it is referred to as image watermarking, representing the application of digital watermarking in static images. This process involves mathematical operations where the original image W is altered to include a watermark C based on a specific embedding function F and a predefined strength α:

$$W' = W + \alpha \times F(W, C). \tag{1}$$

These watermarks, which can be text, images, or other data, are usually imperceptible and embedded in a way that does not significantly alter the original content. The extraction process, represented by function G and potentially requiring a key K, allows for the retrieval of the watermark $\hat{C}$ to confirm content legitimacy and prevent unauthorized use [30]:

$$\hat{C} = G(W', K). \tag{2}$$

Image watermarking is widely used in copyright protection, content authentication, and anti-counterfeiting, providing a means to securely embed and extract watermarks within digital images.

2.2 *Requirements*

1) Imperceptibility.

The watermark embedded as additional information into the carrier media can be visible or invisible [31]. Imperceptibility is crucial in evaluating watermarking systems, ensuring watermarked images appear identical to the original. This means any embedded watermark should be invisible to human eyes, even with minor changes in brightness or contrast. Various methods assess imperceptibility, including the Peak Signal to Noise Ratio (PSNR) [32], Structural Similarity Index (SSIM) [33] and Learned Perceptual Image Patch Similarity (LPIPS) [34], where PSNR is an objective criterion for evaluating image quality, which can be defined as follows:

$$PSNR_{(x,y)} = 10\log_{10}\left(\frac{(MAX_I)^2}{MSE\,(x,y)}\right), \tag{3}$$

where MAX_I represents the maximum possible pixel value of images x and y. $MSE\,(x,y)$ denotes the Mean Squared Error (MSE) calculated between images x and y:

$$MSE = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\|X\,(i,j) - Y\,(i,j)\|^2. \tag{4}$$

$SSIM$ is used to measure the similarity between images x and y, which can be calculated as follows:

$$SSIM\,(x,y) = \frac{\left(2\mu_x\mu_y + C_1\right)\left(2\sigma_{xy} + C_2\right)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)}, \tag{5}$$

where μ_x and μ_y represent the average grey values of images, i.e., are the mean of X and Y. Symbol σ_x and σ_y represent the variances of X and Y. Symbol σ_{xy} represents covariance. $C_1 = (k_1L)^2$ and $C_2 = (k_2L)^2$ are two constants which are used to maintain stability when either $\mu_x^2 + \mu_y^2$ or $\sigma_x^2 + \sigma_y^2$ is very close to 0, where $K_1 = 0.01$ and $K_2 = 0.03$. L is the dynamic range of the pixel values.

LPIPS measures the distance between the deep features of images x and y using a CNN, which can be calculated as follows:

$$LPIPS\,(x,y) = \sum\nolimits_{i=1}^{N} d_F\,(f_i\,(x)\,,f_i\,(y)), \tag{6}$$

where f_i represents the feature extraction at layer i of a pre-trained deep neural network, d_F denotes a distance metric in the feature space, and N is the total number of layers considered.

High PSNR&SSIM and low LPIPS indicate better imperceptibility, ensuring no visible difference between the original and watermarked image. Techniques selecting optimal regions of the cover image for watermark insertion further enhance imperceptibility.

2) Robustness.

Robustness ensures a watermark remains detectable after common image processing operations such as compression, scaling, and rotation. Techniques achieving high robustness include redundant embedding and spread spectrum methods. A robust watermark prevents unauthorized removal and maintains integrity through various attacks, making it suitable for copyright protection and broadcast monitoring [35]. Fragile watermarking, conversely, verify content integrity by revealing tampering. Semi-fragile watermarking withstands some transformations but fail against malicious attacks, used for image authentication [36]. When the embedded watermark is an image, robustness can also be evaluated using PSNR and SSIM, by comparing the similarity between the extracted watermark and the original watermark. When the embedded watermark is a string of characters, robustness is evaluated using the bit error rate (BER) and normalized correlation (NC). These two metrics are calculated to assess the similarity between the embedded watermark and the extracted watermark after applying different attacks to the embedded content, calculated as:

$$BER\left(S, S'\right) = \frac{\sum_i Berr_i}{\sum_i Bstr_i}, \tag{7}$$

$$NC\left(S, S'\right) = \frac{1}{W \times H} \sum_{W-1}^{i=0} \sum_{H-1}^{j=0} \delta\left(S_{i,j}, S'_{i,j}\right), \tag{8}$$

$$\delta\left(S_{i,j}, S'_{i,j}\right) = \begin{cases} 1, & S = S' \\ 0, & otherwise \end{cases}, \tag{9}$$

where $\sum_i Berr_i$ is the number of error bits, $\sum_i Bstr_i$ represents the length of hidden messages. The BER value ranges from 0 to 1. A BER of 0 indicates perfect watermark extraction, meaning the extracted watermark bits are identical to the embedded ones. Conversely, a BER of 1 signifies a complete mismatch between the extracted and embedded watermark bits. While NC is a value in the range [0,1] where a higher value proves a better similarity between media.

3) Other requirements.

Essentially, the design of watermarking methods necessitates both imperceptibility and robustness. Additionally, specific requirements may vary depending on the intended application. Robustness ensures a watermark remains detectable after common image processing operations such as compression, scaling, and rotation.

Security. The security of a watermark pertains to its capacity to withstand hostile attacks, which are any processes intentionally designed to undermine the watermark's purpose. A watermark is considered secure if an unauthorized user/hacker cannot remove it without full knowledge of the embedding algorithm, detector, and watermark composition. Only authorized parties should access the watermark, typically secured through cryptographic keys [37,38]. This ensures that only authorized users can legally detect, extract, or modify the watermark, maintaining its integrity and protection [39]. However, there are still methods that can remove the watermark without prior knowledge. For example, an unauthorized watermark might be embedded into already watermarked media, allowing an illicit user to falsely claim ownership of the content.

Capacity. Capacity refers to the amount of watermark that can be embedded in an image. There is an inherent trade-off: higher capacity often results in decreased image quality and reduced robustness of the method [40]. Therefore, capacity typically represents the maximum quantity of watermark that can be discreetly embedded in the image, while maintaining an optimal balance between imperceptibility and robustness.

In all of the above metrics, there is a complex interplay among the robustness, imperceptibility, and capacity of digital watermarks, which can be illustrated in Fig. 1. When any one of these parameters is fixed, the other two often exhibit a competitive relationship. For example, if the capacity of the watermark is set, enhancing its robustness might require increasing the embedding strength, but this generally leads to significant image distortion. Conversely, reducing the watermark's embedding strength to maintain high image quality can compromise its robustness. Therefore, when designing watermarking algorithms, it is usually necessary to find a balance among robustness, imperceptibility, and capacity, making appropriate adjustments based on specific application requirements.

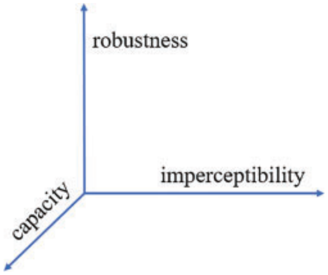

Figure 1: The relationship between watermark performance indicators

2.3 *Classification*

There are various methods for classifying digital watermarking (see Fig. 2), stemming from different criteria, leading to distinct yet interconnected classifications.

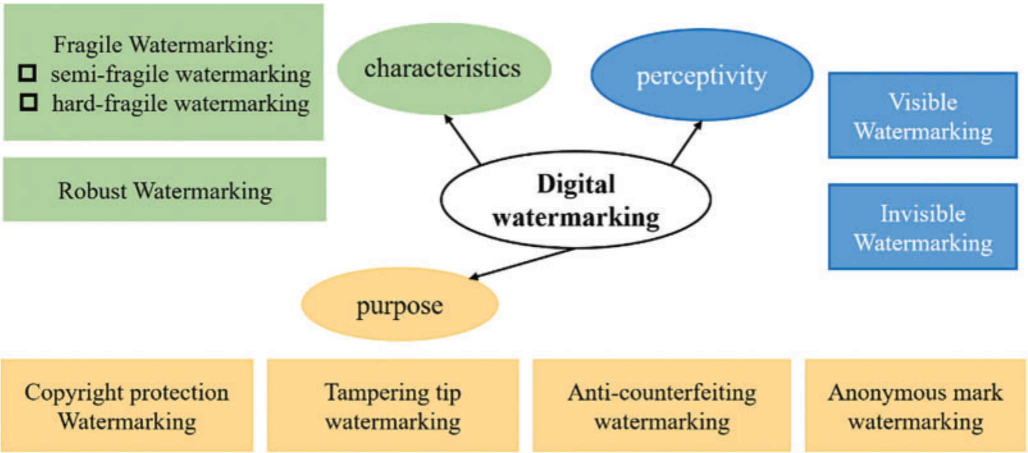
Fragile Watermarking:
□ semi-fragile watermarking
□ hard-fragile watermarking
Robust Watermarking
characteristics
perceptivity
Visible Watermarking
Invisible Watermarking
Digital watermarking
purpose
Copyright protection Watermarking
Tampering tip watermarking
Anti-counterfeiting watermarking
Anonymous mark watermarking

Figure 2: The classification of digital watermarking

1) According to characteristics.

