Mesopotamian journal of Cybersecurity
Vol.5,No.2, **pp**. 453–476
DOI: https://doi.org/10.58496/MJCS/2025/028; ISSN: 2958-6542
https://mesopotamian.press/journals/index.php/cybersecurity

Review Article

Digital watermarking techniques, challenges, and applications: A review

Saif Aldeen S. Naem [1,*], Sarab M. Hameed [1]

[1]*Computer Science Department, College of Science, University of Baghdad, Baghdad, Iraq.*

ARTICLE INFO

Article History

Received 05 Jan 2025
Revised 15 Mar 2025
Accepted 31 Mar 2025
Published 07 Jun 2025

Keywords

Copyright protection

Deep learning-based watermarking

Multimedia security

Traditional watermarking

ABSTRACT

With the rapid advancement of technology, the transmission of digital media over the internet has become easier and more efficient, leading to its widespread use across various fields. However, this progress has also been accompanied by increased risks of breaches, theft, and unethical digital media manipulation. Therefore, watermarking is considered one of the most essential techniques for protecting, verifying, and authenticating digital media by embedding imperceptible information within it. This paper presents a comprehensive literature review that differs from previous studies in its thorough analysis of both traditional and deep learning-based watermarking developed over the last nine years, as well as its adoption of hybrid approaches for adaptive watermarking, accompanied by various image and video datasets. This versatility makes it valuable for numerous applications, including the military, healthcare, and entertainment fields. The results highlight the necessity of adopting adaptive techniques to address the growing digital challenges. Future directions can concentrate on integrating deep learning with dynamic watermarking models to harness the effectiveness and efficiency of watermarking.

1. INTRODUCTION

The enormous volume of multimedia content, such as images, videos, and audio, transferred from one place to another via social media or across other various platforms creates the necessity for the protection of intellectual property, copyright, and content integrity, as well as the prevention of unauthorized distribution [1]. As a result, numerous strategies have emerged, such as encryption, steganography, and digital watermarking, which have become indispensable approaches for protecting digital assets, including health care, entertainment, social media, and the military, because they have the ability to protect, verify, and authenticate digital media. Among these, digital watermarking has gained fame because of its dual functionality of robust media authentication and tamper detection, while it also embeds imperceptible identifiers [2]. Watermarking performance depends on the key metrics of imperceptibility, robustness, and capacity, which are affected by the algorithm, noise, watermarking size, and mode of operation. These aspects need to be carefully balanced; for example, an improvement in robustness can compromise imperceptibility and vice versa [3]. Furthermore, digital media has been progressively manipulated by several advanced techniques; therefore, watermarking techniques must evolve to resist new types of attacks while preserving usefulness.

The components of the watermarking are a host media and a watermark to be embedded into the host media. The watermarked media is transferred across the channel. When the recipient receives it, the watermark is extracted [1]. Figure 1 depicts this process by presenting the general framework of a digital watermarking system, including the embedding and extraction stages. Formally, the embedding process, F, and the extraction process, E, can be represented by Equations 1 and 2, respectively [2].

$$M' = w + \alpha\, F(M, w) \tag{1}$$

where

M represents the host media in which the watermark is embedded,

w is a watermarking that can involve text, images, or any other type of data to be embedded in M,

F is the function or algorithm used to insert the watermark, w , into the media. M.

And α is the weight factor or threshold that determines the extent of the watermark's effect on the media. This factor helps determine the strength of the effect applied to the original data during the insertion process. If α is large, the effect on the original media will be strong, whereas if α is small, the impact will be less noticeable.

*Corresponding author.Email: Sarab.m@sc.uobaghdad, edu.iq

$$w = E(M', k) \quad (2)$$

k is a parameter or value that can be used in the extraction process to determine how to handle the modified media M' during retrieval.

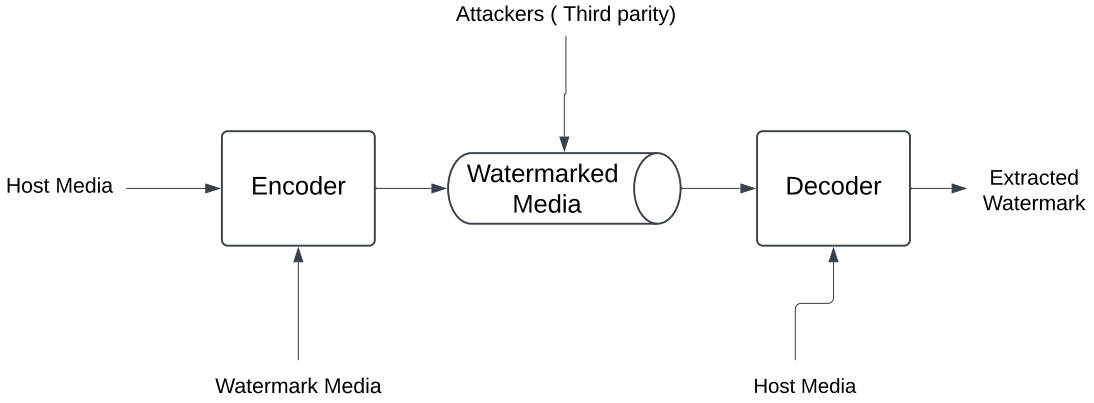

Fig. 1. *General Framework of a Digital Watermarking System*

Early techniques of digital watermarking operated in spatial domains via immediately changing pixel values [2] or frequency domains to embed watermarks in transform coefficients [3]. While these techniques provide imperceptibility, they struggle with robustness. For example, spatial-domain techniques such as least significant bit (LSB) substitution [4] are computationally easy; however, they are distinctly inclined to noise, whereas frequency-domain approaches, although more robust, confront trade-offs between payload ability and distortion [3].

Deep learning has revolutionized digital watermarking by leveraging architectures such as convolutional neural networks (CNNs) and generative adversarial networks (GANs) to optimize robustness and imperceptibility. Unlike traditional methods, these models learn adaptive embedding strategies, embedding watermarks in frequency domains or resilient regions to withstand geometric distortions and compression attacks. However, challenges such as computational complexity, adversarial attacks, and dependency on large-scale datasets persist [5].

Despite significant progress in traditional watermarking techniques as well as deep learning-based techniques, many gaps and challenges remain that have not been addressed. These gaps include several issues. First, we need to realize a balance among imperceptibility, robustness, and capacity since larger watermarks increase robustness and decrease imperceptibility. Second, as deep learning becomes more widespread, watermarking systems must evolve to take advantage of these techniques while being aware of issues such as overfitting, adversarial attacks, and increasing operational complexity. Finally, the volume of digital media has increased, which requires scalable and effective solutions across all platforms.

This review provides a comprehensive analysis of watermarking techniques up to 2025, covering traditional and deep learning-based methods. Additionally, a specialized and actionable set of recommendations for future research is presented, ensuring a more targeted and effective method. Consequently, this paper aims to provide a thorough analysis of digital watermarking techniques, and it differs from previous works in that it focuses on traditional and deep-learning watermarking techniques. Furthermore, it presents the limitations of the practical implementation of deep learning and suggests how to address these limitations. Additionally, it provides a structured taxonomy, different dataset analyses, and assessments of up-to-date watermarking techniques, which makes it a valuable resource for researchers. The contributions of this study are as follows:

1. To provide a detailed analysis of watermarking techniques, including a review of traditional and deep learning-based watermarking models, to provide a basic understanding.
2. To highlight the challenges of watermarking systems, including robustness, imperceptibility, and computational efficiency.
3. To understand the current situation of digital watermarking and identify key areas for future research.

The remainder of this paper is organized as follows: Section 2 presents a taxonomy of watermarking techniques. Sections 3 and 4 discuss common attacks on watermarked content and challenges in watermarking. Section 5 explores watermarking applications in digital media. Section 6 outlines benchmark datasets commonly used in watermarking research. Section 7 reviews traditional and deep learning-based watermarking techniques. Finally, Section 8 summarizes the key recommendations presented in this paper.

2. TAXONOMY OF WATERMARKING

Figure 2 depicts the classification of watermarking techniques on the basis of the host media, process, embedding method, perceptibility, robustness, domain, and embedding location. It provides a structured overview of different watermarking methods in digital media applications. The taxonomy aims to encompass all essential aspects of watermarking techniques, providing a foundation for designing appropriate models. The following subsections present the watermarking taxonomy to understand how they are used, how well they work, and their limitations in different states.

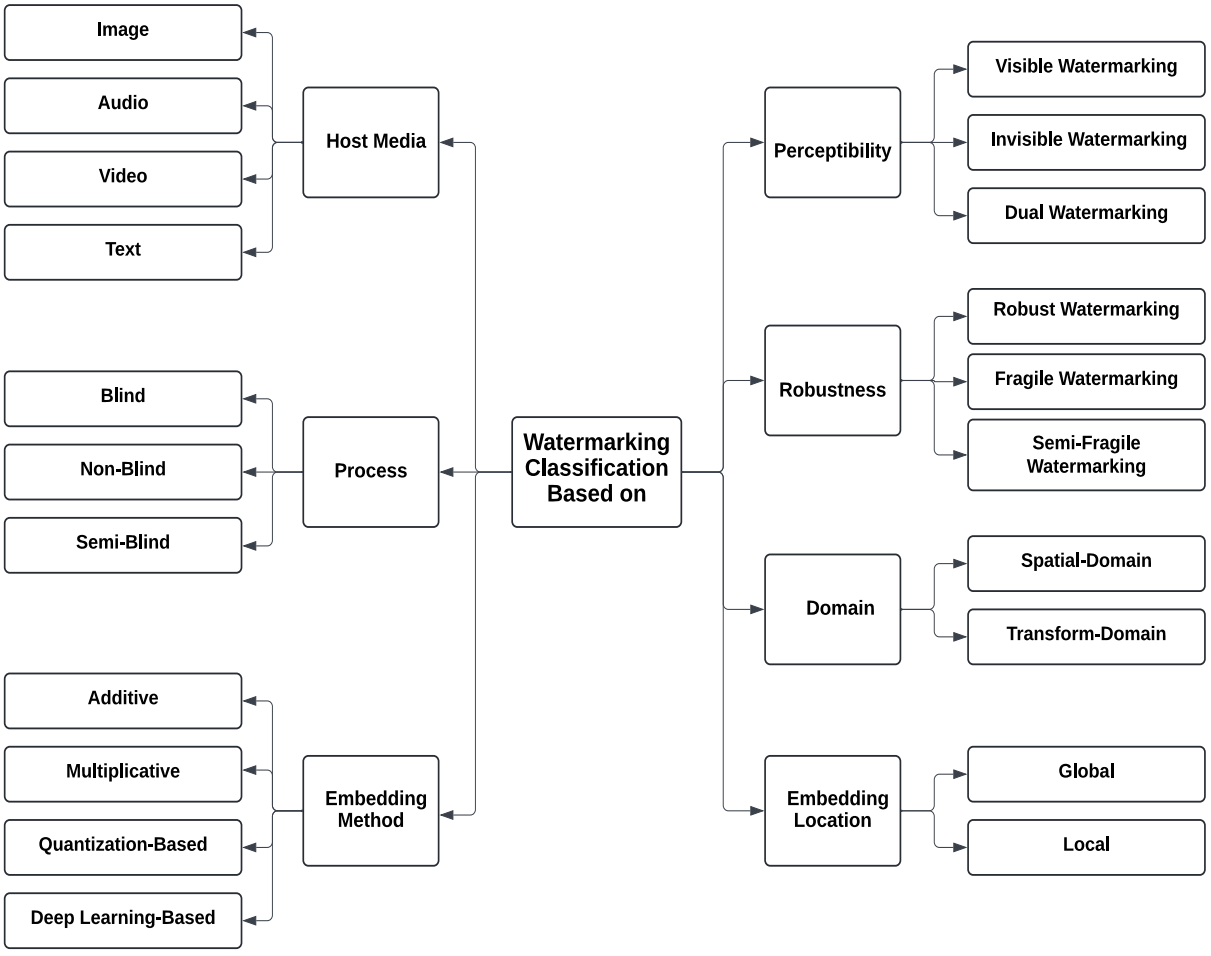

Fig. 2. Classification of Digital Watermarking Techniques Based on Multiple Factors.

