INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 5, May 2026 | ISSN: 2456-4184 | IJNRD.ORG

A Systematic Review of Artificial Intelligence-Based Digital Image Watermarking Techniques: Challenges, Opportunities and Future Directions

Dr. Shalini Animeshbhai Mali
SASCMA English Medium Commerce College (B.Com & M.Com.) & Shri Hasmukhlal Hojiwala BBACollege & Smt. Ushaben Jayvadan Bodawala BCA & M.Sc. (IT) Integrated College & B.Sc. Data Science College

Abstract

Digital images are shared, copied, and re-used at a scale that was never seen before. Because of this, protecting ownership rights and checking whether an image has been tampered with have become genuine problems for researchers, content owners, and industries alike. Digital image watermarking is one of the oldest and most trusted answers to this problem, and in the last few years, Artificial Intelligence, particularly deep learning, has changed the way watermarking is done. Older watermarking methods relied on fixed mathematical rules to hide a mark inside an image, but these rules often failed when the image was compressed, cropped, or attacked by newer image-editing tools, including AI image generators themselves. This paper presents a systematic review of ten recent and well-cited studies on AI-based digital image watermarking, published mostly between 2020 and 2026. The review looks at the methods used, the strengths reported by the authors, and the gaps that are still left open. A comparative summary table is also provided to help readers see the ten studies side by side at a glance. Based on this reading, the paper discusses the major challenges that researchers are facing today, such as balancing robustness with image quality, defending against generative AI attacks, and the heavy computing cost of training deep models. It further points out the opportunities that this field offers, including watermarking for AI-generated content and medical image protection, and closes with possible future directions for researchers working in this area.

Keywords: digital image watermarking; artificial intelligence; deep learning; convolutional neural network; generative adversarial network; image security; copyright protection

1. Introduction

Every day, millions of images travel across social media, cloud storage, and messaging apps. Along with this convenience comes a real worry — how does one prove that an image belongs to a particular person or organisation, and how can one tell whether it has been changed from its original form? Digital watermarking tries to solve exactly this. In simple words, watermarking means hiding a small piece of information, called a watermark, inside a digital image in such a way that it does not disturb how the image looks to the human eye, but it can still be recovered later to check ownership or detect tampering.

For many years, watermarking was done using fixed signal-processing methods, such as working in the frequency domain of an image (for example DCT or DWT based methods) or directly changing pixel values in the spatial domain. These methods worked reasonably well, but they had a common weakness: once an attacker knew the underlying formula, it became easier to remove or damage the watermark. They also struggled when the image went through common real-world changes like compression, resizing, noise, or screenshot capture.

In recent years, Artificial Intelligence, and more specifically deep learning, has entered this field in a big way. Instead of using a fixed formula, a neural network is trained on thousands of images to learn, on its own, the best

way to hide and recover a watermark. Convolutional Neural Networks (CNNs), Generative Adversarial Networks (GANs), and, more recently, Transformer and diffusion-based models have all been tried by researchers for this purpose [6, 12]. These learning-based methods have shown much better resistance to attacks compared to the older fixed methods, because the network can be trained while simulating the very attacks it is expected to survive [13, 14].

At the same time, new problems have come up. AI image generation tools such as diffusion models can create highly realistic fake images, and the same technology can sometimes be used to remove or forge a watermark [11]. Training large watermarking networks also needs heavy computing resources, which is not always available to smaller research groups or organisations, especially in developing countries.

Research gap. Although several individual studies and a few surveys already exist on deep learning-based watermarking [5, 10, 12, 13], most of them either focus narrowly on one type of architecture (only CNN, or only GAN) or do not clearly separate the challenges coming from *traditional* attacks (cropping, compression, noise) from the newer challenge posed by *generative AI* attacks [6, 11]. There is a need for a compact, plainly written review that brings both strands together for readers who are new to the field, such as postgraduate students and early-career researchers.