- Fragile Watermarking: these detect modifications to the host image, indicating whether alterations have occurred [41]. Fragile watermarking can be divided into semi-fragile and hard-fragile categories. Semi-fragile watermarking offers soft authentication with relaxed integrity criteria, allowing for minor modifications. In contrast, hard fragile watermarking enforces strict integrity criteria, resisting all modifications.
- Robust Watermarking: these remain detectable even after non-malicious or certain levels of malicious distortions, ensuring the watermark can still be extracted [42].

2) According to perceptivity.

- Visible Watermarking is perceptible, akin to marks overlaid on an image, similar to paper watermarks [43]. They are primarily used in images and videos to visually indicate ownership, preventing unauthorized commercial use. This is useful for previews in image or video databases available on the Internet.
- Invisible Watermarking, more widely used than visible ones, is not perceptible to the human eye. They can be extracted to prove ownership or integrity when needed [44].

3) According to purpose.

In practice, different watermarking technologies serve various purposes:

- Copyright protection watermarking allows visibility of the watermark on the image, ensuring it remains even after attacks.
- Tampering tip watermarking preserves image integrity, highlights modifications, and resists lossy compression.
- Anti-counterfeiting watermarking is embedded during paper note production and remains detectable after printing and scanning.
- Anonymous mark watermarking conceals important annotations on confidential data, preventing unauthorized access.

2.4 Attacks

In the realm of digital watermarking, attacks constitute a significant research topic, as they pose challenges to the security and efficacy of watermarking technologies. Attacks may aim to disrupt or tamper with the embedded watermark, thereby affecting its detection and authentication capabilities, and impacting its application in areas such as copyright protection and identity verification. The types of attacks generally include:

1) Geometric Attacks.

Geometric attacks are a class of transformations that alter the spatial properties of an image. They can be divided into two main types:

Global Geometric Attacks: These involve operations that affect the entire image uniformly. The most common global attacks include:

- Rotation: The image is rotated by a certain degree, which can disrupt the alignment of features within the image.
- Scaling: The image is resized, either enlarging or reducing its dimensions, potentially causing loss of detail or aliasing effects.
- Translation: The image is shifted in the horizontal or vertical direction, moving features from one part of the image to another.

Local Geometric Attacks: These are more targeted and affect only specific regions within an image. Examples of local geometric attacks are:

- Random Bending Attacks: Portions of the image are warped or bent, which can obscure local features.
- Cropping: Parts of the image are removed, either randomly or systematically, reducing the available image content.
- Row/Column Deletion: Specific rows or columns of pixels are removed, which can lead to a loss of structural information.

2) Signal Processing Attacks.

Signal-processing attacks involve the application of various signal-processing techniques that can potentially degrade the quality or alter the characteristics of an image. These attacks include:

- Compression: Reducing the file size of an image, which can lead to loss of detail or artifacts, especially in lossy compression methods.

- Filtering: Applying filters to the image that can blur details, enhance edges, or remove noise, thus affecting the image's content.
- Printing and Scanning: When an image is printed and then rescanned, the process can introduce artifacts and degrade the image quality, impacting any hidden information.

3) Cross-Media Attacks.

Cross-media attacks encompass a range of challenges that affect the integrity and detection of digital watermarks when content transitions between different media formats. These attacks include, but are not limited to:

Screen Capture Attacks: Involving the process of taking a screenshot or photograph of a digital display, these attacks can lead to:

- Resolution Loss: The captured image may have a different resolution compared to the original, leading to a loss of detail.
- Color Distortion: Differences in display technology can result in color shifts or inaccuracies in the captured image.
- Reflections and Glare: Physical properties of the screen can cause reflections or glare those obscure parts of the image.
- Compression Artifacts: If the screenshot is saved in a compressed format, additional artifacts may be introduced due to the compression process.

Printer-Camera Attacks: The act of printing a watermarked image and then capturing it with a camera introduces challenges such as:

- Degradation of image quality due to the printing process.
- Potential loss of watermark visibility when re-photographed.

Scanner-Printer Attacks: Similar to printer-camera attacks, the process of scanning and reprinting can lead to loss of detail and introduction of noise that may interfere with the watermark's integrity.

Cross-Media Distortions: General distortions that occur when digital content is transferred across different media, including but not limited to, the aforementioned attacks.

These attacks can significantly affect the robustness of steganographic methods, as they must be designed to withstand such transformations while still allowing for the reliable extraction of hidden information.

3 Traditional Image Watermarking and Limitations

Traditional image watermarking techniques fall into two categories: spatial domain and transform domain methods [6,45]. Spatial domain methods operate directly on the image pixels. One common technique is Least Significant Bit (LSB) Embedding, where watermarks are embedded by modifying the image's least significant bits [46]. This method is simple but vulnerable to image processing attacks like compression and cropping. Error Control Coding (ECC) is also used to enhance the robustness against interference [47]. Transform domain methods involve converting the image to a frequency or other transform domain before embedding the watermark. The Discrete Cosine Transform (DCT) is prevalent in JPEG compression and modifies DCT coefficients to embed watermarks, offering good resistance to compression attacks [5,7]. The Discrete Wavelet Transform (DWT) is notable for its ability to withstand scaling and noise addition by embedding watermarks in wavelet coefficients [48]. Discrete Fourier Transform (DFT) [49] and Singular Value Decomposition (SVD) [50] are also used, providing robustness against geometric transformations and maintaining image quality while embedding watermarks. Each method provides varying levels of security and robustness, tailored to specific application needs [51].

Traditional digital watermarking employs manual design for embedding algorithms and utilizes both traditional and deep learning features for extraction. Despite their initial effectiveness, these techniques encounter significant challenges that affect their efficacy and practicality [8]:

1) Attack Resistance Limitations. The robustness of traditional watermarking heavily relies on the chosen algorithm and its parameters. Although some specialized tools exhibit robustness against image manipulations, the overarching trend indicates a susceptibility inherent in many traditional methods. Some traditional approaches are vulnerable to image processing like compression and cropping. Even the more robust transform domain methods can be compromised by geometric attacks such as rotation and scaling [6–8].

2) Impact on Image Quality. Watermark embedding can degrade the perceptual quality of images, particularly when using stronger embedding parameters. This also makes the lossless recovery of the original image challenging, diminishing the methods' practical utility.

3) Scalability and Adaptability Issues. Traditional methods struggle to adapt to new digital formats and require ongoing updates to cope with technological advancements and new security threats.

The limitations of traditional watermarking methods can be attributed to several underlying reasons:

1) Technological Evolution. The rapid pace of technological change often outstrips the ability of traditional methods to adapt, leading to obsolescence as new media formats and compression techniques emerge.

2) Sophistication of Attacks. As attackers become more sophisticated, traditional methods may not be equipped to handle complex and targeted attacks designed to defeat older watermarking techniques.

3) Fixed Rulesets. Traditional watermarking typically relies on static, predefined rules for embedding and extraction, which lack the flexibility to respond to varied and unpredictable real-world conditions.

4) Lack of Learning Capabilities. Unlike modern deep learning approaches, traditional watermarking lacks the ability to learn from data and improve over time, which is crucial for adapting to new challenges.

The advent of deep learning has revolutionized watermarking algorithms by making deep learning features a new standard. Deep neural networks, with their superior feature fitting capabilities, excel in embedding watermarks that are both robust and visually lossless. This integration enhances the seamless coordination between the embedding and extraction processes, offering a promising path forward for overcoming the limitations of traditional watermarking methods.

4 Emerging Deep Watermarking

4.1 General Framework

Deep watermarking typically employs encoders based on deep neural network architectures, such as Convolutional Neural Networks (CNNs) or Invertible Neural Networks (INNs), to train models that embed watermarks robustly and imperceptibly into host images. In contrast to traditional watermarking techniques, which rely on experts to develop embedding methods, deep watermarking leverages retraining to counter various attacks. The foundational framework of deep watermarking encompasses the following components:

- **Encoder:** This component is responsible for embedding the watermark information into the host media, resulting in media that carries the watermark (i.e., images).
- **Noise Layer:** It simulates potential signal distortions, such as JPEG compression and Gaussian noise, thereby enhancing the robustness of the watermark.
- **Decoder:** This extracts the watermark information from the watermarked media, serving to verify copyright or detect tampering.
- **Loss Function:** It is utilized during the training process to optimize the performance of the encoder and decoder, ensuring the imperceptibility and robustness of the watermark.

In addition to these core frameworks, the performance of the watermarking system is further enhanced by incorporating the following supplementary components:

- **Data Preprocessing and Enhancement:** Before watermark embedding and extraction, the raw data may undergo preprocessing steps such as noise reduction and contrast enhancement to improve the robustness and stealth of the watermark.
- **Feature Learning:** Deep learning models autonomously learn feature representations of the host media, which are then employed for watermark embedding and extraction.
- **Multimodal Fusion:** In scenarios involving multiple types of media data (e.g., images and videos), the system must be capable of processing and integrating information from diverse modalities.
- **Adaptive Embedding:** The embedding strategy is dynamically adjusted based on the characteristics of the media content, maximizing the utilization of the host media's capacity.
- **Security Enhancement:** Beyond adversarial training, additional techniques such as encryption and obfuscation may be implemented to strengthen the security of the watermarking system.