2.1 Perceptibility

Watermarking can be categorized according to perceptibility into three types: visible watermarking is designed to be seen by the human eye, such as a logo or text that stands out in the image or video. In invisible watermarking, the watermark is hidden and cannot be seen by the human eye. It is used for data authentication and prevents illegal copying of content. Dual watermarking combines visible and invisible watermarks to provide an extra layer of protection in which a visible watermark is added to the digital media. Then, an invisible watermark is embedded into the watermarked image or video [4].

2.2 Robustness

Robustness is crucial for protecting the copyright, as attackers may try to alter the watermarked image. Robustness means that the watermark can be detected even when digital media changes. The robust procedure often involves embedding the watermark multiple times to make it harder to remove or distort. They are used in copyright where the watermark must withstand manipulations to confirm the owner's authenticity [5].

Fragile watermarks are designed to detect even the smallest changes in an image. If any modifications are made, the watermark will be destroyed, signalling tampering. This type of watermark is mostly used for verifying the integrity of content, as even minor alterations can be detected, making it useful for content authentication [5].

Semifragile watermarks are a balance between robust and fragile types; they can withstand minor changes such as compression but fail if larger modifications occur. These watermarks are often used to validate the authenticity of an image, detecting unauthorized modifications while allowing some acceptable alterations [5].

2.3 Domain

With respect to the domain where the data are embedded, there are two domains: spatial and transform.
Spatial-domain watermarking embeds watermark information directly into the pixels of an image. It alters the image's spatial or time domain. Although spatial domain techniques are easy to implement, provide low complexity, and allow for high embedding capacity, they are often less robust against attacks such as compression or modification. The least significant bit (LSB) approach is the most commonly employed spatial domain. LSB involves incorporating the watermark into LSBs of digital media (images or videos); this method alters the least amount of crucial information with the watermark bits, making them virtually undetectable. The watermark can be placed anywhere within the image or video, either spread out or in one spot, without harming how the image or video looks. However, the watermark is easily removed because of the vulnerability of the least significant bit to various attacks [6]. Furthermore, another approach for a spatial domain is a spatial spread spectrum in which the watermarks spread across the spatial domain depending on the pseudorandom sequence.
On the other hand, transform domain watermarking is characterized by its strength and robustness against common attacks compared with the spatial domain. However, it requires a greater number of computational operations. Therefore, it may not be ideal for applications that require real-time performance and speed or when resources are limited [7]. Watermarks can be embedded into several transforms, as illustrated below.
For example, a watermark can be embedded into frequency components by modifying the frequency coefficients via the discrete cosine transform (DCT). The watermarked image is passed through the inverse discrete cosine transform (IDCT) to obtain the original image and hidden watermark [7]. The discrete Fourier transform (DFT) then decomposes the digital media into a sum of complex exponential functions that are harmonically related. Its output is a periodic discrete signal that is convenient for frequency analysis on digital media [7]. Moreover, to embed a watermark, the discrete wavelet transform (DWT) offers good time and frequency localization and hence provides a better frequency resolution at different frequencies, as digital media are divided into several components called wavelets [7]. Additionally, embedding the watermark in the singular values and subsequently obtaining it via singular value decomposition (SVD) even after tampering is more efficient [7]. Watermarking via SVD is more robust against noise, compression, and attacks. Finally, contourlet transform usage separates digital media into various components to examine essential qualities for watermark concealment. Two primary processes are involved: the Laplacian pyramid (LP) and the directional filter bank (DFB) [8]. The LP process decomposes the media into different frequency bands that effectively capture details at various resolutions, and the DFB process analyses other features of the media in multiple directions.

2.4 Embedding location

A watermark system embeds its information either throughout the entire digital media content or inside selected areas of that media [9]. Local watermarking strategies embed digital watermarks in specific regions of digital content where they can protect certain areas of the media. Specific protection of digital media parts becomes possible through this method in applications. Local attacks such as cropping present risks to this method, but the quality of nonvital areas remains intact. Watermarks provide complete media protection by embedding information throughout the entire content. The technique provides better defense against attacks because it becomes more difficult to remove or alter the data.

2.5 Host media

Various media formats, including images, audio and video files, and text documents, support the watermark application. As a method of image protection, the inclusion of a watermark features transparent copyright marks or pixels embedded with hidden symbols that protect legal documents and photography from unauthorized copying or any type of alteration. Audio files include watermarks that incorporate hidden tones, frequencies, or subtle variations, which preserve the embedded content intact during audio modifications or compression. Videos can incorporate watermarks such as subtitles and logos alongside timestamps, which authenticate original ownership as well as prove video content integrity and nonpiracy to their owners. A text document watermark takes the form of a distinctive barcode, digital signature, or company logo to permanently guard its content against unauthorized use [10].

2.6 Extraction process

Watermarking can be classified into three distinct categories according to process type: nonblind, blind, and semiblind [11] [12]. In nonblind watermarking, both the watermarked content and the original content are required for watermark retrieval. During blind watermarking extraction, only the watermarked media is needed to extract hidden information since access to

the original content is not needed, for example, to protect content that is not accessible, such as copyright protection and multimedia distribution. The process of semiblind watermarking demands the use of both the watermark and the watermarked media to extract hidden information for verifying digital media authenticity.

2.7 Embedding method

Based on the embedding method, watermarking can be classified into two main methods: additive and multiplicative [13]. In additive watermarking, a small amount of noise is added to reveal the unauthorized cover. However, this method can lead to inefficient system implementation in detecting stealthy attacks. On the other hand, in multiplicative watermarking, the system parameters are modified to insert a watermark. This method has the advantage of maintaining the system's performance, and it is better at revealing accuracy. Multiplicative watermarking is generally considered better than incremental watermarking, as it is unable to reduce the performance and provides more effectiveness. Quantization-based watermarking (QIM) is a nonlinear method that is distinct from additive or multiplicative approaches. It embeds watermarks by quantizing host features into predefined intervals. There are other ways to embed watermarks, including deep learning methods and hybrid methods, where deep learning methods such as convolutional neural networks (CNNs), generative adversarial networks (GANs), and transformers take advantage of neural networks to embed watermarks adaptively with host media [14],[15].

3. ATTACKS ON WATERMARKS

With the increasing popularity of watermarks as a means of protecting digital media, they have simultaneously become a target for various attacks, which need to be addressed to maintain the security and integrity of watermarking [16]. Robust countermeasures must be implemented to mitigate the impact of attacks on watermarking systems. Effective techniques include adaptive watermarking, cryptographic security, and advanced detection algorithms. Moreover, successful attacks on watermarking systems may have significant implications, including economic losses for content creators and businesses that rely on copyright protection. As digital media continues to evolve, ongoing research is essential to address emerging threats and enhance the effectiveness of watermarking techniques.

Watermarking attacks can be categorized depending on the aim of the attacker into degradation and removal attacks, as shown in Figure 3. Degradation attacks occur when the watermark is not intentionally targeted for removal but is altered or weakened by various factors. These attacks can increase the detectability of the watermark or integrity and reduce the protective function that should be performed. Common examples of degradation attacks are as follows:

a. Compression: Reducing the file size by eliminating redundant or less important details. When the compression rate is too high, it can distort or even erase the watermark, making it undetectable upon extraction. For example, JPEG compression can negatively impact embedded watermarks, potentially resulting in their permanent loss of detectability.
b. Filtering: Image processing methods such as blurring, sharpening, or median filtering aim to enhance the visual quality of the content. However, these operations may affect the elements containing the watermark, which complicates the watermark extraction process.
c. Noise: Noise introduced by factors such as sensor flaws during capture or transmission errors can degrade the watermark. Additionally, intentional noise addition (e.g., Gaussian or salt-and-pepper noise) for artistic effects may lead to random variations in watermark values, making accurate detection more difficult.
d. Geometric attacks such as rotation, scaling, translation, or cropping change the spatial arrangement of media content. Since many watermarking methods depend on spatial domain integrity, these transformations may misalign the watermark and result in detection failures. Even slight rotations could misplace the watermark.

The removal attacks serve as intentional attacks since their goal is to eliminate or reduce the presence of watermarks from digital content. These attacks work to eliminate or prevent the functioning of embedded watermarks that make the digital content vulnerable and defeat its intended security measures. The main examples of this category of attack include both cryptographic and protocol attacks. The main goal of cryptographic attacks is to protect the watermark's security protocols. Attacks on the authenticity of the original content can be executed by guessing, followed by the extraction or reverse engineering of the encryption keys or algorithms that are used for watermarking purposes unless they claim to compromise the authenticity of the original content [16]. Protocol attacks take advantage of the weaknesses present in the step-by-step processes used in watermarking methods and exploit them. An attacker can eliminate digital watermarks to trick claims of ownership through copy attacks, collusion attacks, or forgery, and the watermarks serve their preventive objectives [17] since they are intended to prevent watermark attacks.

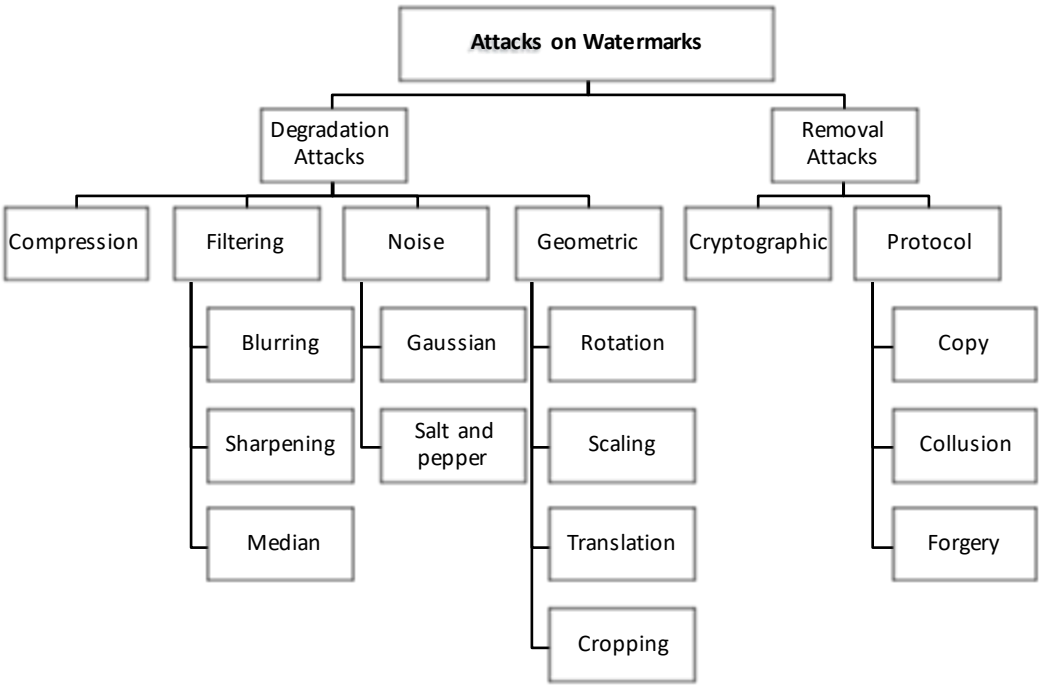

Fig. 3. *Classification of attacks on watermarks*

4. APPLICATIONS OF WATERMARKING IN DIGITAL MEDIA

Digital watermarking can be used for copyright protection by embedding ownership information into digital media. For example, Digimarc's ImageBridge solution adds watermarks to images, which can be identified through software such as Adobe Photoshop. When a user opens a watermarked image, the software detects the watermark and retrieves the owner's contact information, allowing others to request permission for use. This helps protect the rights of the original creator.

A watermark can be used to track the source or recipients of a specific copy of a multimedia file. Each file distributed to different users is embedded with a unique identifier, such as a serial number. These watermarks must remain invisible and resistant to attacks such as compression or filtering. Additionally, the method needs to prevent a "collusion attack," where multiple users might combine their versions to remove the watermark, ensuring that only one unique ID is embedded in any given file.

Watermarks can also prevent illegal copying by embedding a fragile watermark in digital content. When a media player detects this watermark, it will only play the file if a valid watermark is found, stopping unauthorized copying [18].