Objectives of this study. Keeping this gap in mind, the present paper aims to:

1. Identify and review recent, credible studies (2020–2026) on AI-based digital image watermarking.
2. Compare the architectures, contributions, and limitations of these studies in a structured, easy-to-read format.
3. Bring out the common challenges that cut across these studies, with special attention to the threat posed by generative AI.
4. Suggest opportunities and concrete future research directions for scholars entering this field.

Contribution of this paper. This review contributes to the literature in three ways: (i) it provides a side-by-side comparative table of ten recent studies, which is often missing in narrative-only reviews; (ii) it separates challenges into technical, computational, and generative-AI-related categories for clarity; and (iii) it offers future directions that are specific and actionable rather than generic.

The rest of the paper is arranged as follows. Section 2 briefly explains the basic concepts and evaluation metrics used in watermarking research, for readers who are new to the area. Section 3 explains the method followed to select the papers for this review. Section 4 presents the review of the ten selected studies along with a comparative table. Section 5 discusses the major challenges seen across these studies. Section 6 talks about the opportunities this field opens up, and Section 7 lists possible future directions. Section 8 concludes the paper, followed by declarations and references.

2. Background Concepts

Before going into the review, it is useful to explain a few basic ideas in simple terms, since this paper is written for readers who may be new to watermarking research.

Imperceptibility refers to how invisible the watermark is to the human eye. A good watermark should not change the way the image looks. This is usually measured using the Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index (SSIM); higher values on both indicate that the watermarked image looks very close to the original.

Robustness refers to how well the watermark survives after the image undergoes common changes such as compression, cropping, resizing, noise addition, or filtering. Robustness is usually measured using the Normalized Correlation (NC) between the original and extracted watermark, or the Bit Error Rate (BER) of the recovered watermark bits.

Capacity refers to how much information can be hidden inside the image at one time. A higher capacity is generally preferred by content owners, but it usually comes at some cost to imperceptibility or robustness.

Security refers to how hard it is for an unauthorised party to detect, extract, or forge the watermark without knowing the secret key or the trained model used to embed it.

These four properties — imperceptibility, robustness, capacity, and security — pull against each other, and almost every study reviewed in this paper reports some trade-off between them.

---

3. Research Methodology

This review was carried out in a systematic manner, broadly following the usual steps of a literature review: searching, screening, and summarising.

Search strategy. Databases and search platforms such as Google Scholar, Scopus, Web of Science, ScienceDirect, SpringerLink, IEEE Xplore, and MDPI were used to search for papers. Search terms included combinations of "digital image watermarking," "deep learning," "artificial intelligence," "convolutional neural network watermarking," "GAN watermarking," and "AI-generated image watermarking."

Inclusion criteria. Papers were included if they (a) were published between January 2020 and May 2026, (b) were peer-reviewed journal articles, book chapters, or well-cited conference/preprint papers from reputed publishers such as MDPI, Springer, Elsevier, and IEEE, (c) directly dealt with AI-based or deep learning-based digital image watermarking, and (d) were available in English.

Exclusion criteria. Papers focusing only on text or audio watermarking, purely traditional (non-AI) watermarking with no learning component, and papers without sufficient methodological detail were left out.

Selection process. After an initial search, duplicate and clearly unrelated papers were removed. From the remaining pool, ten studies were finally chosen for detailed review on the basis of relevance, citation strength, and how well they represented the range of approaches used in this field — covering surveys as well as method-based studies, and covering CNN, GAN, and specialised architectures.

It must be honestly stated that this review, like most reviews done by a single author or a small team, cannot claim to be fully exhaustive of the very large and fast-growing literature in this area. It is meant to give a representative and fair picture rather than a complete count of every paper published.

Citation style used. In-text citations in this paper are given in square brackets, for example [6], with the number corresponding to the numbered entry in the References section at the end of the paper.