This comprehensive approach ensures that deep watermarking can effectively protect digital media against unauthorized use while maintaining high-quality visual output.

4.2 Deep Watermarking Process

The deep watermarking scheme is decomposed into three main stages. The first stage is the encoder, which trains the encoder network to embed the input message into the original content. The primary objective is to minimize the objective function. This function calculates the difference between the original content and the watermarked content, as well as the difference between the embedded and extracted signatures. The second stage conducts attack simulations by applying various attacks to the watermarked content through a distortion layer. These attacks may include different forms of manipulation, such as cropping and compression. Finally, the decoder utilizes the decoder network to extract the embedded information from the distorted content. Due to the iterative learning process, the embedding is more robust against attacks in the second stage, and the extraction network enhances the integrity of the extracted watermark.

As shown in Fig. 3, the complete watermark embedding process includes: 1) Watermark information encoding: The information to be hidden is encoded into a format suitable for embedding, which may include binary sequences, images, or audio signals, etc. 2) Host media analysis: Analyze the characteristics of the host media, such as texture, color distribution, etc., to determine the best embedding location and strategy. 3) Watermark embedding strategy: Based on the analysis results, select the most appropriate watermark embedding strategy. 4) Watermark embedding: Embed the watermark information into the carrier image features through the constructed encoder. 5) Quality control: Monitor the quality of the host media in real-time during the embedding process to ensure that the addition of the watermark does not significantly degrade the media quality.

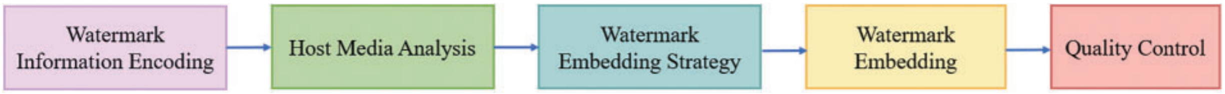

Figure 3: The watermark embedding process

As shown in Fig. 4, the complete watermark extraction process includes: 1) Damage assessment: Before extraction, assess the potential damage or distortion, the host media may have suffered to take appropriate recovery measures. 2) Preprocessing: Preprocess the watermarked media that may have been distorted, such as denoising, color space conversion, etc. 3) Watermark extraction: Extract the watermark information from the carrier image through the decoder. 4) Watermark decoding: Restore the extracted watermark to the format before embedding, which may involve specific decoding algorithms. 5) Integrity verification: Verify the extracted watermark information to confirm its consistency with the original watermark, thereby proving the ownership or integrity of the media.

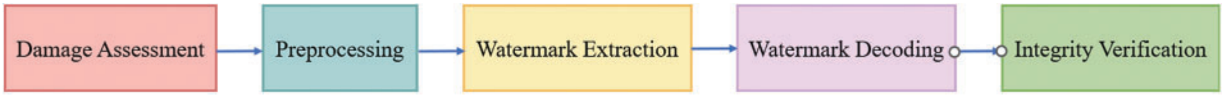

Figure 4: The watermark extraction process

Additionally, the performance of the watermark can be enhanced in the following ways: 1) Embed and extract the watermark at different scales to combat various potential attacks and distortions. 2) Error correction: Introduce error-correcting codes (such as Reed-Solomon codes) to improve the robustness of the watermark system, allowing the watermark information to be recovered even in cases of partial damage. 3) Privacy protection: Ensure that the watermarking process does not disclose sensitive information of the host media. 4) Adversarial robustness: Improve the watermark system's defense against advanced attacks through adversarial training and model robustness research.

5 Review of the Existing Methods

5.1 Methodology

To conduct a comprehensive review of the existing work on deep image watermarking, we meticulously devised the following research plan:

- **Exhaustive Literature Search:** We carried out a thorough literature search for studies directly related to deep image watermarking across multiple authoritative databases, including PubMed, IEEE Xplore, and Google Scholar, ensuring a wide and diverse range of literature sources. Additionally, we manually reviewed the reference lists of included studies and examined related citation records to identify any potentially relevant literature.
- **Stringent Inclusion Criteria:** To ensure the quality and rigor of our review, we conducted a detailed quality assessment of the retrieved literature. We excluded studies that were not related to the theme of deep image watermarking or focused on other forms of digital watermarking, as well as literature from lower-quality journals. We prioritized literature published in high-impact SCI Q1 journals and top-tier CCF-A conferences.

- **Literature Data Extraction:** After selecting the studies, we extracted key data, including authors, publication year, research methods, main contributions, and experimental data. These data were systematically organized into tables for easy comparison among different papers.
- **Literature Summary and Analysis:** Using a narrative synthesis approach, we discussed the common themes and trends among these literatures and conducted a systematic review. Our review aims to provide profound and valuable insights for researchers and practitioners in the field of deep image watermarking.

5.2 Overall

Through the above methodologies, we find that existing deep image watermarking schemes are primarily categorized into two paradigms: the auto-encoder-based paradigm utilizing Convolutional Neural Networks (CNNs) as the backbone network, and the normalizing flow-based paradigm employing Invertible Neural Networks (INNs) as the core architecture.

The CNN-based approach leverages CNNs to train an encoder network and a decoder network for the concealment and recovery of information. A notable advantage of this method is the capacity to employ adversarial networks to enhance the visual quality of the watermarked images [52]. However, a significant drawback is that in these techniques, the encoding and decoding processes are performed sequentially by two non-shared forward networks, leading to the loss of critical information during the forward propagation. Current CNN-based methods often struggle to strike a balance between achieving high-quality watermarked images and the faithful recovery of watermarks, potentially resulting in color distortion and texture-replication artifacts. In contrast, the INN-based scheme processes the hiding and revealing of images through a single invertible neural network, which implies that the entire set of network parameters for both processes can be acquired through a single training session, enabling more imperceptible watermark embedding to compensate for the shortcomings of CNN-based schemes [26]. Nevertheless, the INN-based approach is heavily reliant on its invertibility and cannot incorporate adversarial perturbation training arbitrarily within the network; such distortions may lead to a significant degradation in performance [9].

We initially summarize the main differences between these two paradigms in Table 1, followed by a detailed review of existing schemes in the subsequent two subsections.

Table 1: Comparison table between CNN-based and INN-based deep image watermarking

	Auto-encoder-based method	Normalizing flow-based method
Network structure	CNNs	INNs
Training process	Requires separate training of the encoder and decoder	Trains the network for both hiding and revealing parameters
Running process	Sequentially by two non-shared forward networks	Through the forward and backward passes of the same network

(Continued)

Table 1 (continued)

	Auto-encoder-based method	Normalizing flow-based method
Advantages	1) The use of adversarial networks allows for the improvement of the visual quality of watermarked images, making them more aesthetically pleasing and less noticeable. 2) CNNs can be adapted to various image types and scenarios due to their powerful feature extraction capabilities.	1) Processing is done through a single invertible neural network, which simplifies the training process and allows for a unified set of parameters for both hiding and revealing. 2) The INN-based approach can achieve more imperceptible watermark embedding, which is beneficial for maintaining the original image quality.
Disadvantages	1) The encoding and decoding are performed by separate networks, which can lead to a loss of critical information during the process.	1) The method heavily relies on the network's ability to be invertible, which can be a limitation if the network structure becomes too complex or if adversarial training is required.
	2) Striking a balance between high-quality watermarked images and accurate watermark recovery can be challenging, potentially leading to issues like color distortion and texture replication.	2) The INN-based approach may not easily incorporate adversarial perturbation training, which could be necessary for robustness against certain types of attacks, potentially leading to performance degradation.

5.3 CNN-Based Deep Watermarking

1) Prior Knowledge

Convolutional Neural Networks (CNNs) are extensively applied to a variety of tasks, including image classification and recognition, due to their high efficiency in data representation with a constrained number of parameters. At their core, CNNs represent a specialized form of multi-layer perceptron, designed primarily for extracting and identifying the intricate features of two-dimensional images. As illustrated in Fig. 5, the CNN architecture typically consists of several layers, including: Input Layer, serves as the entry point for the network, typically representing a pixel matrix of an image. The depth corresponds to color channels, with 1 for grayscale and 3 for RGB. Convolutional Layer, a key component of CNNs, where each neuron responds to a small region of the previous layer, resulting in deeper matrices. Pooling Layer (Pooling), maintains matrix depth while reducing its size, effectively downscaling the image resolution. It decreases the number of neurons in fully connected layers, thus reducing network parameters, with no trainable parameters itself. Fully Connected Layer, concludes the CNN with 1–2 layers, using SoftMax for multi-class problems to output the probability distribution across categories. After convolutional and pooling layers abstract image information into higher-level features, fully connected layers complete the classification task. In recent years, with the swift advancement of machine learning tools, especially deep networks across various domains of computer vision and image processing, such as information hiding. Baluja [52] attempted to place a full-size color image within another image of the same size using CNN structure in 2017. In this work, CNNs are simultaneously trained to create the hiding and revealing processes and are designed to specifically work as a pair. He further presented a system to hide a full color image inside another of the same size with minimal quality loss to either image [53]. The system is trained on images drawn randomly from the ImageNet database, and works well on natural images from a wide variety of sources.