Forensic techniques and measures to deter piracy involve embedding situational metadata, such as the recipient's IP address, received format, transmission time, and a distinct forensic watermark, into digital content at one or multiple distribution stages. The forensic watermark can be retrieved as evidence, which can help us determine the source of leakage and take legal action.

Broadcast monitoring allows the ownership of digital content to be proven. By embedding data during production and broadcasting, one can detect who, when, and where the content is broadcast. Additionally, metadata such as the author, content type, and title can be included or linked to a database with more details. The watermark extracted from the content can be quickly analysed, and the broadcasting details can be confirmed with radio and TV stations [18].

Locating content online can be accomplished by embedding a watermark ID into digital content, allowing authors to search for their uniquely watermarked content on the internet. The web pages are constantly scanned for a unique watermark, and any results can be reported to notify the content owner for necessary action [18].

Auditing can be improved by having distributors embed an identifier for each licenced asset. This allows any use of the owner's assets, whether in whole or in part, to be automatically and quickly audited via digital watermarking technology [19].

Recent work in generative AI–driven cyber defense shows that GAN and deep reinforcement learning (DRL) such as CryptoGenSec, outperforms static intrusion detection baselines by 95 % in breach prevention success rate. Regarding security, the watermarking process can be integrated into this framework, providing an additional layer of protection that is continuously updated and evolves in parallel with the ongoing development of threats [20].

5. CHALLENGES OF WATERMARKING

Despite advancements in watermarking techniques, several challenges remain. These challenges arise from the need to balance various factors, such as maintaining the original quality of the content, ensuring the watermark's resilience against tampering or processing, safeguarding it from unauthorized removal or attacks, managing the amount of hidden information, minimizing computational costs, and, in some cases, allowing for the recovery of the original content. In what follows, challenges related to imperceptibility, robustness, security, capacity, computational cost, and reversibility are presented [3].

It is essential first to address the issue of imperceptibility, which means that the watermark must remain hidden while ensuring that the quality of the original content is preserved. Ideally, the watermarked content should look exactly like the original content, with no noticeable changes. To accomplish this, specialized methods must embed the watermark without causing any visible alterations [18].

Next, attention to the robustness of the model that embeds the watermark is important and considered fundamental to the process. The watermark must resist a variety of processing operations, such as compression, resizing, or filtering, without losing its detectability. Watermarking techniques differ in their ability to withstand these types of changes. Some worked well against attacks but failed against others. Therefore, the continuous development and improvement of watermarking techniques to counter different attacks are needed [2].

In addition, the security of watermarking is an important issue to consider. Powerful and new encryption methods are needed to maintain security, which requires improvements to continue combating attacks [4].

Furthermore, capacity refers to the amount of information to be embedded in the host. The challenge is to embed a watermark without introducing visible distortions. As more data are embedded, the probability of degradation increases, which is particularly concerning in applications where clarity and precision are critical [5].

Furthermore, computational cost is a significant issue in the embedding and extraction processes. It is possible to produce robust watermarks with high capacity, but the computational load can be heavy, especially in real-time and online applications. Therefore, a balance between accuracy, watermark strength, and cost is required [5].

Finally, a critical issue in watermarking is reversibility, in which the watermark can be extracted and the host can be restored [4].

6. BENCHMARK DATASETS OF WATERMARKING

Watermarking techniques can be evaluated on different benchmark datasets, so it is important to examine the dataset that is used in assessing and comparing watermarking techniques. Figure 4 depicts the datasets commonly used in watermarking. There are several image and video datasets with different characteristics, as reported in Table I.

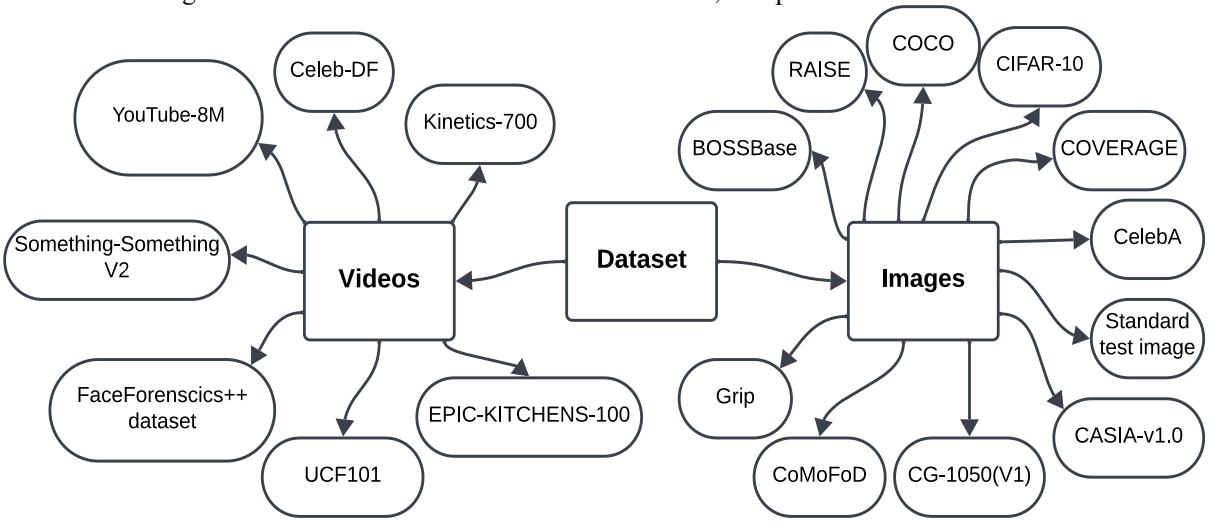

Fig. 4. *Categorization of Video and Image Datasets Used for Watermarking*

TABLE I. AVAILABLE IMAGE AND VIDEO DATASETS FOR WATERMARKING

Dataset	*Type*	*Size*	*Year*	*Link*	*Access level*

Standard Test Image	Image	~1,700 images (multiple categories)	1997	http://sipi.usc.edu/database/	Public
CG-1050(v1)	Image	1,050 images (computer-generated vs. photographic)	2005	Shared by authors/research groups	Public/Request
CIFAR-10	Image	60,000 32×32 color images (10 classes)	2009	https://www.cs.toronto.edu/~kriz/cifar.html	Public
GRIP	Image	Varies (often a smaller set for image forensics)	2010	Shared by research groups	Public/Request
BOSSBase	Image	10,000 grayscale images (for steganalysis)	2010	https://www.kaggle.com/datasets/lijiyu/bossbase	Public
CoMoFoD	Image	Up to 260 base images (several subsets; thousands total)	2012	http://www.vcl.fer.hr/comofod	Public
CASIA-v1.0	Image	1,721 images (800 authentic + 921 tampered)	2013	http://forensics.idealtest.org/	Public
Common Objects in Context	Image	>330,000 images	2014	https://cocodataset.org/	Public
RAISE	Image	8,156 RAW images	2015	https://loki.disi.unitn.it/RAISE/	Public
COVERAGE	Image	~100 base images (copy–move forgery dataset)	2017	https://github.com/wenbihan/coverage	Public/Request
CelebA-HQ	Image	30,000 high-resolution face images	2017	https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html	Public
UCF101	Video	13,320 video clips	2012	https://www.crcv.ucf.edu/research/data-sets/ucf101/	Public
YouTube-8 M	Video	6.1 million videos	2016	https://research.google.com/youtube8m/	Public
Something-Something V2	Video	220,847 video clips	2017	https://paperswithcode.com/dataset/something-something-v2	Public
FaceForensics ++	Video	1,004 videos (500,000+ frames)	2018	https://niessnerlab.org/projects/roessler2018faceforensics.html	Request
Celeb-DF	Video	~5,639 videos	2019	https://cse.buffalo.edu/~siweilyu/celeb-deepfakeforensics	Public
Kinetics-700	Video	~650,000 video clips	2019	https://github.com/cvdfoundation/kinetics-dataset	Public
EPIC-KITCHENS-100	Video	~100 hours of egocentric video	2020	https://epic-kitchens.github.io/	Public

6.1 Image datasets

Image datasets for watermarking research play a crucial role in evaluating and comparing the performance of various watermarking techniques. In this work, the datasets selected for experimentation were meticulously selected because of their broad acknowledgement and regular use in prominent watermarking research publications.

One of the prominent datasets is the Common Objects in Context (COCO) dataset, which offers diverse 1000-color images showing complex scenes with various objects in real-world settings [21]. Another significant dataset is RAISE, in which native images are captured via a Nikon D90 camera, comprising 8170 high-luminance, uncompressed images [22]. The

COVERAGE dataset was forged with copy-move images, including annotations, with dimensions of 400 × 486 pixels, consisting of 200 images (100 original and 100 tampered) [23].
Additionally, the Standard Test Image (USC-SIPI Image) dataset offers images in various dimensions, such as 256 × 256, 512 × 512, or 1024 × 1024 pixels, and includes grayscale images at 8 bits/pixel and color images at 24 bits/pixel that are used to support image processing and machine vision research. It is a benchmark for comparing new watermarking techniques with existing methods on standard images such as Lena, Baboon, and Barbara. SIPI also includes 44 benchmark images, 16 colors, and 28 monochromes of different sizes: 14 at 256 × 256, 26 at 512 × 512, and 4 at 1024 × 1024 pixels [24].
In the domain of deep learning-based watermarking, CASIA-v1.0 is a commonly used dataset consisting of 800 original and 921 tampered JPEG images with dimensions of 384 × 256 pixels [25]. Another notable dataset is CG-1050 (v1), which contains 1050 images of various dimensions organized into two main directories: training and validation with subdirectories for original and manipulated cropped images, where the manipulated area ranges from 25% to 75% of the cropped region [26]. Moreover, CoMoFoD consists of multiple sets of forged images available at two resolutions: smaller images at 512 × 512 and larger images at 3000 × 2000, which are saved in JPEG and PNG formats. Images are grouped into five manipulation types: translation, rotation, scaling, combination, and distortion. Both authentic and manipulated images undergo various postprocessing techniques, such as JPEG compression, blurring, noise addition, and color reduction [27]. The GRIP dataset is another important resource that includes ground truth images that align with various forgery types, including copy-paste, rotation, noise, scaling, and JPEG compression [28].
Among datasets focusing on human imagery, CelebA-HQ is a high-quality dataset consisting of 30,000 images at 1024×1024 resolution, images of human faces that represent 6,217 unique identities. It contains three groups for training, validation, and testing [29]. Additionally, the CIFAR-10 dataset contains 60,000 color images with a size of 32x32 pixels that are categorized into 10 different classes, with 6,000 images per class. It includes 50,000 images for training and 10,000 images for testing [30]. Finally, BOSSBase consists of 10,000 grayscale images of size 512x512 that were taken via seven different cameras in portable gray map format [31].

6.2 Video Datasets

In digital watermarking, video datasets serve as foundational tools for evaluating the effectiveness and resilience of watermarking techniques. Video datasets typically encompass a range of resolutions, frame rates, and motion dynamics, ensuring comprehensive assessments of watermarking performance.
Among the notable datasets in this field is Celeb-DF, a large DeepFake dataset with 5,639 videos and over 2 million frames from YouTube clips of 59 celebrities. The videos were produced via better synthesis techniques with fewer visual issues to make Celeb-DF a challenging dataset for testing and advancing DeepFake detection algorithms [32]. The dataset serves as a common evaluation tool for researchers to test detection systems. [33].
Another significant resource is Kinetics-700, with 650,000 clips that cover 700 human action classes. The videos show people interacting with objects, such as playing instruments, as well as people interacting with each other, such as shaking hands and hugging. Each action class has at least 700 video clips. Each clip is labelled with an action class and is approximately 10 seconds long [34].
Additionally, among the largest video datasets available to researchers, YouTube-8 M contains 6.1 million YouTube videos. These videos cover 3,862 categories, from cooking to extreme sports, which we have aggregated into 24 high-level verticals for easier browsing. This dataset contains a subset of more than 237,000 meticulously annotated portions of video, 5 seconds long, from 1,000 categories. YouTube-8 M was originally announced with "8 million" videos in it (hence the name) and has become a default resource for anyone looking for a source for video classification and understanding research, owing to rich content without needing to store terabytes of data [35].
Something-Something V2 provides a large collection of labelled video clips that capture basic human actions with everyday objects, including 220,847 videos, with 168,913 for training, 24,777 for validation, and 27,157 for testing, covering 174 labels [36]. EPIC-KITCHENS-100 is a collection of 100 hours, 20 M frames, and 90K actions in 700 variable-length videos [37]. UCF101 is an expanded version of UCF50, comprising 13,320 video clips into 101 categories, including body movements, human interactions, human-object interactions, playing musical instruments, and sports, with a total video time of over 27 hours, a fixed frame rate of 25 FPS and a resolution of 320x240, sourced from YouTube [38].
Finally, FaceForensics++ is a dataset of 1000 original videos modified via four automated facial manipulation techniques: Deepfakes, Face2Face, FaceSwap, and NeuralTextures. The videos were sourced from 977 YouTube videos, and all contained mostly frontal, trackable faces without occlusions, enabling the creation of realistic forgeries [39].

