IBIMA Publishing
Communications of International Proceedings
https://ibimapublishing.com/p-articles/46AI/2025/4628125/
Vol. 2025 (34), Article ID 4628125

Lightweight AI Watermarking: Challenges and Research Directions*

Marta BISTROŃ and Zbigniew PIOTROWSKI

Military University of Technology, gen. Sylwestra Kaliskiego 2, 00-902 Warsaw, Poland

Correspondence should be addressed to: Marta BISTROŃ, marta.bistron@wat.edu.pl

* Presented at the 46th IBIMA International Conference, 26-27 November 2025, Ronda, Spain

Abstract

Generative AI has revolutionized the creation of digital media, but it has also exacerbated concerns about authenticity and ownership. Digital watermarking remains a key mechanism for addressing these challenges. However, modern transformation- and diffusion-based approaches, while highly robust, incur significant computational costs that hinder their use in real-time environments. This paper argues that the future of watermarking lies in lightweight methods that combine robustness with efficiency. It reviews the computational burden of current AI-based approaches, presents strategies such as model compression, knowledge distillation, and lightweight architectures, and highlights application areas such as streaming, videoconferencing, gaming, and IoT multimedia. The study highlights the need for watermarking frameworks that are both secure and deployable under strict latency and resource constraints.

Keywords: generative AI, model compression, real-time content protection, lightweight watermarking

Introduction

The rapid development of generative AI has transformed the dynamics of digital media creation and distribution. Modern architectures such as diffusion models and transformers enable the creation of high-quality images, video, and audio on an unprecedented scale. While these advances open up new possibilities for creativity and communication, they also pose a number of challenges and reinforce concerns about authenticity, ownership, and regulatory compliance in digital ecosystems.

Digital watermarking remains one of the most widely used mechanisms for addressing these issues. By embedding imperceptible signals into content, watermarks facilitate provenance tracking and copyright protection. However, practical implementation of watermarking in today's environment faces new challenges. In particular, streaming platforms, live video conferencing, and real-time content delivery require solutions that are not only effective but also computationally efficient.

Modern AI-based watermarking approaches demonstrate impressive resilience to distortion and manipulation, but they rely on architectures that require significant computational power. Watermark embedding and extraction using diffusion or transformer methods often require significant processing resources and introduce latency, making them unsuitable for real-time or large-scale scenarios.

In this paper, the authors argue that future research must place greater emphasis on lightweight watermarking solutions -approaches that combine robustness with efficiency and can be implemented in real-time environments. The following sections discuss: Section 2 - the computational burden of current AI watermarking methods; Section 3 - strategies for developing lightweight alternatives to expensive architecture computations; Section 4 - key application scenarios where such methods are urgently needed; and Section 5 - summary and key conclusions.

Cite this Article as: Marta BISTROŃ and Zbigniew PIOTROWSKI, Vol. 2025 (34) "Lightweight AI Watermarking: Challenges and Research Directions " Communications of International Proceedings, Vol. 2025 (34), Article ID 4628125, https://doi.org/10.5171/2025.4628125

Computational Burden of AI-Based Watermarking

AI-based watermarking methods have demonstrated remarkable resilience to both traditional distortions and modern attacks, but they incur significant computational costs. Three main groups of approaches can be distinguished:

- early deep learning frameworks based on convolutional neural networks or autoencoders and incorporating GAN architecture elements (Kandi, Mishra and Gorthi, 2017; Zhu et al., 2018);
- attention-based algorithms, including in particular transformers, vision transformers (ViTs), and swin transformers (shifted windows transformers) (Aberna, Agilandeeswari and Aashich, 2023);
- diffusion-based solutions (Zhang et al., 2024).

Early watermarking systems based on convolutional neural networks (CNNs) or autoencoder architectures, sometimes incorporating elements of generative adversarial networks (GANs), ushered in the first wave of deep learning in watermarking. These models offered improved performance and robustness compared to classical signal processing methods while maintaining acceptable image quality. However, the computational demands of training and inference were already significant. Data embedding and extraction pipelines relied on deep convolutional layers, resulting in high memory consumption and relatively long processing times, limiting their practicality for large-scale or real-time deployments.