---

4. Review of Selected Studies

4.1 Review 1 — Bistroń, Żurada, and Piotrowski (2026) [6]

This paper gives a wide-ranging survey of both classical and deep learning-based image watermarking. The authors trace the history of the field, starting from the first use of CNNs in watermarking around 2017, moving to the arrival of GAN-based models, and finally to the recent use of Vision Transformers and diffusion models. A useful point made in this paper is that diffusion-based watermarking, because it works in the latent space of the image, is showing good early results against attacks coming from generative AI tools themselves. The paper is helpful for anyone wanting a quick but detailed timeline of how the field has grown.

4.2 Review 2 — Singh and Singh (2023, Journal of Electronic Imaging) [13]

Written by researchers from the National Institute of Technology, Patna, this review looks closely at watermarking done inside deep-learning environments. It explains basic ideas of traditional watermarking first, before moving to deep-learning-based methods, and then compares several recent contributions side by side. The authors are honest about the fact that most deep watermarking models still find it hard to balance three things together — keeping the image looking clean, keeping the watermark hidden, and making the watermark strong enough to survive attacks.

4.3 Review 3 — Ben Jabra and Ben Farah (2024) [5]

This paper reviews deep-learning-based watermarking for both images and videos, and compares them side by side. One of its interesting findings is that while image watermarking has moved quite far with deep learning, video watermarking is still behind, mostly because video brings extra problems like frame dropping and motion artefacts that image-only methods are not built to handle. The paper is useful for readers who want to understand where video watermarking research still needs catching up.

4.4 Review 4 — Li, Wang, and Barni (2021) [12]

This is one of the earlier and widely cited surveys covering deep neural network watermarking broadly. The authors classify the field along three lines: the type of neural network used, the structure of the watermarking model, and the training strategy followed. They also provide a useful statistical breakdown of the datasets, loss functions, and types of attacks that different studies have used, which is quite handy for a new researcher trying to get a lay of the land quickly.

4.5 Review 5 — Hosny, Magdi, El-Komy, and Hamza (2024) [10]

Published in Computer Science Review, this survey takes stock of digital image watermarking methods built using deep learning, with a good coverage of medical image watermarking as well. It brings together studies on encrypted-image watermarking, CNN and GAN-based schemes, and points out that, despite so much work already done, very few studies test their models against a genuinely wide variety of attacks together, which limits how much we can trust the reported robustness numbers.

4.6 Review 6 — Singh and Singh (2023, Multimedia Tools and Applications) [14]

Different from their earlier review, this paper by the same authors proposes an actual CNN-based watermarking method. An encoder network pulls out features from the cover image and the secret watermark, combines them, and a denoising autoencoder is used on the receiving side to clean up noise before the watermark is extracted. Their results show good visual quality along with reasonable robustness, and this paper is a good example of how review-type researchers also contribute original technical work in the same space.

4.7 Review 7 — Anand and Singh (2022) [3]

This study looks at covert communication using deep learning, an area closely related to watermarking. It studies how deep networks can be used to hide information inside cover media in a way that is very hard to detect, and discusses the trade-off between how much data can be hidden and how well it survives compression and other common changes. The paper is useful for understanding the connection between watermarking and the broader field of information hiding.

4.8 Review 8 — Anand, Singh, and Zhou (2023) [4]

Focused specifically on medical images, this book chapter reviews watermarking techniques meant to protect patient data and medical records. Medical images bring their own special demands — the watermark must not disturb the diagnostic quality of the image even slightly, since a doctor's judgement depends on it. The authors

discuss zero-watermarking and reversible watermarking approaches that try to meet this strict requirement, and they call for more deep-learning-based solutions designed specifically for the healthcare setting.

4.9 Review 9 — Anand, Bedi, Aggarwal, Khan, and Rida (2024) [2]

This paper presents a deep learning-based zero-watermarking approach built specially for authenticating and securing healthcare records. Zero-watermarking is a slightly different idea — instead of changing the pixels of the image at all, a "fingerprint" is generated from the image's own features and stored separately, so that the original medical image remains completely untouched. The study reports strong results for tamper detection while keeping the image completely lossless, which is important in clinical settings.