7. REVIEW OF THE EXISTING METHODS

7.1 Methodology

This review was conducted by performing a literature search of existing studies on digital watermarking techniques from 2016--202525. A complete literature search was carried out through Google Scholar, IEEE Xplore, and Scopus via Boolean-formatted key phrases (e.g., "digital watermarking," "deep learning-based watermarking," and "copyright safety"). The initial results (250 studies) were screened for relevance on the basis of titles, abstracts, and publication quality. Nonpeer-reviewed articles, nonEnglish papers, and studies missing empirical validation data were excluded. After duplicates were removed and rigorous filtering was applied, 150 outstanding papers were retained for full-textual content overview. To ensure credibility, priority was given to legitimate journals (e.g., IEEE, Elsevier, and Springer) and recent peer-reviewed meetings. From those, 69 formative works were selected. We extracted the primary data from the selected studies, including the year of publication, research methods, main contributions, and empirical data. These data were organized into tables to facilitate comparisons between different studies. Next, we discuss the research, trends, benefits, and limitations, as the aim of the presented review research is to provide deep and valuable insights to researchers in the field of watermarking.

7.2 Literature on Traditional Techniques

Several traditional watermarking techniques exist; however, they typically fail to address the balance of imperceptibility, robustness, and capacity. This section identifies and addresses significant current gaps. Table II summarizes the main differences between these techniques.

A method for embedding a watermark into the host image via fractals, which are mathematical shapes that exhibit self-similarity and irregularity, was proposed in [40]. The watermark is encoded and embedded into selected blocks of the image via fractal coding. The blue component of the image is used for the watermarking process because it is less sensitive to noise attacks. The results were strong and stable, and the famous Lena and Baboon images with signal-to-noise ratios above 40 dB were used. The use of fractals for watermarking takes advantage of their self-similarity and irregularity, making them resistant to noise attacks. However, the method's performance is measured via images such as Lena and Baboon images, which may not fully represent the diversity of images encountered in real-world applications.

The quantization index modulation (QIM) technique for hiding watermarks in 3D meshes to protect ownership was proposed. It modifies geometric properties, including distances between vertices and directions, to embed hidden information [41]. Techniques, including transformation, rotation, and noise addition, were used to increase the resistance of the watermark to attacks. The Princeton 3D mesh benchmark dataset was used for evaluation, and the results revealed a high capacity for watermark concealment, distortion control, and strong resistance to attacks. However, the reliance on geometric transformations such as rotation and noise addition to resist attacks could raise concerns about the visual quality of 3D models.

A patchwork technique for audio watermarking [42] that embeds a secret image as a watermark into the audio signal via LSB encoding and Base64 was proposed. The WAV audio files "future bells.wav" and "bars.wav" were used, and the results revealed high audio and image quality. The PSNR for the audio data was 30 dB, and for the image, it was 40 dB, with a very low bit error rate of 0.036%. The method might struggle with the trade-off between robustness and imperceptibility, especially when subjected to more severe audio transformations such as compression, reverb, or filtering.

DWT and DCT were used to embed a watermark within the host media to take advantage of the frequency domain properties of both DWT and DCT to secure copyright protection [43]. The image is divided via the DWT method into subbands, with a focus on the HL1 and HH1 bands in the process of embedding the watermark with the host image. This allows the watermark to be embedded in the midfrequency band coefficients of 16x16 blocks via the DCT method. The dataset includes the famous “Lena” image sized 512x512 as the cover image and another grayscale image sized 256x256 as the watermark. The results of the hybrid method showed a significant performance improvement compared with using either DWT or DCT separately, especially in the embedding and extraction of the watermark under different conditions. The method's reliance on specific subbands (HL1 and HH1) for embedding could limit its adaptability to different image types or sizes.

In [44], the spread spectrum technique was used to hide the watermark within the image while preserving its quality. The spread spectrum works by distributing the watermark across a wide frequency range, making it robust against noise attacks. Researchers have also used the discrete wavelet transform (DWT) to embed the watermark in the low-frequency subbands of the DWT. The IIUC image was used as the dataset, and after applying attacks such as noise and rotation, strong tamper resistance with an effectiveness of approximately 80% was achieved.

SVD and DWT were used to embed watermarks in printed and scanned images. First, the image was transferred into the DWT, and the watermark was embedded in the singular values of the low frequency of the DWT [45]. The experiments were conducted on Lena and Baboon images and demonstrated that the method strongly resists distortion, with PSNR values reaching 30 dB, ensuring minimal image degradation. However, this method's effectiveness in real-world settings where scanning and printing introduce variations in image quality remains to be fully examined.

A method for embedding watermarks in images that combines DFT and particle swarm optimization (PSO) to balance robustness against attacks while preserving the imperceptibility of the embedded watermark was proposed [46]. The PSO algorithm was used to improve watermarking by optimizing the embedding region. The dataset used in this work includes

1000 color images from the COCO dataset, and the results show that PSO harnesses the robustness of watermarking compared with other methods.

[47] produced a novel image watermarking technique that benefits the integer wavelet transform (IWT) power and parity-bit checking to detect tampering effectively and ensure authentication. The IWT is based on the lifting scheme to provide a direct way to convert image pixels' integer values to integer coefficient values rather than floating point coefficients. This method produces a robust approach to safeguarding the integrity of the original image by embedding the watermark bits in the low- and mid-frequency components of a two-level wavelet decomposition and maintaining even parity in each block. The results obtained on the Lena, Girl, and Barbara datasets, with PSNR values reaching up to 46.56 dB and SSIM values exceeding 0.99, underscore the efficacy of this technique. The technique's robustness and high PSNR and SSIM values suggest its effectiveness, but further comparison with other advanced methods and its scalability to handle large datasets or high-resolution images would be beneficial.

DCT and linear modulation were used for embedding the watermark in grayscale images [48]. First, the image was divided into 8x8 blocks, and then the DCT transform was applied. The watermark is then embedded using the least significant bits (LSB). This hybrid technique ensures both imperceptibility and robustness against attacks. The experiments were conducted on Lena and Baboon, and the quality was high, with PSNR values exceeding 40 dB. The results demonstrated that the watermark extraction was successful without any significant degradation in its quality or the host image, maintaining a balance between the imperceptibility and robustness of the watermark. However, the scalability of this technique for larger or more complex images and its resistance to other attacks require further study.

The least significant bit (LSB) technique was used to hide the watermark integrated with Canny edge detection [49]. Canny's method selects the best locations in the image for embedding the watermark to avoid embedding in smooth areas of the image. Additionally, an extra layer of protection was added by scrambling the watermark via a chaotic substitution box. This method remains simple but offers an enhancement in overall performance. The SIPI image dataset with grayscale images. The results show that the method provides minimal visual distortion and strong robustness against attacks, and the watermark was successfully extracted without degradation. However, the method may struggle with highly textured or complex images.

A perceptual watermarking system was proposed to combat deepfake face-swapping techniques by embedding invisible watermarks by personal identity information into the original image, allowing the detection of facial manipulation and tracing of the source of the image. The CelebA-HQ and LFW datasets are used, and the results demonstrate a 96% success rate in watermark extraction and a 97% success rate in deepfake detection [50]. However, this method may be limited in its scope, as it focuses primarily on face-swapping detection.

[51] introduced a video watermarking approach that leverages Galois field (GF) multiplication tables, with a focus on three irreducible polynomials, and an adaptive thresholding technique for scene-based frame selection. By watermarking only 20% of the video frames, the method enhances efficiency while preserving video quality. Experiments were conducted on six videos. The results demonstrate that using the irreducible polynomial $x^4 + x^3 + 1$ yields the highest PSNR values, reaching over 53 dB when embedding smaller watermarks. This method remains strong against various attacks while maintaining a high level of similarity and normalized correlation (NC) between the original and watermarked videos. On the other hand, the adaptive threshold selects the most relevant frames without making the watermark noticeable. Overall, it provides a strong balance between watermark invisibility, computational efficiency, and resilience across diverse video scenarios. However, the method was tested on a limited set of videos, and it may not perform as effectively with highly dynamic or fast-moving scenes.

The hybrid method was proposed in [52], which combines SVD and the Mojette transform to enhance the trade-off between imperceptibility, robustness, and capacity in image watermarking. By applying Mojette projections to the singular values of both the host and the watermark images, the method embeds the watermark into geometrically distributed bins that maximize payload capacity while minimizing perceptual impact. The results achieved a high imperceptibility of PSNR equal to 45.2373 dB and robustness against geometric and nongeometric attacks, with normalized cross-correlation (NCC) values exceeding 0.95. However, the testing was confined primarily to gray images, and the impact of payload size on performance remains undefined.

[53] proposed a method for hiding the blind watermark inside color images via the fast four-dimensional qubit decomposition method FQSD, which is performed by encoding the color channels of the RGB image jointly in a four-dimensional matrix. The watermark is hidden above the energy coefficients by modifying the quantization index modulation, as FQSD allows the matrix to be decomposed into upper triangular blocks. The watermark is encrypted by a two-dimensional logistic-adjusted sine map (LASM) chaotic map to mix the pixel distributions. The method achieves a PSNR close to 38 dB, an SSIM of 0.95, and a robustness of 0.96 NC against multiple attacks (JPEG compression, noise). The drawback of this method is that it incurs a high computational cost compared with other spatial methods.

A robust method for digital watermarking that combines the dither modulation (DM) algorithm with just noticeable distortion (JND) perceptual models to address modern and advanced attack challenges was proposed in [54]. The algorithm enhances traditional quantization-based watermarking by adaptively adjusting the embedding strength via Weber’s law and

Watson-based JND thresholds, which consider the adaptation of brightness and contrast within and between blocks and the sensitivity of the DCT coefficient. The method achieved success when tested on the BOSSbase dataset, and the results were as follows: PSNR greater than 22 dB and bit error rate (BER) less than 0.3 when traditional attacks were used. For DnCNN attacks based on artificial intelligence, the result was a BER of 0.26, outperforming DC-only. However, it does not include the extent to which the JND model affects the quality of visual perception of the watermark in terms of visual invisibility. A watermarking method was proposed to embed a watermark in the DFT amplitude–frequency coefficients via polar coordinate mapping, and a machine learning model was used to estimate the scaling parameters to accurately recover the watermark [55]. The proposed method was tested on the COCO2017 and BOSSBase-1.0 datasets. It achieved high visual quality with a PSNR of ~39 dB and an SSIM of 0.9707 on COCO2017. The results for BOSSBase were a PSNR of ~44 dB and an SSIM of 0.9816. Watermarks are successfully extracted from very small regions up to 10% of the original image. The proposed method uses both polar coordinate transformation and a machine learning model to estimate the scaling parameters. This may result in an increased computing load.