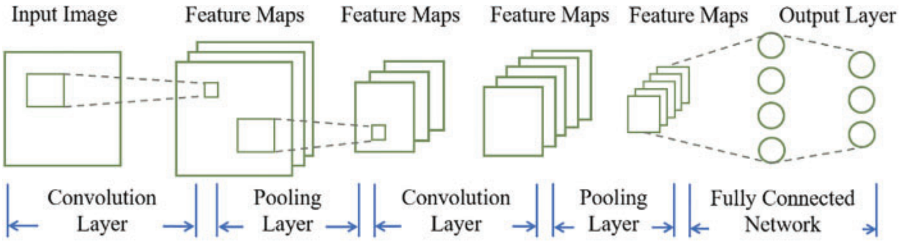

Figure 5: The basic structure of Convolutional Neural Networks (CNNs)

2) Main Schemes

Based on the successful application of deep learning to hiding images, CNNs have been increasingly applied to watermarking tasks. In 2018, Zhu et al. introduced HiDDeN, a pioneering deep watermarking framework utilizing encoder-decoder architectures [12]. This CNN-driven approach (also known as auto-encoder framework), depicted in Fig. 6, offers flexible trade-offs among capacity, secrecy, and noise robustness by adjusting training parameters or noise layers. HiDDeN outperformed traditional watermarking methods, both quantitatively and qualitatively, and its end-to-end design allows for easy adaptation to new distortions without the need for specialized algorithm redesign.

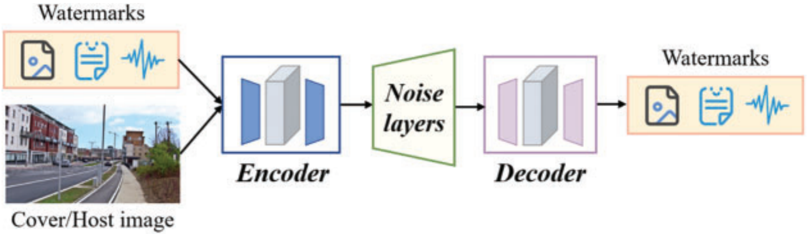

Figure 6: General framework of dependent CNN-based deep image watermarking

Building on this, in 2019, Liu et al. [14] proposed a novel two-stage separable deep learning (named TSDL) framework for practical blind watermarking to improve the practicability and robustness of algorithms. Specifically, the TSDL framework incorporates two distinct training components: noise-free end-to-end adversary training (FEAT) and noise-aware decoder-only training (ADOT). In FEAT, a sophisticated multi-layer feature encoding network is designed to develop the encoder, while ADOT focuses on creating a decoder that is robust and versatile enough to handle various types of noise. Extensive testing reveals that this framework not only achieves superior stability, enhanced performance, and quicker convergence than existing state-of-the-art OET methods but also effectively withstands high-intensity noises not previously examined in earlier studies.

Luo et al. proposed a distortion-agnostic watermarking technique (TSDL) in 2020 that bypasses the need for explicit image distortion modeling during training [15]. This method, which integrates adversarial training and channel coding, enhances system robustness and demonstrates comparable performance to explicit distortion models on trained distortions while generalizing better to new ones.

In the same year, Ahmadi et al. presented ReDMark [18], an end-to-end diffusion watermarking framework capable of learning watermarking algorithms in any transform space. With two Fully Convolutional Neural Networks featuring a residual structure, ReDMark performs real-time embedding and extraction, trained to conduct secure watermarking while simulating attacks as differentiable layers for end-to-end training. ReDMark's diffusion of watermark data across the image enhances security and robustness, outperforming contemporary methods in imperceptibility, robustness, and processing speed.

Yu [16] introduced ABDH, leveraging generative adversarial networks for data generation, proposing an end-to-end framework that extends their utility to data hiding. This framework includes a discriminative model that simulates the detection process, understanding the cover image's sensitivity to semantic changes, and an attention model that generates masks to produce high-quality target images without disturbing key features.

Zhong et al. presented a robust, blind image watermarking scheme using deep neural networks' fitting capabilities to automate the watermarking process [11]. The architecture is tailored for watermarking tasks and trained unsupervised to minimize domain knowledge requirements. It demonstrates flexibility and robustness against a variety of distortions without prior knowledge of potential image alterations, showcasing an unsupervised training approach for watermarking that achieves robustness without human intervention or annotation.

Addressing the shortcomings of existing works that often fail to preserve image quality or robustness against perturbations or are too complex to train, Bui et al. proposed RoSteALS [54] in 2023, a practical technique using frozen pre-trained autoencoders. This model offers a small model size, modular design, and state-of-the-art secret recovery capability with comparable image quality. With a light-weight secret encoder of just 300k parameters, it is easy to train and has perfect secret recovery performance.

In summary, these deep learning-based watermarking techniques have made significant strides in enhancing the robustness and adaptability of digital watermarking, providing efficient and secure methods for image authentication and copyright protection while maintaining high visual quality.

The proposals of these schemes have effectively bolstered the resilience of deep watermarking solutions in the face of geometric, signal processing, and deep learning attacks. Beyond the aforementioned prevalent attack vectors, the realm of Cross-Media Attacks has emerged as an area of considerable public interest. This is due to the omnipresence of camera phones and digital displays, capturing digitally displayed images with camera phones is becoming widely practiced.

In 2019, Wengrowski et al. [55] developed an advanced Light Field Messaging (LFM) system using deep learning for digital watermarking in photographic domains. It embeds hidden video information on screens for capture by handheld cameras, aiming for minimal visual distortion and maximal recovery accuracy. The system addresses the complex interaction of screen radiation through a Camera-Display Transfer Function (CDTF) trained on millions of images. This results in undetectable message embedding and highly reliable recovery, outperforming current methods with superior BER scores.

Inspired by the powerful feature learning capacity of deep neural networks, Fang et al. [56] proposed a deep template-based watermarking algorithm in 2020, which is achieved by leveraging the special properties of the human visual system, i.e., insensitivity to specific chrominance components, the proximity principle, and the oblique effect. At the extracting side, a novel two-stage extracting network is further proposed, which first tries to recover the distorted images with an enhancing subnetwork then classify the watermark patterns into bits. Thanks to this work's strong extracting ability, both digital editing resilience and camera-shooting resilience are considered for the first time.

At the same year, Tancik et al. [57] introduced StegaStamp, a state-of-the-art steganographic algorithm for embedding digital data into printed and digital photos. This system allows data to be encoded and decoded in a way that is nearly invisible to the human eye and retrievable by internet-connected imaging systems. At the heart of StegaStamp is a deep neural network, trained to be resilient against image distortions typical of real-world photography and printing. The inclusion of an image perturbation module ensures that the model is versatile and adaptable to various real-world imaging conditions.

Jia et al. [17] proposed a novel approach (named RIHOOP) to embed hyperlinks into common images, making the hyperlinks invisible for human eyes but detectable for mobile devices equipped with a camera. This method employs an end-to-end neural network with an encoder for message embedding and a decoder for extraction, fortified by a distortion network that simulates camera-induced image distortion in printing and display. The network's 3-D rendering operations enhance robustness against camera capture. Experiments demonstrate its superiority over prior methods in robustness and image quality, enabling applications like "image hyperlinks" for multimedia advertising and "invisible watermarks" for digital copyright protection.

In 2021, Fang et al. introduced "TERA," a screen-to-camera image code that emphasizes transparency, efficiency, robustness, and device adaptability [58]. By harnessing human vision properties and an advanced attention-guided network, TERA excels in robustness and visual quality, with potential applications in 2D image codes and watermarks for tracking and warning. However, it is not immune to cropping attacks and has limitations in terms of message capacity.

Continuing the innovation, in 2022, Jia et al. proposed a method for learning and detecting invisible markers in offline-to-online photography scenarios, focusing on the comprehensive process of hiding, locating, correcting, and recovering information [59]. This approach embeds data in sub-images within an integrated localization module in an end-to-end framework, ensuring high visual quality and robustness in recovery, and demonstrating resilience to common shooting conditions and geometric distortions.

Fang et al. introduced a novel insight in 2022, suggesting that it is unnecessary to simulate the entire screen-shooting process quantitatively; instead, focusing on the most influential distortions is sufficient to create a robust noise layer [60]. They presented PIMoG, a screen-shooting noise layer that simulates key distortions—perspective, illumination, and moiré—in a differentiable manner, demonstrating exceptional performance and robustness under various screen-shooting conditions.