4.10 Review 10 — Gao, Kang, and Chen (2021) [8]

This paper works on video rather than still images but is included here because it uses a similar deep CNN and self-organizing map approach that has since influenced image watermarking research as well. The authors build a zero-watermarking scheme working in the polar complex exponential transform domain and report good robustness against common video-processing attacks. It is a good example of how ideas first tested on video are often later carried back into still-image watermarking research.

4.11 Comparative Summary Table

Table 1 places the ten reviewed studies side by side so that a reader can compare their focus area, the type of AI architecture used, the main contribution, and the limitation noted during this review.

Table 1. Comparative summary of the ten reviewed studies on AI-based digital image watermarking

S. No.	Study (Author, Year)	Focus Area	AI Architecture Used	Main Contribution	Limitation Noted
1	Bistroń et al. (2026) [6]	General image watermarking survey	CNN, GAN, Transformer, Diffusion	Traces the full timeline of AI watermarking and highlights diffusion-based latent-space embedding	Being a survey, it does not test any single method experimentally
2	Singh & Singh (2023, JEI) [13]	Deep-learning watermarking survey	CNN, GAN	Compares many recent contributions and lays out basic concepts clearly	Notes but does not solve the imperceptibility–robustness–capacity trade-off
3	Ben Jabra & Ben Farah (2024) [5]	Image and video watermarking	CNN, GAN	Shows that video watermarking lags behind image watermarking in deep-learning adoption	Video-specific attacks (frame drop, motion) remain under-studied
4	Li, Wang, & Barni (2021) [12]	Deep neural network watermarking survey	CNN, GAN, DNN (general)	Statistical breakdown of datasets, loss functions, and	Being an earlier survey, it does not cover Transformer

S. No.	Study (Author, Year)	Focus Area	AI Architecture Used	Main Contribution	Limitation Noted
				attack types used across the literature	or diffusion-based work
5	Hosny et al. (2024) [10]	Deep-learning watermarking survey incl. medical images	CNN, GAN	Wide coverage including encrypted-image and medical watermarking	Points out that few studies test against a genuinely wide variety of combined attacks
6	Singh & Singh (2023, MTAP) [14]	Method-based (not a survey)	CNN + denoising autoencoder	Proposes an encoder–decoder CNN scheme with good visual quality and robustness	Tested mainly on standard datasets, not on real-world social-media compression
7	Anand & Singh (2022) [3]	Covert communication (related field)	Deep learning (general)	Studies the capacity–robustness trade-off in information hiding	Focused on covert communication rather than watermarking specifically
8	Anand, Singh, & Zhou (2023) [4]	Medical image watermarking	Zero-watermarking, reversible watermarking	Reviews methods that protect diagnostic image quality	Calls for more deep-learning solutions built specifically for healthcare
9	Anand, Bedi, Aggarwal, Khan, & Rida (2024) [2]	Healthcare record watermarking	Deep learning-based zero-watermarking	Keeps the medical image completely untouched while still enabling tamper detection	Zero-watermarking needs a separate secure database to store the fingerprint
10	Gao, Kang, & Chen (2021) [8]	Video zero-watermarking	Deep CNN + self-organizing map	Good robustness against common video-processing attacks	Built for video; not directly a still-image method

4.12 Comparative Discussion

Reading Table 1 as a whole, three broad observations stand out. First, CNN-based encoder–decoder designs (rows 2, 4, 5, 6) remain the backbone of most practical watermarking systems, while newer Transformer and diffusion-based ideas (row 1) are only beginning to be adopted widely. Second, zero-watermarking (rows 8, 9, 10) is clearly becoming the preferred approach whenever the cover image cannot be altered even slightly, such as in medical or forensic use, because it stores a separately generated fingerprint rather than modifying pixel data. Third, almost none of the ten studies report testing their method against attacks generated by modern diffusion-based image

editing tools, which suggests that this specific gap, discussed further in Section 5, is still wide open across the literature and not just a limitation of one or two papers.