TABLE II. COMPARISON OF DIFFERENT TRADITIONAL WATERMARKING TECHNIQUES

Reference	*year*	*Technique*	*Description*	*Dataset*	*Advantage*	*Disadvantage*	*Results*
[40]	2016	Fractal Coding	Using self-similar fractal shapes to embed watermarks in specific image blocks.	Lena and Baboon images	robust embedding and blue component utilized to minimize noise sensitivity	Limited to specific image blocks	High stability with SNR above 40 dB, demonstrating robustness against noise
[41]	2017	Quantization Index Modulation	Embedding hidden information by modifying geometric properties in 3D meshes.	Princeton 3D mesh benchmark dataset	Strong watermark concealment, distortion control, and resistance to transformation, rotation, and noise attacks	Limited to 3D mesh applications	High capacity for watermark concealment with strong resistance to various types of attacks
[42]	2019	Patchwork	embedding a secret image in audio signals using LSB encoding and Base64.	"future bells.wav" and "bars.wav" WAV audio files	High audio and image quality with low bit error rate; adaptable for audio watermarking on low-power devices like Raspberry Pi	Limited to audio applications	PSNR of 30 dB for audio, 40 dB for image, with very low bit error rate of 0.036%
[43]	2019	Hybrid DWT-DCT	Combining DWT and DCT techniques to embed watermarks in mid-frequency band coefficients.	Lena image (512x512) and grayscale image (256x256)	Utilizes the frequency domain properties of DWT and DCT, improving robustness and efficiency in embedding and extraction	Complexity in combining two transforms may increase the computational load	Significant performance improvement in watermark embedding and extraction compared to using DWT or DCT alone
[44]	2020	Spread Spectrum	Distributing watermark across a wide frequency range; implemented using DWT.	IIUC image library	Distributes watermark across a wide frequency range, enhancing robustness against noise	higher computational resources due to the frequency spread	Achieved approximately 80% resistance to tampering, with strong effectiveness against noise and rotation attacks
[45]	2020	Hybrid SVD-DWT	Embedding watermarks using SVD and DWT, maintaining robustness in printed/scanned images.	Lena and Baboon images	Strong resistance to printing and scanning distortions; minimal impact on image quality	Limited to low-frequency embedding, may reduce flexibility in high-frequency areas	Achieved PSNR values above 30 dB, ensuring minimal image degradation even after printing and scanning
[46]	2021	DFT with Particle Swarm	Optimizing embedding region to balance	COCO (1000 color images)	Balances robustness and imperceptibility; resists geometric	require high computational resources for optimization	Significant improvement in robustness and image quality,

		Optimizatio n	robustness and imperceptibility		distortions (e.g., rotation, scaling); strong against JPEG compression and noise		outperforming traditional methods in real-world scenarios
[47]	2021	Integer Wavelet Transform	Utilizing the lifting scheme for watermarking in low/mid-frequency components.	Standard grayscale images: Lena, Girl, Barbara	Utilizes IWT for integer-to-integer transformations, avoiding floating-point rounding; maintains high imperceptibility and robustness; detects tampered regions effectively	Some tampered blocks may not be detected if parity remains unchanged; limited robustness due to parity bit vulnerabilities.	PSNR ≈ 44.4367 dB (Lena: 44.2416, Girl: 46.5576, Barbara: 42.5110); SSIM ≈ 0.9956; Embedding time ≈ 15.7 seconds (Lena), Extraction time ≈ 16.4 seconds (Lena)
[48]	2023	DCT with Linear Modulation	Dividing images into 8x8 blocks, applying DCT, and embedding watermarks using LSB.	Lena, Baboon (grayscale images)	Ensures imperceptibility and robustness; effective against attacks like JPEG compression, Gaussian noise, and rotation	Limited to grayscale images	High-quality results with PSNR > 40 dB; successful watermark extraction without significant degradation in quality or host image
[49]	2023	LSB Watermarki ng with Canny Edge Detection	Embedding watermark in selected areas using Canny edge detection, with added scrambling for protection.	SIPI (grayscale images 512x512, watermark 32x32)	Simple and efficient; improved security through Canny edge detection and chaotic substitution; resistant to noise and geometric attacks	Limited to grayscale images; potentially weak to direct detection in smooth areas without Canny.	High PSNR, SSIM, and NC values; minimal visual distortion; robust against salt-and-pepper noise, Gaussian noise, cropping, and rotation; successful watermark extraction without quality loss
[50]	2024	Perceptual	Embedding invisible watermarks containing personal identity information for deepfake detection.	CelebA-HQ, LFW	Enables detection of face manipulation and tracing of image source; high success rates in extraction	Limited application scope, primarily focused on face-swapping detection	96% success in watermark extraction, 97% success in deepfake detection
[51]	2024	Video Watermarki ng with Galois Field Multiplicati on	Utilizing Galois Field multiplication for watermarking scene changes in videos; adaptive thresholding employed.	Private Camera Data and Public Web Data (Foreman, Akiyo, Coastguard videos; watermark images with 225×225 and 100×100 resolutions)	High imperceptibility, robustness, adaptive scene selection, dynamic thresholding, low computational time	Limited performance with very short videos due to insufficient scenes; trade-off between imperceptibility and robustness	PSNR values: ~54.54 dB (Akiyo), ~53.25 dB (Foreman), ~52.91 dB (Coastguard); Robustness metrics with NC > 0.98 under various attacks
[52]	2024	SVD with Mojette Transform	Combining Singular Value Decomposition and Mojette Transform to embed	Standard test images (e.g., Lena, Cameraman)	High imperceptibility (PSNR ~45 dB. Robust to noise, filtering, compression,	Payload capacity requires further analysis; Tested only on grayscale images.	PSNR of 45.24 dB at 1% scale factor; Extracted watermarks show NCC >95% under attacks

			watermarks in projection bins of host images. Mojette Transform enhances robustness through geometric projections.		rotation, and cropping (NCC >95%).		(noise, filtering, compression, rotation, etc.).
[53]	2024	Fast Quaternion Schur Decomposition (FQSD)	Embedding watermarks using quaternion Schur decomposition and quantization index modulation (QIM), leveraging color channel correlations.	Standard images (Athens, Peppers, Butterfly1, etc.) from CVG-UGR/USC-SIPI	High efficiency (structure-preserving algorithms), robustness to common/geometric attacks, and color synchronization.	Requires image correction for geometric attacks	PSNR >38 dB, SSIM >0.95, NC >0.96 (common attacks), NC >0.90 (geometric).
[54]	2025	DCT-domain Dither Modulation (DM) + JND	Combining DCT quantization with perceptual JND models (Weber/Watson) to embed watermarks adaptively. Resists AI-driven attacks (DnCNN), volumetric scaling, and histogram equalization.	BOSSbase101	Robust against modern AI attacks; adaptive embedding strength via HVS models.	Computationally intensive due to JND calculations; limited to grayscale images.	PSNR >22 dB (imperceptibility), BER <0.3 (traditional attacks), BER 0.26 (DnCNN attacks).
[55]	2025	DFT-based polar coordinate mapping + ML	Embedding watermarks in DFT frequency domain; uses ML for scaling parameter estimation and padding for arbitrary resolutions.	BOSSBase-101, COCO2017	Robust to arbitrary scaling/cropping and supports any resolution; it has a high PSNR.	Performance degrades with <10% cropped area; ML adds computational overhead.	PSNR ~44 dB (BOSSBase), ~39 dB (COCO); BEQ as low as 0.0090.

7.3 Literature on Deep Learning-Based Watermarking

This section presents the watermarking literature, which is based on deep learning algorithms that have been recently used to improve the performance of watermarking algorithms. Table III summarizes the main differences between these techniques.

In [56], an end-to-end trainable deep learning framework was presented as HiDDeN for data hiding that is applicable to steganography and digital watermarking. This framework relies on CNNs for encoding and decoding data, ensuring that hidden information remains invisible and can withstand distortions such as noise. This makes it suitable for real-world applications where images are compressed, cropped, or otherwise modified. The framework integrates an encoder network to embed the watermark into the cover image and produces a watermarked image and a decoder network to reconstruct and extract the watermark, even when the image has been distorted. The addition of an adversary network allows the model to be trained to detect whether an image contains hidden data, providing an adversarial loss that improves the quality of watermarked images by making them harder for adversaries to detect. Image distortion loss (L2) ensures that the watermarked image closely resembles the original image, and message distortion loss (L2) maximizes the accuracy of the decoded message. The adversarial loss further ensures that the watermarked image is indistinguishable from the original image in terms of detection by adversaries. The framework was tested against various types of noise, such as Gaussian blurring, pixelwise dropout, cropping, and JPEG compression, with noise layers placed between the encoder and decoder to ensure robustness. The COCO dataset was used for training, and the BOSS dataset was used for testing in steganalysis

experiments, with the results showing that the method can hide large amounts of data 0.203 bits per pixel (BPP) with minimal distortion, demonstrating high capacity and robustness against common noise types such as Gaussian blurring and JPEG compression. However, this method may require large datasets for training and could benefit from optimization for real-time applications where computational efficiency is crucial.

In [57], an image watermarking method with unsupervised deep learning was proposed to automate watermark embedding and extraction. The watermarking was treated as an image fusion task, blending the watermark and cover image features for invisibility while ensuring resilience. The GAN network includes an embedder for embedding, an extractor for retrieval, and an invariance layer that enhances fidelity, robustness, and correlation between the watermark and fused image features. This unsupervised training enhances the system's resilience against image-processing distortions and enables challenging applications to extract watermarks from camera-captured images. Using ImageNet (128x128 datasets for training, COCO for cover images, and binary CIFAR as watermark test sets), real-world testing involved mobile phone-captured image resamples. The results revealed high fidelity with a PSNR of 39.72 dB and low BER rates under distortion: 11.6% for JPEG (Quality 10), 7.8% for 65% cropping, 32.2% for 20% Gaussian noise, and 12.3% for 90% salt-and-pepper noise. However, the method's performance may decrease when it is exposed to more complex distortions or images with more intricate details.

A CNN-based watermarking method was used in [58] to protect the intellectual property of digital images. It is blind, robust, and invisible. These networks adjust the resolution of the watermark to match that of the host image. The BOSSbase dataset contains 10,000 images for training purposes, and the watermarks are generated randomly. The results report an average PSNR of 43.23 dB during training and 40.58 dB during evaluation. The method was tested against several attacks, such as pixel value changes and geometric attacks, Gaussian filtering, median filtering, and salt-and-pepper noise. The results show the method's resistance to various attacks with BERs lower than 0.1. However, the random generation of watermarks could result in challenges in watermark consistency, and further refinement might be required for fine-tuning watermark embedding.

A deep learning-based watermarking model named ReDMark in [59] combined two fully convolutional neural networks (FCNs) with residual structures for embedding and extraction to improve speed and robustness. CIFAR10 and Pascal VOC2012 are adopted for training, and Granada's dataset is used for testing. Additionally, various simulated attacks, including JPEG compression, Gaussian noise, and other real-world distortions, were applied to the dataset to allow the watermark data to diffuse across a wide area of the image. The results show that the model has high resistance to attacks, with a PSNR of 35 dB and a structural similarity index (SSIM) of 0.96. While the model’s results showed promising performance under simulated attacks, it may face challenges when exposed to more advanced adversarial techniques or large-scale data.

In [60], a supervised GAN watermarking model was presented to embed invisible watermarks into images to verify ownership and protect intellectual property. The structure of a GAN consists of a pretrained encoder-decoder network that injects a watermark into each image, allowing for efficient and seamless ownership verification. The decoder remains frozen during GAN training, guiding the watermark’s embedding, whereas a combined loss function ensures that the watermark is preserved in generated images, even under fine-tuning. The model’s robustness is further strengthened by data augmentations—such as JPEG compression, noise addition, and color transformations—applied during training, enabling the watermark to withstand various postprocessing distortions. The CelebA, LSUN-Bedroom, FFHQ, and VGG Flowers datasets are used for model evaluation. The model achieves watermark bit accuracy (approximately 99%) and retains high image quality, with PSNR values above 45 and SSIM values above 99%. However, its reliance on pretrained networks could limit its adaptability to new types of distortions and data.

In [61], a deepfake prevention method combining a GAN and 3D CNN was proposed to invisibly embed encrypted watermarks in video frames. This method starts by creating an attention model by training the 3D CNN to identify features within frames, guiding the optimal watermark placement. The GAN then uses this attention model to seamlessly embed a human-invisible watermark into the frames. UCF101 was used for training, Hollywood2 was used for validation, A2D was used for testing, and TikTok trending videos were used for additional short-form social media validation. Testing included watermarked videos subjected to deepfake attacks with DeepFaceLab 2.0, where tampering was detected by comparing original and altered watermark probabilities. This method demonstrated impressive performance, achieving 99.7% accuracy in embedding, a 100% success rate in deepfake prevention, a PSNR of 13.98 (indicating low noise and high quality), and an SSIM of 0.301 (showing high similarity and low distortion).