While these proposals have made significant strides in enhancing the resilience and applicability of deep watermarking solutions, they are not without limitations. For instance, some methods may still be susceptible to cropping attacks and have limited message capacity. Nevertheless, these advancements open up new horizons for research and practical applications in the field of digital security and intellectual property protection.

It is important to note that in the watermarking methods we discussed, the encoder's input is related to both the watermark and the carrier image. This main framework is referred to as the Dependent Deep Watermarking (DDW) framework (see Fig. 6). Additionally, there is another notable setup where the encoder's input is related only to the watermark. This type of framework is called the Universal Deep Watermarking (UDW) framework (see Fig. 7). In 2020, Zhang et al. [20] pioneered the inaugural universal meta-architecture, designated as Udh. In their scholarly article, they employed their proprietary framework to dissect characteristics of DNN-based information embedding, encompassing the following aspects: a) The encoding of information exhibits cross-channel and spatial locality traits; b) Information is encoded as high-frequency entities, integrated into the host image, with the decoding process hinging on the variance in high-frequency content to extricate the watermark. Exploiting its universal property, Udh is applied for efficient watermarking. Leveraging the simplicity, efficacy, and versatility of UDW, Luo et al. [1] introduced an adaptive patch-level watermarking framework (named DIPW) for copyright protection in 2023. This work assessed the shortcomings of current watermarking methods for copyright protection and introduced a new framework that enhances imperceptibility and secrecy. The framework features a dual-loss function, a watermark-sensitive classifier, an object-attention heuristic for patch elaboration, and a plagiarism-resistant learning approach, increasing robustness in watermark extraction. Its effectiveness and performance were confirmed through experiments and visualization analyses. Table 2 summarizes the main difference between these schemes. Notably, we select some representative schemes [1,12,20] to demonstrate the effects of CNN-based deep watermarking (see Fig. 8).

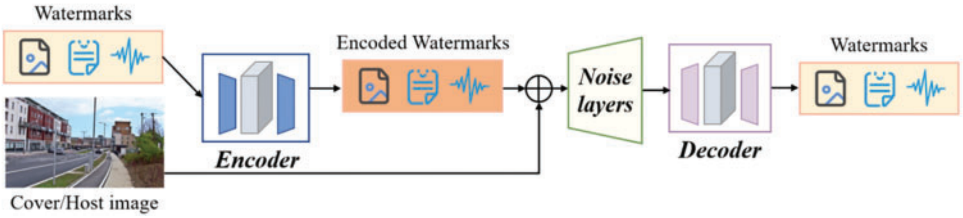

Figure 7: General framework of universal CNN-based deep image watermarking

Table 2: Comparison table between CNNs-based deep image watermarking schemes

Method	Years	Watermarks	Optimizations	Resistant attacks
HiDDeN [12]	2018	Message	Noise adversarial training	Signal processing
TSDL [14]	2019	Message	Two-stage separable training approach	Signal processing
Luo et al. [15]	2020	Message	Integrates adversarial training and channel coding	Geometric; signal processing
ReDMark [18]	2020	gray-scale Images	End-to-end diffusion watermarking framework	Signal processing

(Continued)

Table 2 (continued)

Method	Years	Watermarks	Optimizations	Resistant attacks
ABDH [16]	2020	Image	Attention mask	Geometric; signal processing
Zhong et al. [11]	2020	gray-scale Images	Unsupervised training approach	Geometric; signal processing
RoSteALS [54]	2023	Image	Leveraging frozen pre-trained autoencoders/latent space of a pre-trained auto encoder	Geometric; signal processing
LFM [55]	2019	Message/barcode	Camera-Display Transfer Function (CDTF)	Screen capture
Fang et al. [56]	2020	Message	Two-stage extracting network	Signal processing; Cross-Media
Stegastamp [57]	2020	Hyperlink	Retrievable by internet-connected imaging systems	Geometric; signal processing; Cross-Media
RIHOOP [17]	2020	Hyperlinks	Simulates camera-induced image distortion in printing and display	Signal processing; Cross-Media
TERA [58]	2021	Message	Advanced attention-guided strategy	Geometric; signal processing; Cross-Media
Jia et al. [59]	2022	Data matrix	Sub-images embedding trategy with integrated localization	Signal processing; Cross-Media
PIMoG [60]	2022	Message	Simulates key screen-shooting distortions in a differentiable manner.	Signal processing; Cross-Media
Udh [20]	2020	Image	Streamlines the encoder input to be only associated with the watermark	Signal processing
DIPW [1]	2023	Image	Patch-level embedding strategy	Geometric; signal processing; plagiarism

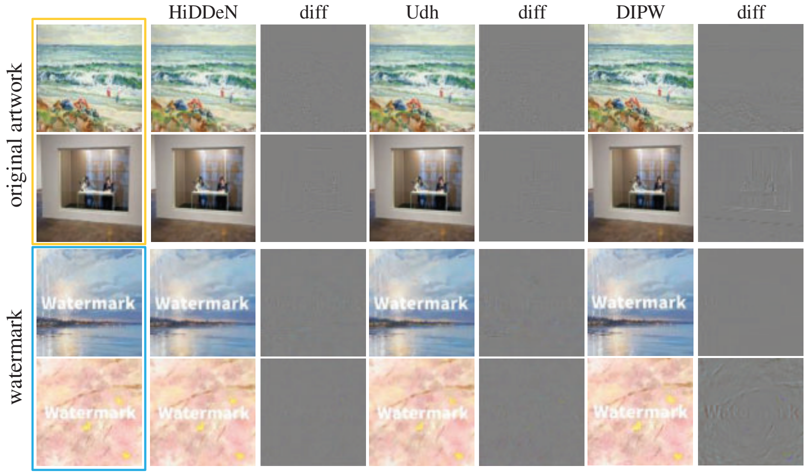

Figure 8: The effects of CNN-Based deep watermarking

5.4 INN-Based Deep Watermarking

1) Prior Knowledge

Invertible Neural Networks (INNs) are a distinct class of neural network architectures characterized by their reversibility. Each operation within the network is invertible, meaning that the network can generate an output from a given input and can also reconstruct the original input data from this output using the same network parameters [61]. In INNs, the forward and backward propagation processes share identical parameters, ensuring consistency in both the forward and reverse operations [62]. Due to their reversible nature, theoretically, no information is lost during the forward and backward propagation, making INNs particularly well-suited for applications requiring precise reversibility, such as lossless data compression, generative modeling, and information hiding—where the hidden data must be perfectly recoverable. As illustrated in Fig. 9, INNs comprise a series of reversible layers or blocks, such as coupling layers or other types of reversible structures. These designs allow the network to encode information during forward propagation and decode it during reverse propagation. In the context of information hiding, Jing et al. [26] introduced an invertible neural network called HiNet for image hiding that significantly enhances both security and accuracy in data recovery. HiNet employs the same network parameters for both the concealing and revealing of images, treating these tasks as the forward and backward operations of an invertible network. This unified parameter usage means the network requires only a single training session to optimize parameters for both the concealing and revealing processes. The introduction of HiNet lays a foundational basis for the development of subsequent INNs-based deep watermarking techniques, also known as normalizing flow-based approaches (Fig. 10).

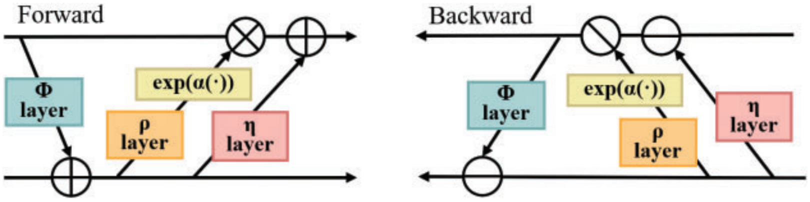

Figure 9: The basic structure of Invertible Neural Networks (INNs)

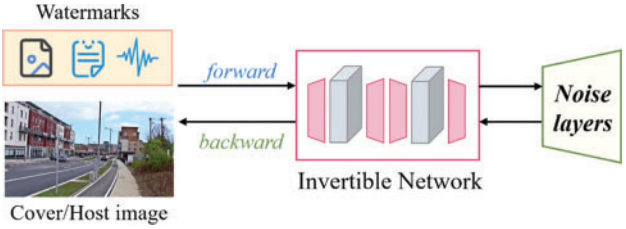

Figure 10: General framework of INN-based deep image watermarking

2) Main Schemes

In 2022, Xu et al. [24] unveiled a groundbreaking framework, RIIS. This novel approach elegantly integrates a conditional normalizing flow to capture the nuanced distribution of the high-frequency component, contingent upon the characteristics of the container image. RIIS is further bolstered by a meticulously crafted Container Enhancement Module (CEM) that significantly contributes to the robustness of the reconstruction process. To address the challenge of varied distortion levels, RIIS incorporates a Distortion-Guided Modulation (DGM) mechanism, which adeptly adjusts the network parameters across flow-based blocks. This innovation positions RIIS as a versatile solution, capable of maintaining imperceptibility and capacity while enhancing robustness against distortions. Extensive experimental evidence has confirmed RIIS's superior performance, solidifying its potential for practical applications.