---

5. Challenges in AI-Based Image Watermarking

Balancing image quality and robustness. The single biggest challenge that keeps appearing across the literature is what researchers call the invisibility-robustness trade-off [12, 13, 14]. If the watermark is hidden very deeply and made hard to detect, it often becomes weak and gets destroyed easily by attacks. If it is made stronger to resist attacks, it starts to slightly change how the image looks, which defeats the whole purpose of watermarking.

Generative AI as a new kind of attacker. Diffusion-based image generation and editing tools, which have become extremely common since 2023, are now able to regenerate or "clean up" images in a way that can wipe out an invisible watermark without the attacker even trying to target it specifically [6, 11]. This is a genuinely new and difficult problem that older watermarking research never had to deal with.

Heavy computational cost. Training deep watermarking networks, especially GAN and diffusion-based ones, needs a lot of GPU time and large datasets [5, 6]. This puts smaller labs, universities in developing regions, and independent researchers at a real disadvantage compared to big technology companies.

Lack of common testing standards. Different papers test their models on different datasets and against different sets of attacks, which makes it very hard to fairly compare one method against another [10]. A method that looks excellent in one paper's test conditions may not perform the same way under another paper's conditions.

Security of the watermarking model itself. Since the "rules" of hiding and extracting the watermark are now learned inside a neural network rather than a fixed formula, there is a new risk — if an attacker gets access to the trained model itself, they may find it easier to remove watermarks made by that same model [12].

Domain-specific needs. Medical, legal, and satellite images each have their own strict quality requirements, and a general-purpose watermarking model built for ordinary photographs does not always transfer well to these specialised domains [2, 4, 8].

---

6. Opportunities

Despite the challenges, this field is opening up several promising directions of opportunity. AI-generated content is exploding in volume, and there is now a genuine market and social need for reliable ways to mark and trace such content, which watermarking research is well placed to serve [11]. Healthcare is another strong opportunity area, since hospitals and diagnostic centres need trustworthy, tamper-proof, and lossless ways to protect patient image records, and zero-watermarking research is already showing useful early results here [2, 4]. There is also room for combining watermarking with blockchain-based record-keeping, so that ownership proof does not depend on a single central authority. Lightweight and efficient deep models, designed to run on mobile phones or low-power devices, could also make watermarking accessible to smaller businesses and individual creators, not just large companies [7]. Finally, cross-domain approaches that borrow ideas from steganography, digital forensics, and deepfake detection could push watermarking research forward at a faster pace than working in isolation [3].

---

7. Future Research Directions

Building on the gaps found in this review, a few directions look particularly worth pursuing for future researchers. Watermarking schemes need to be tested and built specifically to survive attacks from generative AI tools, not just traditional attacks like cropping or compression [6, 11]. Researchers could also work towards common, shared benchmark datasets and a standard set of attack tests, so that different studies can be compared fairly with each other [10, 12]. There is a need for lighter and less resource-hungry model designs that do not compromise too much on robustness, making the technology usable for smaller organisations as well. More attention should also go towards domain-specific watermarking, particularly for medical, legal, and satellite images, where the acceptable margin of error is very small [2, 4, 8]. Lastly, combining watermarking with explainable AI methods could help build trust, since users and courts of law would be able to understand, in simple terms, why a particular image was flagged as watermarked or tampered with.

---

8. Conclusion

This paper carried out a systematic review of ten recent studies on Artificial Intelligence-based digital image watermarking, published largely between 2020 and 2026. It is clear from this review that deep learning has genuinely improved watermarking compared to older fixed-formula methods, offering better robustness and adaptability to real-world image changes. At the same time, several challenges remain unresolved, most notably the trade-off between image quality and robustness, the rise of generative AI as a new kind of threat, and the heavy computing resources needed to train these models. Opportunities are equally real, especially around protecting AI-generated content and securing medical images. Researchers working in this field would do well to focus future efforts on generative-AI-resistant designs, shared testing standards, and lightweight models that can reach a wider set of users. It is hoped that this review gives a useful and honest starting point for students and researchers who are new to this fast-growing area.