In [62], the ARWGAN model was developed for watermarking via an encoder-decoder GAN framework enhanced by attention mechanisms and feature fusion to overcome limitations. The model’s feature fusion module (FFM) integrates shallow and deep features across layers, boosting watermark robustness. Moreover, an attention module (AM) generates an attention mask that highlights regions ideal for embedding and focuses on textured and less noticeable areas for minimal visual impact. Additionally, the noise subnetwork simulates attacks such as cropping, blurring, and JPEG compression to improve robustness against real-world distortions. The ARWGAN was trained on COCO with over 100,000 images and tested on 3,000 images, with further cross-dataset tests on Pascal VOC2012, ImageNet, and EUVP. The model excelled in

both image quality and robustness: on Pascal VOC2012, it achieved an average PSNR of 36.83 dB and an SSIM of 0.9698; on COCO, it scored 35.87 dB in PSNR and 0.9688 in SSIM. ARWGAN's robustness averaged 98.39% on COCO, outperforming competitors such as MBRS (96.92%), and on Pascal VOC2012, it maintained 98.89% robustness, achieving 94.45% under JPEG compression (quality 50) and 100% against Gaussian noise ($\sigma = 2.0$), consistently surpassing other models across all conditions.

In [63], a hybrid architecture for the encoder-decoder framework that combines CNNs and transformers was introduced. The encoder extracts primary features from the cover image via convolutional layers, whereas the decoder integrates CNNs and Swin transformers to capture both local and global image features. The multiscale attentional feature fusion module (MA-FFM) merges local and global feature contexts, allowing more effective interaction between local and nonlocal pixels. This method was trained on the COCO dataset. The results demonstrated that the imperceptibility achieved PSNR > 40 dB and SSIM > 0.95 across various noise types, including Gaussian, salt and pepper, dropout, and JPEG compression. While this method has significant potential, further exploration of its performance across different media types, such as video or audio, and its computational efficiency could help assess its real-world applicability.

The DNN watermarking introduced in [64] includes the Kuribayashi white-box approach to embedding watermarks into convolution layers (e.g., VGG16, ResNet50) via constant weight codes (CWCs). By encoding watermarks into fixed-weight codewords and adjusting thresholds for weight parameters, the method achieved an accuracy of 97% on flower datasets. However, the method focused on convolutional layers, which restricts its applicability, especially for architectures with fewer parameters in these layers. Furthermore, the dependency on specific models and datasets raises concerns about generalization across diverse contexts.

[65] used a DNN watermarking method that prevents the modification of model weights by encoding watermarks through secret key-based matching of weight binary codes in convolutional layers. The experiments revealed that the watermark remained robust against Gaussian noise attacks, with a bit error rate (BER) of 6.2% when noise occurred with a standard deviation of 0.5 times. The limitation of the method was not tested against advanced pruning and fine-tuning, and it depends on specific model architectures.

In [66], a CNN with a quantized activation function (QAF) uses a series of hyperbolic tangent (tanh) functions across the image's DCT. For training, 50,000 images from the CIFAR-10 dataset were used, and for testing, a set of 49 grayscale images (512x512 pixels) was used. The evaluation metrics, including the structural similarity index measure (SSIM) and peak signal-to-noise ratio (PSNR), are used to assess the image quality of the watermarked images. The QT-QAF-Net method demonstrated the effectiveness of the QAF layer in enhancing robustness while maintaining high image quality. However, while the method shows promise, it may benefit from more extensive testing with larger datasets and additional distortions to assess its scalability and robustness in real-world applications, particularly for complex images or videos.

In [67], a GAN-based information-hiding technique for digital images named GANMarked was proposed. It consists of three main components: an autoencoder network that securely encodes watermarks by merging two watermark images into one, adding an extra layer of security and efficiency to watermarking, and a decoder that allows the recovery of each watermark from the encoded representation. A distinctive feature of GANMark is its secure trigger key, which protects the model itself, not just the watermark, ensuring model ownership verification and making it highly useful for protecting intellectual property in artificial intelligence (AI) and machine learning models. The datasets used include the DIV2K (a high-quality image set), COCO (for diverse robustness testing), and COVID-19 and Cats & Dogs datasets, adding variety for cross-domain testing and evaluating metrics such as the peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), normalized correlation (NC), and bit error rate (BER), and GANMarked achieves high PSNR and SSIM scores, indicating that the watermark is invisible to human eyes. With an NC of 0.999 and a BER of 0, it displayed perfect robustness and resilience against hybrid attacks, demonstrating its strength and practicality for real-world applications. Despite these strengths, future research could explore the model's application to more varied datasets and evaluate its performance under extreme conditions, such as highly compressed images or video data, to confirm its resilience in real-world scenarios.

In [68], the proposed Npix2Cpix method combines a U-Net-based conditional generative adversarial network (GAN) for denoising watermarked images with a Siamese network designed for one-shot classification. The GAN component effectively removes handwriting and background noise, achieving a PSNR of approximately 24 dB and an SSIM of approximately 0.85, demonstrating its ability to manage class variances while leveraging one-shot learning to contend with limited datasets. However, the method encounters difficulties in extremely noisy environments and struggles to scale effectively to accommodate various watermark designs.

[69] introduced neural network-based watermarking that integrates a quantized activation function (QAF) into a conditional GAN framework, allowing it to accurately mimic the standard JPEG quantization process. The U-Net-based generator embeds the watermark into the blue channel of the image, whereas the discriminator is optimized via both SSIM and cross-entropy loss. The experimental results show that the method achieves an approximately 24 dB PSNR, an approximately 0.85 SSIM, and a bit error rate as low as 0.003% under common distortions. Although the technique significantly enhances

watermark invisibility and adapts well to the conditions of social network distribution, its reliance on standardized quantization tables restricts its flexible performance.

Zhu et al. [70] developed a combination of U-Net++ attention and dense skip connections to enhance the model's ability to localize features and embed messages precisely. The architecture also uses a quadratic nonlinear loss weight strategy inspired by the WGAN to optimize the balance between invisibility and robustness. The experimental results revealed advantages in terms of both visual quality and resilience to simulated and real-world attacks. However, using a complex architecture with attention mechanisms and dense layers may lead to a significant increase in the computational load.

A method to embed watermarks in audio was proposed in [71]. It relies on deep learning and resisting audio-recording (AR) distortions and embeds watermark bits in low-frequency spectrum frames to ensure imperceptibility. DeepAWR was tested and evaluated on the FMA and Aidatatang-200zh datasets. The model extracted a near-perfect watermark with an ACC of approximately 1.0 under AR distortions and at varying distances of 50–150 cm. However, this method suffers from the complexity of high decoding and a capacity limit of 100 bps.

Deep-based watermarking methods have made significant progress in enhancing watermark robustness and imperceptibility, utilizing advanced deep learning architectures such as GANs, CNNs, U-Nets, DNNs, and transformers. However, there remains an area for improvement in terms of computational efficiency, overfitting, and further robustness testing in real-world, complex environments. The complexity and resource demands may limit their applicability in real-time scenarios or on devices with restricted processing capabilities.

TABLE III. COMPARISON OF DIFFERENT DEEP LEARNING-BASED WATERMARKING TECHNIQUES

Reference	*Year*	*Technique*	*Description*	*Dataset*	*Advantage*	*Disadvantage*	*Results*
[56]	2018	HiDDeN	End-to-end deep learning framework for data hiding; uses CNNs for encoding and decoding watermark.	COCO (training), BOSS (testing)	High capacity and robustness; can withstand distortions like noise, cropping, and JPEG compression; adversarial loss enhances watermark invisibility	require high computational resources due to adversarial training	0.203 bits per pixel (BPP) data hiding with minimal distortion, demonstrating robustness against Gaussian blurring and JPEG compression
[57]	2019	Unsupervised GAN	Automates watermark embedding and extraction as an image fusion task; uses an embedder and extractor network.	ImageNet, COCO, CIFAR	Fully automated; resilient in real-world distortions, suitable for camera-captured images	Higher BER for severe noise levels; requires adaptation for specific distortion types.	PSNR: 39.72 dB; BER: 11.6% (JPEG Q10), 7.8% (cropping), 32.2% (Gaussian noise), 12.3% (salt-and-pepper)
[58]	2020	CNN	Blind and robust watermarking method; adjusts watermark resolution to fit host images.	BOSS dataset (10,000 images, 512 × 512 pixels, downscaled to 128 × 128 for training)	Blind, robust, and invisible watermarking; adaptable to any resolution; includes a scaling factor to balance invisibility and robustness; provides attack simulations.	Lower scaling factor values may reduce robustness	High invisibility with PSNR of 43.23 dB during training and 40.58 dB during evaluation; BERs lower than 10% under attacks like Gaussian filtering, median filtering, and salt-and-pepper noise
[59]	2020	ReDMark	Combining two fully convolutional networks for embedding and	CIFAR10, Pascal VOC2012 (training),	High resistance to attacks (JPEG compression, Gaussian noise,	Limited generalization information may require high	PSNR > 35 dB, SSIM ≈ 0.96, robust against real-

			extraction improves speed and robustness.	Granada (testing)	cropping, grid-based removals)	computational power for training	world distortions
[60]	2022	Supervised GAN	Embedding invisible watermarks into images for ownership verification; uses a pretrained encoder-decoder structure.	CelebA, LSUN-Bedroom, FFHQ, VGG Flowers	Efficient ownership verification with minimal computational cost; robust to postprocessing distortions; protects pretrained GANs	Potentially limited to GAN-generated images; requires specific training setup.	Near-perfect watermark accuracy (~99%); PSNR > 45, SSIM > 99%
[61]	2022	GAN-3D CNN	Invisibly embed encrypted watermarks in video frames to enhance security and ownership verification.	UCF101 (training), Hollywood2 (validation), A2D, TikTok trending videos (testing)	High accuracy in deepfake prevention; invisible watermarking; adaptable to various video platforms	PSNR and SSIM scores indicate some level of noise and lower visual quality in certain cases	99.7% embedding accuracy, 100% deepfake prevention success, PSNR: 13.98, SSIM: 0.301
[62]	2023	ARWGAN	An attention-guided robust watermarking model that uses a GAN framework with feature fusion	COCO (training), Pascal VOC2012, ImageNet, EUVP (testing)	High robustness; minimal visual impact; adaptable to various real-world attacks	Computationally intensive due to attention mechanisms and feature fusion	COCO: PSNR 35.87 dB, SSIM 0.9688; Pascal VOC2012: PSNR 36.83 dB, SSIM 0.9698; Robustness: 98.39% on COCO, 98.89% on Pascal VOC2012, 94.45% under JPEG compression, 100% against Gaussian noise
[63]	2023	Hybrid CNN-Transformer	Combining CNNs and Swin Transformers for enhanced feature extraction,	COCO dataset (10,000 images, resized to 128x128, split into training, validation, and testing sets)	High imperceptibility and robustness	Potentially high computational cost due to hybrid model complexity	Achieved PSNR > 40 dB, SSIM > 0.95; Low BER; robust against Gaussian, Salt and Pepper, Dropout, and JPEG compression
[64]	2023	DNN with CWC & NFTs	Embedding watermark into convolution layers (VGG16, ResNet50) using Constant Weight Codes (CWC).	Flower dataset, ImageNet	99% pruning resilience; Maintains 97% model accuracy; NFT integration enhances security	Limited to convolution-layer architectures; Untested generalization across diverse models/datasets	99% robustness against pruning attacks; 97% accuracy on flower dataset; Statistical MSE detection
[65]	2023	Weightless White-box Watermarking	Embedding watermarks via code matching between watermark bits	CIFAR-10	No training overhead; Undetectable via weight analysis; Robust to	Vulnerable to LSB-targeted attacks; Untested on nonconvolutional	BER <20% under Gaussian noise (std ≤2); 0% accuracy