Building upon this foundation, in 2023, Luo et al. [9] introduced the IRWArt watermarking framework, a sophisticated parameter-sharing Invertible Neural Network (INN) tailored for the protection of artwork copyrights. The IRWArt framework is particularly noteworthy for its strategic use of INN's forward and backward processes to embed and retrieve watermarks with precision. It skillfully integrates a Frequency Domain Transformation Module (FDTM), a Quality Enhancement Module (QEM), noise layers, and astute training strategies to direct watermark embedding into the least obtrusive regions of the artwork. This ensures minimal impact on the artwork's quality while fortifying against plagiarism. The effectiveness and exceptional performance of IRWArt have been corroborated through experimentation and visual analysis.

In their latest contribution, Lan et al. [63] have developed a robust end-to-end system that leverages the power of an Invertible Neural Network (INN). This system distinguishes itself by prioritizing the minimization of information loss during the forward operation, a strategic departure from traditional spatial domain applications of INN. It introduces a Mutual Information Loss to meticulously regulate information flow and employs a Two-Way Fusion Module (TWFM) that adeptly utilizes spatial and Discrete Cosine Transform (DCT) domain features to bolster message extraction. These pioneering features ensure the lossless recovery of secret messages from DCT coefficients. The experimental results are nothing short of impressive, demonstrating a marked reduction in error rates and setting a new benchmark.

Fang et al. [27] proposed a flow-based architecture that ensures the encoded features are in high alignment with those required by the decoder, thereby minimizing the embedding of redundant features. This is achieved through the shared parameters of the architecture, which enhance the consistency between the encoded and required decoder features. Additionally, they developed an INN-based noise layer (INL) specifically designed to simulate and manage black-box distortions effectively. The INL's forward operation serves as a noise layer during training, while its backward operation enables denoising prior to extraction. Comprehensive testing has demonstrated that this method surpasses other state-of-the-art techniques in terms of robustness against both white-box and black-box distortions, as well as invisibility. This innovative approach marks a significant advancement in the field of digital watermarking. Table 3 summarizes the main difference between these schemes. Notably, we select some representative schemes [24,9] to demonstrate the effects of CNN-based deep watermarking (see Fig. 11).

Table 3: Comparison table between CNNs-based deep image watermarking schemes

Method	Years	Watermarks	Optimizations	Resistant attacks
RIIS [24]	2022	Image	Container enhancement module and distortion-guided modulation	Signal processing
IRWArt [9]	2023	Image	Contrastive loss functions	Geometric; signal processing; plagiarism g
Lan et al. [63]	2023	Message	Mutual information loss and two-way fusion module	Signal processing
Fang et al. [27]	2023	Data matrix	Simulate and manage black-box distortions	Geometric; signal processing; black-box

6 Challenges in Deep Watermarking Technology

1) Challenge of Model Generalization.

In the field of deep watermarking, the ability of models to generalize is crucial for practical applications. Current methods often depend on distortion models used during training, which limits their adaptability to unknown distortions. To overcome this challenge, future research should explore new training strategies such as adaptive learning rate adjustments, meta-learning, and the introduction of more complex network architectures like attention mechanisms and graph neural networks.

2) Balancing Image Quality and Robustness.

Deep watermarking technology must maintain image quality while ensuring the robustness of watermark information. This challenge involves optimizing encoding strategies to achieve effective information hiding with minimal visual impact. Future research could investigate content-based encoding methods and multi-scale, multi-band encoding techniques to enhance the stealthiness and robustness of watermarks.

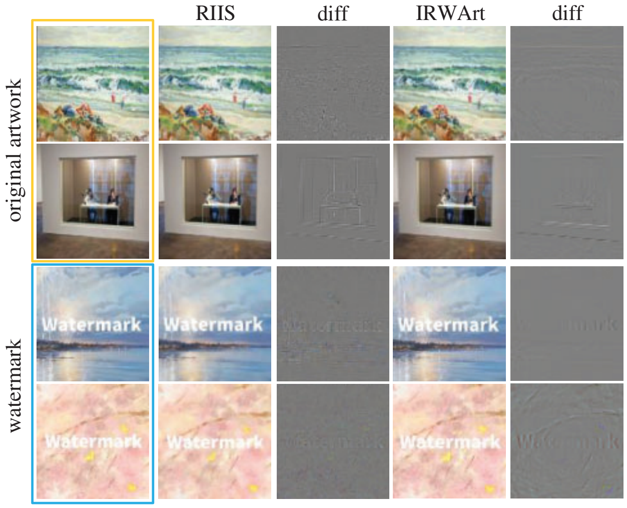

Figure 11: The effects of INN-based deep watermarking

3) *Defense Against Adversarial Attacks*.

Another challenge for deep watermarking systems is defending against adversarial attacks, which attempt to compromise the detectability of watermarks through subtle perturbations. To enhance system security, future studies could develop new adversarial training approaches and leverage generative adversarial networks (GANs) to bolster the model's defensive capabilities.

4) Computational Complexity and Real-Time Processing.

In applications requiring swift responses, such as video streaming or real-time monitoring, the computational efficiency and real-time processing capabilities of deep watermarking algorithms are essential. Future research could focus on algorithm optimization to reduce computational overhead and explore parallel computing techniques on specific hardware platforms.

5) Diversity and Scale of Datasets.

Training robust deep watermarking models necessitates large-scale and diverse datasets. This involves not only data collection and processing but also addressing privacy protection and copyright issues. Future research could investigate data augmentation techniques and synthetic data generation methods to construct richer training datasets.

6) Model Interpretability and Transparency.

In legal and regulatory contexts, the interpretability and transparency of deep watermarking models are vital for ensuring legality and trustworthiness. Future studies could develop new visualization tools and techniques to enhance model transparency and meet legal and regulatory requirements.

7 Future Research Directions in Deep Watermarking Technology

1) Enhancing Model Generalization

Future research could explore novel training methodologies aimed at improving model adaptability to unseen data. Techniques like adaptive learning rates adjust the learning rate during training based on the model's performance, potentially reducing overfitting and enhancing generalization. Incorporating advanced network architectures such as attention mechanisms, which focus on relevant features of input data, and graph neural networks, which excel in data structured as graphs, may also boost the generalization capabilities of watermarking models.

2) Optimizing the Balance between Image Quality and Robustness

Enhancing the trade-off between maintaining high image quality and ensuring robust watermarking could involve the development of innovative encoding and decoding strategies. Content-based encoding adjusts the embedding strength based on the content of the image, potentially preserving important details while embedding data. Additionally, multi-scale multi-band encoding could leverage different frequency bands and scales to optimize the visibility and robustness of the watermark, allowing for finer control over the embedding process.

3) Exploring Adversarial Training and Defense Mechanisms

To enhance robustness against sophisticated attacks, future research should focus on adversarial training techniques that involve training models against examples specifically designed to deceive them. Using Generative Adversarial Networks (GANs), where a generator and discriminator work in tandem, can simulate adversarial attacks during training, thus preparing the watermarking system to withstand real-world tampering and manipulation.

4) Development of Real-Time Watermarking Algorithms

Efforts to develop real-time watermarking algorithms should focus on optimizing algorithms to reduce computational demands and latency. Investigating parallel computing frameworks and specialized hardware accelerations, such as GPUs or TPUs, can significantly decrease processing times, enabling real-time watermark embedding and extraction in streaming media applications.

5) Construction of Multimodal Datasets

To improve model performance across diverse scenarios, creating large-scale, multimodal datasets that encompass various image types, sources, and conditions is essential. Research into advanced data augmentation techniques, such as synthetic image generation and simulation of various attack scenarios, can further enhance model robustness and generalization.

6) Improving Model Interpretability

Developing tools and methodologies for better model interpretability is crucial for ensuring transparency and trust in watermarking applications. Visualization techniques that elucidate how models make decisions can help in fine-tuning the models and in validating their effectiveness, which is especially important in meeting regulatory and compliance standards.

7) Cross-Modal Watermarking Techniques

Expanding the application of watermarking technology to other media types such as video and audio, and exploring cross-modal synchronization and fusion techniques, can open new avenues for copyright protection across multimedia content. Such research could lead to more robust and versatile watermarking systems that ensure content protection in diverse media formats.

8) Security and Privacy Protection

Research into integrating advanced encryption methods, secure communication protocols, and privacy-preserving techniques into the watermarking process will be critical. Ensuring the confidentiality and integrity of the embedded data during transmission and storage is paramount for the widespread adoption of watermarking technologies.

9) Integration into Practical Applications

Integrating watermarking technology into practical applications requires research into compatibility with existing digital infrastructure, such as digital rights management systems and social media platforms. Studies on seamless integration strategies that do not compromise system performance or user experience are vital for the adoption of watermarking technologies in commercial products.