References

1. Ahmadi, M., Norouzi, A., Karimi, N., Samavi, S., & Emami, A. (2020). ReDMark: Framework for residual diffusion watermarking based on deep networks. *Expert Systems with Applications, 146*, Article 113157.

2. Anand, A., Bedi, J., Aggarwal, A., Khan, M. A., & Rida, I. (2024). Authenticating and securing healthcare records: A deep learning-based zero watermarking approach. *Image and Vision Computing, 145*.

3. Anand, A., & Singh, A. K. (2022). A comprehensive study of deep learning-based covert communication. *ACM Transactions on Multimedia Computing, Communications, and Applications, 18*(2s), 1–19.

4. Anand, A., Singh, A. K., & Zhou, H. (2023). A survey of medical image watermarking: State-of-the-art and research directions. In *Medical Information Processing and Security: Techniques and Applications* (pp. 325–360). Institution of Engineering and Technology. https://doi.org/10.1049/PBHE044E_ch14

5. Ben Jabra, S., & Ben Farah, M. A. (2024). Deep learning-based watermarking techniques challenges: A review of current and future trends. *Circuits, Systems, and Signal Processing, 43*(7), 4339–4368. https://doi.org/10.1007/s00034-024-02651-z

6. Bistroń, M., Żurada, J. M., & Piotrowski, Z. (2026). Deep learning for image watermarking: A comprehensive review and analysis of techniques, challenges, and applications. *Sensors, 26*(2), Article 444. https://doi.org/10.3390/s26020444

7. Fkirin, A., Attiya, G., El-Sayed, A., & Shouman, M. A. (2022). Copyright protection of deep neural network models using digital watermarking: A comparative study. *Multimedia Tools and Applications, 81*(11), 15961–15975.

8. Gao, Y., Kang, X., & Chen, Y. (2021). A robust video zero-watermarking based on deep convolutional neural network and self-organizing map in polar complex exponential transform domain. *Multimedia Tools and Applications, 80*(4), 6019–6039.

9. Ge, S., Xia, Z., Fei, J., Tong, Y., Weng, J., & Li, M. (2023). A robust document image watermarking scheme using deep neural network. *Multimedia Tools and Applications*, *82*, 38589–38612.

10. Hosny, K. M., Magdi, A., El-Komy, O., & Hamza, H. M. (2024). Digital image watermarking using deep learning: A survey. *Computer Science Review, 53*, Article 100662. https://doi.org/10.1016/j.cosrev.2024.100662

11. Hu, Y., Jiang, Z., Guo, M., & Gong, N. Z. (2025). *Secure and robust watermarking for AI-generated images: A comprehensive survey* (arXiv:2510.02384). arXiv. https://arxiv.org/abs/2510.02384

12. Li, Y., Wang, H., & Barni, M. (2021). A survey of deep neural network watermarking techniques. *Neurocomputing, 461*, 171–193.

13. Singh, H. K., & Singh, A. K. (2023). Comprehensive review of watermarking techniques in deep-learning environments. *Journal of Electronic Imaging, 32*(3), Article 031804.

14. Singh, H. K., & Singh, A. K. (2023). Digital image watermarking using deep learning. *Multimedia Tools and Applications*. Advance online publication. https://doi.org/10.1007/s11042-023-15750-x

15. Wang, X., Ma, D., Hu, K., Hu, J., & Du, L. (2021). Mapping based residual convolution neural network for non-embedding and blind image watermarking. *Journal of Information Security and Applications, 59*, Article 102820.

16. Wei, Q., Wang, H., & Zhang, G. (2020). A robust image watermarking approach using cycle variational autoencoder. *Security and Communication Networks, 2020*, 1–9.

17. Zhong, X., Huang, P. C., Mastorakis, S., & Shih, F. Y. (2020). An automated and robust image watermarking scheme based on deep neural networks. *IEEE Transactions on Multimedia, 23*, 1951–1961.