			and LSBs of weight binaries without modifying weights.		pruning/fine-tuning.	models (e.g., transformers).	loss; Survives pruning; Fails at std=5 (46% BER, 52.8% accuracy drop).
[66]	2024	Quantized Activation Function	The proposed neural network employs a quantized activation function (QAF) using hyperbolic tangent curves to approximate JPEG compression's quantization table, integrated into a GAN framework, enhancing robustness against JPEG attacks and reducing watermark extraction errors compared to noise-based simulations.	CIFAR-10 (50,000 images, 32x32 pixels); 49 grayscale images (512x512 pixels) with embedded 4x4 bit watermarks	QAF layer simulates JPEG compression more accurately, improving robustness and preserving image quality; outperforms standard noise addition layers	Complex architecture due to integration of QAF layer	Higher SSIM and PSNR scores indicate better image quality and less visual degradation; Lower BER, particularly at JPEG compression levels Q ≤ 70,
[67]	2024	GANMarked	A secure watermarking technique that utilizes a secure trigger key for model ownership verification,	DIV2K, COCO, COVID-19, Cats & Dogs	High security with a key-based autoencoder; handles diverse robustness with cross-domain testing; suitable for IP protection.	Computational cost due to GAN	High PSNR and SSIM; NC of 0.999; BER of 0, showing robustness and resilience to hybrid attacks
[68]	2024	U-Net GAN + Siamese One-Shot	Combining a modified U-Net-based conditional GAN (Npix2Cpix) for denoising historical watermarks with Siamese networks for one-shot classification.	Large-scale historical watermark dataset (>16,000 images)	95% one-shot classification accuracy; Robust to handwriting/noise; Preserves watermark spatial consistency	Computationally intensive; Struggles with severe degradation scenarios	PSNR: 32.4 dB; SSIM: 0.89; 95% classification accuracy; Outperforms SOTA by 15-20% accuracy
[69]	2024	Template-based NN extraction + Polar Codes	Combining neural network-based watermark extraction with polar codes for error correction. Employing CA-SCL decoding with CRC for robustness.	Custom dataset (1,000 HD images from movie trailers; tested on 7 platforms: Facebook, VK, Telegram, etc.)	High robustness to real-world social network distortions (BER ≤ 0.01); adaptable to multiple platforms via nested polar codes; no auxiliary data needed.	Computational overhead from polar code decoding (SCL list size = 8); dependency on prefiltering (JPEG QF=90) to exclude unstable containers.	PSNR=39.66, SSIM=0.987; watermark capacity 650-1600 bits for 1920×1080 images; BER < 3% on Telegram/Snapchat, <20% on Pinterest after transmission.
[70]	2024	Attention U-Net++	A digital image watermarking scheme based on attention U-Net++ structure	MIRFLICKR	Superior visual quality and robustness effectively extract image features and find optimal	Better performance requires careful tuning of the loss function weights; it relies on	Watermarked images show higher PSNR (up to 37.39 dB) and SSIM (up to

					pixel space for embedding messages.	simulated noise layers during training.	0.982) values; maintain bit-accuracy above 98% for various attacks including perspective warp, Gaussian noise, and JPEG compression; outperform HiDDeN, StegaStamp, and RIHOOP in both visual quality and robustness metrics.
[71]	2025	Deep Learning + ReDS Simulator	Embedding bits in low-frequency spectrogram frames; uses ReDS for AR distortion simulation.	FMA, Aidatatang-200zh	Robust to AR (ACC ≈1.0); high SNR (34–36 dB).	High decoder complexity; limited to 100 bps.	ACC >0.98 under AR; bACC ≈1.0 (50 cm), 0.9997 (150 cm); SNR 34–36 dB.

7.4 Comparison of Watermarking Techniques

Traditional and deep learning-based watermarking methods differ significantly in methodology, robustness, imperceptibility, and computational complexity. Table IV provides a comparative analysis based on key characteristics. Traditional and deep learning-based watermarking techniques have emerged as crucial tools for protecting digital content, each offering distinct advantages and challenges. While traditional methods remain suitable for stable environments requiring fast performance with limited capacity to withstand advanced attacks, deep learning approaches represent a significant leap in adapting to evolving threats. Owing to their ability to resist attacks, offer superior imperceptibility, and handle diverse datasets efficiently, these methods signify the future of digital content protection. Although deep learning demands more computational resources, continuous advancements in acceleration techniques and performance optimization are making it the ideal solution for advanced applications in the long run. This research suggests a hybrid approach that combines the reliability of traditional embedding with the adaptive power of deep learning to address complex threats and provides a balanced strategy, enhancing digital security and safeguarding intellectual property rights.

TABLE IV. COMPARISON BETWEEN TRADITIONAL AND DEEP-LEARNING-BASED WATERMARKING

Characteristic	***Traditional watermarking method***	***Deep-learning based watermarking***
Method	DCT, DWT, SVD, Fractal Coding, and LSB encoding	CNNs, GANs, Transformers
Data Types Used	Primarily images, some audio, and 3D models	Images, videos, and possibly other media types
Robustness	Good against basic attacks (e.g., noise, compression). Vulnerable to specific attacks	High robustness, capable of coping with complex and adaptive attacks
Imperceptibility	Often high, but can be compromised with larger watermarks	Highly imperceptible due to adversarial training techniques
Computational Complexity	Generally, lower, suitable for real-time applications	Higher computational costs may not be ideal for real-time usage
Performance Metrics	PSNR is typically above 40 dB; SSIM scores depend on the method	High PSNR (e.g., >43 dB) and varied BER performance depending on the model used
Scalability	May struggle with large datasets and diverse applications	More adaptable to varying data sizes and types
Application	Suitable for stable datasets with predictable attacks	High adaptability in dynamic, varied real-world conditions

Adaptability	Limited adaptability; often tailored for specific types of media	Strong adaptability across various media and conditions
Ease of Implementation	Simpler to implement due to established techniques	Requires expertise in deep learning frameworks and models

8. CONCLUSIONS

This paper examines several state-of-the-art advancements in digital watermarking from 2016--2025. It presents traditional watermarking methods in the spatial and frequency domains, which shows their simplicity and efficiency. However, they face challenges in robustness and adaptability, making them vulnerable to advanced threats and manipulations. Therefore, watermarking techniques have evolved to develop robust watermarking methods that keep up with the evolution of watermarking technologies. One of the efficient techniques for watermarking is the use of deep learning; however, overcoming challenges related to high resource consumption, ensuring model efficiency, and developing reliable detection mechanisms are necessary. These challenges require the development of new solutions that provide intellectual property protection while maintaining the effectiveness and performance of deep learning models. The future of watermarking focuses on adopting hybrid models that combine the power of deep learning, adaptive techniques, and dynamic schemes to improve imperceptibility, robustness, and security and address the continuing diversity of challenges posed by unauthorized data manipulation and access. Furthermore, establishing standardized datasets is necessary for benchmarking performance and fostering meaningful comparisons across different watermarking techniques. Collaboration across academia and industry will be crucial to driving these advancements and addressing the ongoing challenges in digital watermarking.

Conflicts of interest

The authors declare that they have no conflicts of interest.

Funding

No funding.

Acknowledgement

None.

References

[1] P. Aberna and L. Agilandeeswari, “Digital image and video watermarking: Methodologies, attacks, applications, and future directions,” Multimedia Tools and Applications, vol. 83, pp. 5531–5591, June 2023.

[2] Y. Luo, X. Tan, and Z. Cai, 'Robust deep image watermarking...' CMC-Comput. Mater. Contin., vol. 81, no. 1, 2024, doi: 10.32604/cmc.2024.055150

[3] M. Begum and M. S. Uddin, “Digital image watermarking techniques: A review,” Journal of Imaging, vol. 6, no. 2, Art. no. 24, 2020. https://doi.org/10.3390/info11020110.

[4] A. Malanowska, W. Mazurczyk, T. Koohpayeh Araghi, D. Megías, and M. Kuribayashi, "Digital watermarking—A meta-survey and techniques for fake news detection," IEEE Access, vol. 12, pp. 36311–36345, 2024, doi: 10.1109/ACCESS.2024.3374201.

[5] M. Gupta and R. R. Kishore, “A survey of watermarking technique using deep neural network architecture,” in 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), 2021, pp. 630–634.

[6] Y. A. Hassan and A. M. S. Rahma, (2023). An overview of robust video watermarking techniques. Iraqi Journal of Science, vol. 64, no. 7, 3613-3624. https://doi.org/10.24996/ijs.2023.64.7.38

[7] O. Evsutin, A. Melman, and R. Meshcheryakov, “Digital steganography and watermarking for digital images: A review of current research directions,” IEEE Access, vol. 8, pp. 166575–166611, 2020. https://doi.org/10.1109/ACCESS.2020.3022779.

[8] R. Thomas and M. Sucharitha, “Reversible color image watermarking scheme based on contourlet transform and principal component analysis,” in 2021 7th International Conference on Electrical Energy Systems (ICEES), 2021, pp. 1–6. IEEE. https://doi.org/10.1109/ICEES51510.2021.9383746.

[9] X. Zhong, A. Das, F. Alrasheedi, and A. Tanvir, “A brief, in-depth survey of deep learning-based image watermarking,” Applied Sciences, vol. 13, no. 21, Art. no. 11852, 2023. https://doi.org/10.3390/app132111852.

[10] D. Awasthi, A. Tiwari, P. Khare, and V. K. Srivastava, “A comprehensive review on optimization-based image watermarking techniques for copyright protection,” Expert Systems with Applications, vol. 242, Art. no. 122830, 2024. https://doi.org/10.1016/j.eswa.2023.122830.

[11] S. Sharma, J. J. Zou, G. Fang, P. Shukla, and W. Cai, “A review of image watermarking for identity protection and verification,” Multimedia. Tools Applications vol. 83, no. 11, pp. 31829-31891, 2024.

[12] S. Khalilidan, M. Mahdavi, A. Balouchestani, Z. Moti, and Y. Hallaj, “A semi-blind watermarking method for authentication of face images using autoencoders,” in 2020 6th International Conference on Web Research (ICWR), 2020, pp. 123–128. IEEE. https://doi.org/10.1109/ICWR49608.2020.9122291.

[13] R. M. G. Ferrari and A. M. H. Teixeira, "A switching multiplicative watermarking scheme for detection of stealthy cyber-attacks," IEEE Transactions on Automatic Control, vol. 66, no. 6, pp. 2558–2573, 2021. https://doi.org/10.1109/TAC.2020.3013850.

[14] S. Wadhera, D. Kamra, A. Rajpal, A. Jain, and V. Jain, "A comprehensive review on digital image watermarking," WCNC-2021: Workshop on Computer Networks & Communications, 2021.

[15] L. Geng, W. Zhang, H. Chen, H. Fang, and N. Yu, "Real-time attacks on robust watermarking tools in the wild by CNN," Journal of Real-Time Image Processing, 2020. https://doi.org/10.1007/s11554-020-00941-8.

[16] S. Boujerfaoui, R. Riad, H. Douzi, F. Ros, and R. Harba, "Image watermarking between conventional and learning-based techniques: A literature review," Electronics, vol. 12, no. 1, Art. no. 74, 2023. https://doi.org/10.3390/electronics12010074.

[17] H. M. Al-Dabbas, R. A. Azeez, and A. E. Ali, "High-accuracy models for iris recognition with merging features," Int. J. Adv. Appl. Sci., vol. 11, no. 6, pp. 89–96, 2024, doi: 10.21833/ijaas.2024.06.010.

[18] H. Tao, L. Chongmin, J. M. Zain, and A. N. Abdalla, "Robust image watermarking theories and techniques: A review," International Journal of Advanced Research in Computer and Communication Engineering, vol. 12, no. 2, 2014.

[19] M. A. Nematollahi, C. Vorakulpipat, and H. Gamboa Rosales, Digital watermarking techniques and trends. Springer, 2017. https://doi.org/10.1007/978-981-10-2095-7.

[20] G. Al-Kateb, I. Khaleel, and M. Aljanabi, "CryptoGenSec: A hybrid generative AI algorithm for dynamic cryptographic cyber defence," *Mesopotamian Journal of Cybersecurity*, vol. 4, no. 3, pp. 22–35, 2024. https://doi.org/10.58496/MJCS/2024/013.

[21] Common Objects in Context (COCO) Dataset. Available: https://cocodataset.org/. Accessed: Dec. 30, 2024.

[22] RAISE Dataset. Available: http://loki.disi.unitn.it/RAISE. Accessed: Dec. 30, 2024.

[23] COVERAGE Dataset. Available: https://www.coverage-db.org/. Accessed: Dec. 30, 2024.