10) Development of International Standards and Regulations

Participating in the development of international standards and regulations is crucial for ensuring the compatibility and uniformity of watermarking technologies across borders. Establishing clear guidelines and standards can help in fostering global acceptance and enhancing the trustworthiness of watermarking solutions in international markets.

8 Conclusion

This paper offers an extensive review of deep image watermarking, an advanced technique for protecting the copyright and integrity of digital media, particularly images, in the rapidly evolving digital landscape. The review begins by outlining the fundamental concepts of digital watermarking and highlighting the limitations of traditional methods. The paper then delves into the architecture and processes of deep watermarking, showcasing how recent advancements in deep learning have enabled the development of robust watermark encoders and decoders. These deep learning-based watermark algorithms, referred to as "deep watermarking," have demonstrated significant potential due to their simplified framework and enhanced performance capabilities. A comparative analysis of various deep watermarking paradigms reveals the current technological challenges and identifies key areas for future research. The paper emphasizes the importance of model generalization, balancing image quality with robustness, and defending against adversarial attacks. It also underscores the need for real-time processing algorithms, diverse and large-scale datasets, model interpretability, and the integration of security and privacy protections.

Looking ahead, the future of watermarking lies in enhancing model generalization, balancing image quality with robustness, and developing real-time algorithms. It also involves constructing multimodal datasets for broader applicability, improving model interpretability, and expanding into cross-modal techniques. Additionally, integrating security and privacy protections, seamless practical application, and the development of international standards will be pivotal. These collective efforts will ensure the continued innovation and practical relevance of deep image watermarking in the realm of digital rights management and content protection.

Acknowledgement: The authors would like to express our sincere gratitude and appreciation to each other for our combined efforts and contributions throughout the course of this research paper.

Funding Statement: This work is supported by the National Natural Science Foundation of China (Nos. 62072465, 62102425), the Science and Technology Innovation Program of Hunan Province (Nos. 2022RC3061, 2023RC3027).

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Yuanjing Luo, Zhiping Cai; data collection: Xichen Tan; analysis and interpretation of results: Yuanjing Luo, Xichen Tan; draft manuscript preparation: Yuanjing Luo. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data and materials used in this review are derived from publicly accessible databases and previously published studies, which are cited throughout the text. References to these sources are provided in the bibliography.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] Y. Luo, T. Zhou, S. Cui, Y. Ye, F. Liu and Z. Cai, "Fixing the double agent vulnerability of deep watermarking: A patch-level solution against artwork plagiarism," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 3, pp. 1670–1683, 2023. doi: 10.1109/TCSVT.2023.3295895.

[2] A. Ray and S. Roy, "Recent trends in image watermarking techniques for copyright protection: A survey," *Int. J. Multimed. Inf. Retr.*, vol. 9, no. 4, pp. 249–270, 2020. doi: 10.1007/s13735-020-00197-9.

[3] X. H. Jiang, "Digital watermarking and its application in image copyright protection," presented at 2010 Int. Conf. on Intell. Comput. Techn. Autom., Changsha, China, Autom, May 11–12, 2010.

[4] P. Kadian, S. M. Arora, and N. Arora, "Robust digital watermarking techniques for copyright protection of digital data: A survey," *Wirel. Pers. Commun.*, vol. 118, pp. 3225–3249, 2021. doi: 10.1007/s11277-021-08177-w.

[5] A. Piva, M. Barni, F. Bartolini, and V. Cappellini, "DCT-based watermark recovering without resorting to the uncorrupted original image," presented at 1997 Int. Conf. on Img. Pro., Santa Barbara, CA, USA, Oct. 26–29, 1997, pp. 26–29.

[6] I. Cox, M. Miller, J. Bloom, and C. Honsinger, "Digital watermarking," *J. Electron. Imaging.*, vol. 11, no. 3, pp. 414–414, 2002. doi: 10.1117/1.1494075.

[7] J. S. Mei, S. K. Li, and X. M. Tan, "A digital watermarking algorithm based on DCT and DWT," presented at 2009 6th Web Inf. Syst. App. Conf., Xuzhou, China, Sep. 18–20, 2009.

[8] 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, 2019. doi: 10.1016/j.neucom.2019.01.067.

[9] Y. Luo, T. Zhou, F. Liu, and Z. Cai, "IRWArt: Levering watermarking performance for protecting high-quality artwork images," presented at ACM Web Conf. 2023, Austin, TX, USA, Apr. 30–May 4, 2023.

[10] Y. Meng, X. Chen, X. Sun, Y. Liu, and G. Wei, "A dual model watermarking framework for copyright protection in image processing networks," *Comput. Mater. Contin.*, vol. 75, pp. 831–844, 2023. doi: 10.32604/cmc.2023.033700.

[11] X. Zhong, P. -C. Huang, S. Mastorakis, and F. Y. Shih, "An automated and robust image watermarking scheme based on deep neural networks," *IEEE Trans. Multimed.*, vol. 23, pp. 1951–1961, 2020. doi: 10.1109/TMM.2020.3006415.

[12] J. Zhu, R. Kaplan, J. Johnson, and L. Fei-Fei, "HiDDeN: Hiding data with deep networks," presented at 15th Eur. Conf. Comput. Vis., Munich, Germany, Sep. 8–14, 2018, pp. 8–14.

[13] R. Zhang, S. Dong, and J. Liu, "Invisible steganography via generative adversarial networks," *Multimed. Tools Appl.*, vol. 78, pp. 8559–8575, 2019. doi: 10.1007/s11042-018-6951-z.

[14] Y. Liu, M. Guo, J. Zhang, Y. Zhu, and X. Xie, "A novel two-stage separable deep learning framework for practical blind watermarking," presented at 27th ACM Int. Conf. Multimed., Nice, France, Oct. 21–25, 2019.

[15] X. Luo, R. Zhan, H. Chang, F. Yang, and P. Milanfar, "Distortion agnostic deep watermarking," presented at 2020 IEEE Conf. Comput. Vis. Pattern Recognit., Seattle, WA, USA, Jun. 14–19, 2020.

[16] C. Yu, "Attention based data hiding with generative adversarial networks," presented at 34th AAAI Conf. Artif. Intell., New York, NY, USA, Feb. 7–12, 2020.

[17] J. Jia *et al.*, "RIHOOP: Robust invisible hyperlinks in offline and online photographs," *IEEE Trans. Cybern.*, vol. 52, no. 7, pp. 7094–7106, 2020. doi: 10.1109/TCYB.2020.3037208.

[18] M. Ahmadi, A. Norouzi, N. Karimi, S. Samavi, and A. Emami, "ReDMark: Framework for residual diffusion watermarking based on deep networks," *Expert Syst. Appl.*, vol. 146, 2020, Art. no. 113157. doi: 10.1016/j.eswa.2019.113157.

[19] H. Zhang and Y. Li, "Digital watermarking via inverse gradient attention," presented at the 2022 9th IEEE Int. Conf. Behav. Social Comput., Matsuyama, Japan, Oct. 29–31, 2022.

[20] C. Zhang, P. Benz, A. Karjauv, G. Sun, and I. S. Kweon, "UDH: Universal deep hiding for steganography, watermarking, and light field messaging," presented at 34th Adv. Neural Inf. Process. Syst., Dec. 6–12, 2020.

[21] Z. Zheng, Y. Hu, Y. Bin, X. Xu, Y. Yang, and H. T. Shen, "Composition aware image steganography through adversarial self-generated supervision," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 11, pp. 9451–9465, 2023. doi: 10.1109/TNNLS.2022.3175627.

[22] A. C. Gilbert, Y. Zhang, K. Lee, Y. Zhang, and H. Lee, "Towards understanding the invertibility of convolutional neural networks," presented at 26th Int. Joint Conf. Artif. Intell., Melbourne, Australia, Aug. 19–25, 2017.

[23] T. F. van der Ouderaa and D. E. Worrall, "Reversible GANs for memory efficient image-to-image translation," presented at 2019 IEEE Conf. Comput. Vis. Pattern Recognit., Long Beach, CA, USA, Jun. 15–20, 2019.

[24] Y. Xu, C. Mou, Y. Hu, J. Xie, and J. Zhang, "Robust invertible image steganography," presented at 2022 IEEE Conf. Comput. Vis. Pattern Recognit., New Orleans, LA, USA, Jun. 18–24, 2022.

[25] S. -P. Lu, R. Wang, T. Zhong, and P. L. Rosin, "Large-capacity image steganography based on invertible neural networks," presented at 2021 IEEE Conf. Comput. Vis. Pattern Recognit., Nashville, TN, USA, Jun. 20–25, 2021.

[26] J. Jing, X. Deng, M. Xu, J. Wang, and Z. Guan, "HiNet: Deep image hiding by invertible network," presented at 2021 IEEE Int. Conf. Comput. Vis., Montreal, QC, Canada, Oct. 10–17, 2021.

[27] H. Fang, Y. Qiu, K. Chen, J. Zhang, W. Zhang, and E. -C. Chang, "Flow-based robust watermarking with invertible noise layer for blackbox distortions," presented at 37th AAAI Conf. Artif. Intell., Washington, DC, USA, Feb. 7–14, 2023.

[28] F. Li, Y. Sheng, X. Zhang, and C. Qin, "iSCMIS: Spatial-channel attention based deep invertible network for multi-image steganography," *IEEE Trans. Multimed.*, vol. 26, pp. 3137–3152, 2023. doi: 10.1109/TMM.2023.3307970.

[29] R. G. Van Schyndel, A. Z. Tirkel, and C. F. Osborne, "A digital watermark," presented at 1st Int. Conf. Img. Pro., Austin, TX, USA, Nov. 13–16, 1994.

[30] N. Nikolaidis and I. Pitas, "Digital image watermarking: An overview," presented at 6th IEEE Int. Conf. Multimed. Comput. Syst., Florence, Italy, Jun. 7–11, 1999.

[31] B. Wang, L. Shen, J. Zhang, Z. Xu, and N. Wang, "A text image watermarking algorithm based on image enhancement," *Comput. Mater. Contin.*, vol. 77, no. 1, pp. 1183–1207, 2023. doi: 10.32604/cmc.2023.040307.

[32] A. Tanchenko, "Visual-psnr measure of image quality," *J. Vis. Commun. Image R.*, vol. 25, no. 5, pp. 874–878, 2014. doi: 10.1016/j.jvcir.2014.01.008.

[33] D. R. I. M. Setiadi, "PSNR vs SSIM: Imperceptibility quality assessment for image steganography," *Multimed. Tools Appl.*, vol. 80, no. 6, pp. 8423–8444, 2021. doi: 10.1007/s11042-020-10035-z.

[34] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," presented at 2018 IEEE Conf. Comput. Vis. Pattern Recognit., Salt Lake City, UT, USA, Jun. 18–22, 2018.

[35] J. Liu and X. He, "A review study on digital watermarking," presented at 2005 Int. Conf. on Intell. Comput. Techn. Autom., Karachi, Pakistan, Aug. 27–28, 2005.

[36] J. Sang and M. S. Alam, "Fragility and robustness of binary-phaseonly-filter-based fragile/semifragile digital image watermarking," *IEEE Trans. Instrum. Meas.*, vol. 57, no. 3, pp. 595–606, 2008. doi: 10.1109/TIM.2007.911585.

[37] S. Tyagi, H. V. Singh, R. Agarwal, and S. K. Gangwar, "Digital watermarking techniques for security applications," presented at 2016 Int. Conf. Emerg. Trends Elect. Elec. Sustainable Eng. Syst., Sultanpur, India, Mar. 11–12, 2016.

[38] P. H. Wong, O. C. Au, and Y. M. Yeung, "Novel blind multiple watermarking technique for images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 8, pp. 813–830, 2003. doi: 10.1109/TCSVT.2003.815948.

[39] P. Singh and R. S. Chadha, "A survey of digital watermarking techniques, applications and attacks," *Int. J. Eng. Techno.*, vol. 2, no. 9, pp. 165–175, 2013.

[40] K. Huang, X. Tian, H. Yu, M. Yu, and A. Yin, "A high capacity watermarking technique for the printed document," *Electronics*, vol. 8, no. 12, 2019, Art. no. 1403. doi: 10.3390/electronics8121403.

[41] J. Fridrich, "Security of fragile authentication watermarks with localization," in *Security and Watermarking of Multimedia Contents IV*, USA: SPIE, 2002, vol. 4675, pp. 691–770.

[42] H. Tao, L. Chongmin, J. M. Zain, and A. N. Abdalla, "Robust image watermarking theories and techniques: A review," *J. Appl. Res. Tech.*, vol. 12, no. 1, pp. 122–138, 2014. doi: 10.1016/S1665-6423(14)71612-8.

[43] M. S. Kankanhalli, Rajmohan, and K. Ramakrishnan, "Adaptive visible watermarking of images," presented at 6th IEEE Int. Conf. Multimedia Comput. Syst., Florence, Italy, Jun. 7–11, 1999.

[44] M. M. Yeung and F. Mintzer, "An invisible watermarking technique for image verification," presented at 1997 Int. Conf. Img. Pro., Santa Barbara, CA, USA, Oct. 26–29, 1997.

[45] F. M. Boland, J. J. O'Ruanaidh, and C. Dautzenberg, "Watermarking digital images for copyright protection," presented at 15th Int. Conf. on Img. Pro. App., Edinburgh, Scotland, Jul. 4–6, 1995.

[46] S. Gupta, A. Goyal, and B. Bhushan, "Information hiding using least significant bit steganography and cryptography," *Int. J. Mod. Educ. Comput. Sci.*, vol. 4, no. 6, pp. 27–34, 2012. doi: 10.5815/ijmecs.2012.06.04.

[47] A. Anand and A. K. Singh, "Joint watermarking-encryption-ecc for patient record security in wavelet domain," *IEEE Trans. Multimed.*, vol. 27, no. 3, pp. 66–75, 2020. doi: 10.1109/MMUL.2020.2985973.

[48] D. R. Huang, J. F. Liu, J. W. Huang, and H. M. Liu, "A DWT-based image watermarking algorithm," presented at IEEE Int. Conf. Multimed. Expo, Tokyo, Japan, Aug. 22–25, 2001.

[49] M. Urvoy, D. Goudia, and F. Autrusseau, "Perceptual dft watermarking with improved detection and robustness to geometrical distortions," *IEEE Trans. Inf. Forensics Secur.*, vol. 9, no. 7, pp. 1108–1119, 2014. doi: 10.1109/TIFS.2014.2322497.

[50] C. -C. Chang, P. Tsai, and C. -C. Lin, "SVD-based digital image watermarking scheme," *Pattern Recogn. Lett.*, vol. 26, no. 10, pp. 1577–1586, 2005. doi: 10.1016/j.patrec.2005.01.004.

[51] W. Li, H. Wang, Y. Chen, S. M. Abdullahi, and J. Luo, "Constructing immunized stego-image for secure steganography via artificial immune system," *IEEE Trans. Multimed.*, vol. 25, no. 2, pp. 8320–8333, 2023. doi: 10.1109/TMM.2023.3234812.

[52] S. Baluja, "Hiding images in plain sight: Deep steganography," presented at 31st Adv. Neural Inf. Process. Syst., Long Beach, CA, USA, Dec. 4–9, 2017.

[53] S. Baluja, "Hiding images within images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 7, pp. 1685–1697, 2019. doi: 10.1109/TPAMI.2019.2901877.

[54] T. Bui, S. Agarwal, N. Yu, and J. Collomosse, "RoSteALS: Robust steganography using autoencoder latent space," presented at 2023 IEEE Conf. Comput. Vis. Pattern Recognit., Vancouver, BC, Canada, Jun. 17–24, 2023.

[55] E. Wengrowski and K. Dana, "Light field messaging with deep photographic steganography," presented at 2019 IEEE Conf. Comput. Vis. Pattern Recognit., Long Beach, CA, USA, Jun. 15–20, 2019.

[56] H. Fang *et al.*, "Deep template-based watermarking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 4, pp. 1436–1451, 2020. doi: 10.1109/TCSVT.2020.3009349.

[57] M. Tancik, B. Mildenhall, and R. Ng, "StegaStamp: Invisible hyperlinks in physical photographs," presented at 2020 IEEE Conf. Comput. Vis. Pattern Recognit., Seattle, WA, USA, Jun. 14–19, 2020.

[58] H. Fang *et al.*, "Tera: Screen-to-camera image code with transparency, efficiency, robustness and adaptability," *IEEE Trans. Multimed.*, vol. 24, pp. 955–967, 2021. doi: 10.1109/TMM.2021.3061801.

[59] J. Jia, Z. Gao, D. Zhu, X. Min, G. Zhai, and X. Yang, "Learning invisible markers for hidden codes in offline-to-online photography," presented at 2022 IEEE Conf. Comput. Vis. Pattern Recognit., New Orleans, LA, USA, Jun. 18–24, 2022.

[60] H. Fang, Z. Jia, Z. Ma, E. -C. Chang, and W. Zhang, "PIMoG: An effective screen-shooting noise-layer simulation for deep-learning-based watermarking network," presented at 30th ACM Int. Conf. on Multimed., Lisboa, Portugal, Oct. 10–14, 2022.

[61] T. Teshima, I. Ishikawa, K. Tojo, K. Oono, M. Ikeda and M. Sugiyama, "Coupling-based invertible neural networks are universal diffeomorphism approximators," presented at 34th Adv. Neural Inf. Process. Syst., Dec. 6–12, 2020.

[62] G. A. Padmanabha and N. Zabaras, "Solving inverse problems using conditional invertible neural networks," *J. Comput. Phys.*, vol. 433, no. 6, 2021, Art. no. 110194. doi: 10.1016/j.jcp.2021.110194.

[63] Y. Lan, F. Shang, J. Yang, X. Kang, and E. Li, "Robust image steganography: Hiding messages in frequency coefficients," presented at 37th AAAI Conf. Artif. Intell., Washington, DC, USA, Feb. 7–14, 2023.