[24] USC-SIPI Image Database. Available: https://sipi.usc.edu/database/. Accessed: Dec. 31, 2024.

[25] Modified CASIA Dataset. IEEE Dataport. Available: https://ieee-dataport.org/open-access/modified-casia. Accessed: Dec. 31, 2024. Dataset 28xhc4kyfp.

[26] CG-1050 (v1). Available: https://data.mendeley.com/datasets/28xhc4kyfp/1. Accessed: Dec. 31, 2024.

[27] CoMoFoD: Copy-Move Forgery Detection Dataset. Visual Computing Lab. Available: https://www.vcl.fer.hr/comofod/. Accessed: Jan. 1, 2025.

[28] Global Roads Inventory Project (GRIP) Dataset. Netherlands Environmental Assessment Agency, GLOBIO. Available: https://www.globio.info/download-grip-dataset. Accessed: Jan. 2, 2025.

[29] T. Wang, M. Huang, H. Cheng, B. Ma, and Y. Wang, "Robust identity perceptual watermark against deepfake face swapping," arXiv, 2023]. Available: https://doi.org/10.48550/arXiv.2311.01357. Accessed: Jan. 3, 2025.

[30] CIFAR Dataset. GitHub. Available: https://github.com/EN10/CIFAR. Accessed: Jan. 3, 2025.

[31] C. Mo, F. Liu, M. Zhu, G. Yan, B. Qi, and C. Yang, "Image steganalysis based on deep content features clustering," Computers, Materials & Continua, vol. 73, no. 3, pp. 5641–5659, 2023. Available: https://doi.org/10.32604/cmc.2023.039540. Accessed: Jan. 3, 2025.

[32] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for deepfake forensics," arXiv, 2020. Available: https://doi.org/10.48550/arXiv.1909.12962. Accessed: Dec. 30, 2024.

[33] Z. F. Hussain and H. R. Ibraheem, "Novel convolutional neural networks based Jaya algorithm approach for accurate deepfake video detection," Mesopotamian Journal of Cybersecurity, vol. 2023, pp. 35–39, 2023. https://doi.org/10.58496/MJCS/2023/007.

[34] Kinetics-700 Dataset. Papers with Code. Available: https://paperswithcode.com/dataset/kinetics-700. Accessed: Dec. 31, 2024.

[35] Google. (2016). YouTube-8M: A large-scale video classification benchmark. Available: https://research.google.com/youtube8m/. Accessed: Jan. 2, 2025.

[36] Something-Something V2 Dataset. Papers with Code. Available: https://paperswithcode.com/dataset/something-something-v2. Accessed: Jan. 1, 2025.

[37] EPIC-KITCHENS-100 Dataset. University of Bristol. Available: https://epic-kitchens.github.io/. Accessed: Mar. 13, 2025.

[38] UCF101 Action Recognition Dataset. Center for Research in Computer Vision, University of Central Florida. Available: https://www.crcv.ucf.edu/data/UCF101.php. Accessed: Dec. 29, 2024.

[39] FaceForensics++. Papers with Code. Available: https://paperswithcode.com/dataset/faceforensics-1. Accessed: Dec. 28, 2024.

[40] R. S. R. Channapragada and M. V. N. K. Prasad, "Digital watermarking using fractal coding," in Advances in Signal Processing and Intelligent Recognition Systems, S. M. Thampi, A. Abraham, and B. J. Prabhu, Eds., Springer, 2016, pp. 109–118. https://doi.org/10.1007/978-3-319-28658-7_10.

[41] S. Borah and B. Borah, "Quantization index modulation (QIM) based watermarking techniques for 3D meshes," in 2017 Fourth International Conference on Image Information Processing (ICIIP), 2017, pp. 284–289. IEEE. https://doi.org/10.1109/ICIIP.2017.8313726.

[42] Y. Chincholkar and S. Ganorkar, "Audio watermarking algorithm implementation using patchwork technique," in 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), 2019. https://doi.org/10.1109/I2CT45611.2019.9033771.

[43] N. H. Barnouti, Z. S. Sabri, and K. L. Hameed, “Digital watermarking based on DWT (discrete wavelet transform) and DCT (discrete cosine transform),” International Journal of Engineering & Technology, vol. 7, no. 4, pp. 4825–4829, 2019. https://doi.org/10.14419/ijet.v7i4.25085.

[44] M. R. Islam, M. A. Hasan, S. M. T. Islam, N. Islam, and M. L. Hossain, “Analysis and implementation of digital watermarking using spread spectrum method,” in 2020 International Conference on System, Computation, Automation and Networking (ICSCAN), Pondicherry, India, 2020. https://doi.org/10.1109/ICSCAN49426.2020.9262285.

[45] S. Han and W. L. Zhang, “SVD-based digital watermarking algorithm for print-and-scan process,” in 2020 IEEE 20th International Conference on Communication Technology (ICCT), 2020, pp. 1427–1430. IEEE. https://doi.org/10.1109/ICCT50939.2020.9295673.

[46] M. Cedillo-Hernandez, A. Cedillo-Hernandez, and F. J. Garcia-Ugalde, “Improving DFT-based image watermarking using particle swarm optimization algorithm,” Mathematics, vol. 9, no. 15, Art. No. 1795, 2021. https://doi.org/10.3390/math9151795.

[47] J. H. Awad and B. D. Majeed, "Image watermarking based on IWT and parity bit checking," Iraqi J. Sci., vol. 62, no. 8, pp. 2726–2739, 2021, doi: 10.24996/ijs.2021.62.8.26.

[48] W. Alomoush, O. A. Khashan, A. Alrosan, H. H. Attar, A. Almomani, F. Alhosban, and S. N. Makhadmeh, “Digital image watermarking using discrete cosine transformation based linear modulation,” Journal of Cloud Computing, vol. 12, Art. no. 96, 2023. https://doi.org/10.1186/s13677-023-00413-1.

[49] Z. B. Faheem, A. Ishaq, F. Rustam, I. de la Torre Díez, D. Gavilanes, M. Masias Vergara, and I. Ashraf, “Image watermarking using least significant bit and Canny edge detection,” Sensors, vol. 23, no. 3, Art. no. 1210, 2023. https://doi.org/10.3390/s23031210.

[50] T. Wang, M. Huang, H. Cheng, B. Ma, and Y. Wang, "Robust identity perceptual watermark against deepfake face swapping," arXiv preprint arXiv:2311.01357, 2023. doi: 10.48550/arXiv.2311.01357.

[51] Y. A. Hassan and A. M. S. Rahma, “Improving video watermarking through Galois field GF(2^4) multiplication tables with diverse irreducible polynomials and adaptive techniques,” Computers, Materials & Continua, vol. 78, no. 2, Art. no. 046149, 2024. https://doi.org/10.32604/cmc.2023.046149.

[52] A. Marjuni, G. F. Shidik, A. Syukur, D. R. I. M. Setiadi, and N. Rijati, “Toward achieving a trade-off for SVD-based image watermarking using Mojette transform,” IEEE Access, 2024. https://doi.org/10.1109/ACCESS.2024.3368528.

[53] Y. Qiu, S. Jiao, and Q. Su, “Enhancing color image watermarking via fast quaternion Schur decomposition: a high-quality blind approach,” The Visual Computer, 2024. https://doi.org/10.1007/s00371-024-03674-y.

[54] B. D. Majeed, A. H. Taherinia, H. S. Yazdi, and A. Harati, “CSRWA: Covert and severe attacks resistant watermarking algorithm,” Computers, Materials & Continua, 2025. https://doi.org/10.32604/cmc.2024.059789

[55] S. Wu, W. Lu, X. Yin, and R. Yang, “Robust watermarking against arbitrary scaling and cropping attacks,” Signal Processing, 2025. https://doi.org/10.1016/j.sigpro.2024.109655.

[56] J. Zhu, R. Kaplan, J. Johnson, and L. Fei-Fei, “HiDDeN: Hiding data with deep networks,” in Computer Vision – ECCV 2018, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., Springer, Cham, Lecture Notes in Computer Science, vol. 11219, pp. 682–697, 2018. https://doi.org/10.1007/978-3-030-01267-0_40.

[57] X. Zhong and F. Y. Shih, “A robust image watermarking system based on deep neural networks,” arXiv preprint arXiv:1908.11331, 2019. https://doi.org/10.48550/arXiv.1908.11331.

[58] J.-E. Lee, Y.-H. Seo, and D.-W. Kim, “Convolutional neural network-based digital image watermarking adaptive to the resolution of image and watermark,” Applied Sciences, vol. 10, no. 19, Art. no. 6854, 2020. https://doi.org/10.3390/app10196854.

[59] M. Ahmadi, A. Norouzi, N. Karimi, S. Samavi, and A. Emami, “ReDMark: Framework for residual diffusion watermarking based on deep networks,” Expert Systems with Applications, vol. 146, Art. no. 113157, 2020. https://doi.org/10.1016/j.eswa.2019.113157.

[60] J. Fei, Z. Xia, B. Tondi, and M. Barni, “Supervised GAN watermarking for intellectual property protection,” in 2022 IEEE International Workshop on Information Forensics and Security (WIFS), Shanghai, China, Dec. 12–16, 2022. IEEE. https://doi.org/10.1109/WIFS55849.2022.9975409.

[61] I. Noreen, M. S. Muneer, and S. Gillani, "Deepfake attack prevention using steganography GANs," PeerJ Comput. Sci., vol. 8, Art. no. e1125, 2022, doi: 10.7717/peerj-cs.1125.

[62] J. Huang, T. Luo, L. Li, G. Yang, H. Xu, and C.-C. Chang, “ARWGAN: Attention-guided robust image watermarking model based on GAN,” IEEE Transactions on Instrumentation and Measurement, vol. 72, Art. no. 5018417, 2023. https://doi.org/10.1109/TIM.2023.3285981.

[63] B. Wang, Z. Song, and Y. Wu, “Robust blind watermarking framework for hybrid networks combining CNN and Transformer,” in *Proceedings of Machine Learning Research*, vol. 222, ACML 2023, B. Yanıkoğlu and W. Buntine, Eds., 2023. https://proceedings.mlr.press/v222/wang24a/wang24a.pdf.

[64] M. Kuribayashi, T. Yasui, and A. Malik, “White box watermarking for convolution layers in fine-tuning model using the constant weight code,” Journal of Imaging, vol. 9, no. 6, p. 117, 2023. https://doi.org/10.3390/jimaging9060117.

[65] H. Deng, X. Wang, G. Yu, X. Dang, and R. P. Liu, “A novel weights-less watermark embedding method for neural network models,” IEEE 22nd International Symposium on Communications and Information Technologies (ISCIT), 2023. https://doi.org/10.1109/ISCIT57293.2023.10376108.

[66] S. Yamauchi and M. Kawamura, “A neural-network-based watermarking method approximating JPEG quantization,” *Journal of Imaging*, vol. 10, no. 6, Art. no. 138, 2024. https://doi.org/10.3390/jimaging10060138.

[67] H. K. Singh, N. Baranwal, K. N. Singh, and A. K. Singh, "GANMarked: Using secure GAN for information hiding in digital images," *IEEE Transactions on Consumer Electronics*, 2024. https://doi.org/10.1109/TCE.2024.3406956.

[68] U. Saha, S. Saha, S. A. Fattah, and M. Saquib, "Npix2Cpix: A GAN-based image-to-image translation network with retrieval-classification integration for watermark retrieval from historical document images," IEEE Access, 2024. https://doi.org/10.1109/ACCESS.2024.3424662.

[69] O. Evsutin, F. Ivanov, and K. Dzhanashia, "Watermarking for social network images with improved robustness through polar codes," IEEE Access, 2024. https://doi.org/10.1109/ACCESS.2024.3446489.

[70] L. Zhu, Y. Zhao, Y. Fang, and J. Wang, "A novel robust digital image watermarking scheme based on attention U-Net++ structure," The Visual Computer, vol. 40, pp. 8791–8807, 2024. https://doi.org/10.1007/s00371-024-03271-z

[71] G. Lin, W. Luo, P. Zheng, and J. Huang, "An audio watermarking method against re-recording distortions," Pattern Recognition, 2025. https://doi.org/10.1016/j.patcog.2025.111366.