The emergence of attention-based architectures brought further advances. Vision transformers (ViTs) and their derivatives, such as swin transformers, introduced greater flexibility and scalability, enabling the watermarking of multimodal data and higher-resolution content. Their ability to model long-range dependencies provided a natural advantage in tasks such as video watermarking. However, this capability came at a price: transformers are parameter-heavy, requiring significant computational power and memory. The quadratic complexity of attention operations with respect to sequence length further exacerbates these problems, making transformer-based watermarking impractical in time-sensitive scenarios such as streaming or live videoconferencing.

Recently, diffusion-based watermarking has set a new standard for robustness. By embedding signals in a latent space and iteratively denoising, these methods are resistant to compression, noise, and even regeneration attacks from advanced generative models. At the same time, the very design that makes diffusion effective also introduces high latency: multiple forward and backward passes are required for a single sample, often resulting in very long processing times for each image. This makes diffusion methods impractical for continuous data streams. So, while diffusion watermarking offers state-of-the-art robustness, it highlights a continuing trade-off in this field: greater robustness typically comes at a higher computational cost, reinforcing the need for research into alternative, lightweight solutions. Table 1 summarizes the advantages and capabilities of each solution and their limitations from a computational cost perspective.

Table 1. Computational cost perspective of AI-based watermarking approaches.

Approach	Advantages	Limitations
CNN/Autoencoder/GAN	Improved robustness over classical methods; good visual quality; relatively mature.	Heavy training; high memory usage; limited efficiency for large-scale or real-time tasks.
Attention	Flexible, scalable; well-suited for multimodal and high-resolution content; strong modeling of dependencies.	Parameter-heavy; high memory demand; quadratic complexity; impractical for streaming or long sequences.
Diffusion	State-of-the-art robustness; resistant to compression, noise, and regeneration attacks.	Multiple forward/reverse passes; high latency; energy-intensive; unsuitable for real-time deployment.

Strategies for Lightweight Watermarking

Addressing the computational burden associated with AI-based watermarking requires methods that maintain robustness while reducing model complexity and inference time. There are many strategies developed in other frameworks that can be successfully adapted to the field of watermarking.

One of the most common approaches is model compression, which is used in virtually every AI-based industry, especially for mobile devices. Compression reduces the size and cost of models without requiring retraining from scratch. Techniques such as:

- pruning - removing redundant or irrelevant weights (Cheng, Zhang and Shi, 2024),
- quantization - representing parameters using lower-precision integers (e.g., INT8 instead of FP32) (Nagel et al., 2021),

are widely used to deploy deep models on edge devices. In watermarking, compression can significantly reduce the latency of embedding and extraction, but caution must be exercised to avoid compromising the watermark's robustness in the event of an attack.

The second strategy is knowledge distillation, in which a large, efficient "teacher" model is used to train a smaller "student" model that mimics its behavior. This method is primarily used for diffusion models, where a teacher model is trained to achieve the measured result in hundreds of iterations, and then a student model is trained based on this, which achieves the same result in just a few iterations. In the case of watermarking, this can enable efficient extraction networks that retain the accuracy of diffusion-based and transformer-based pattern detection while operating on limited hardware. Distillation is particularly promising in real-time scenarios, offering a balance between robustness and efficiency.

The main limitation of both strategies is the costly and time-consuming training phase, with efficiency benefits visible only at deployment.

Another promising direction, effective both in training and deployment, is the use of lightweight architectures specifically designed for performance. Architectures such as MobileNet (Howard et al., 2017), ShuffleNet, and EfficientNet demonstrate that high accuracy is achievable with drastically fewer parameters and lower FLOPs. Integrating watermarking targets with such frameworks can result in models that are naturally suited for implementation in streaming, IoT, or mobile environments.

Another emerging area worth considering is hardware-aware design. Implementing watermarking methods directly on accelerators such as FPGAs or ASICs can provide real-time performance while reducing power consumption. This direction aligns with the growing demand for watermarking in live streaming and videoconferencing, where latency constraints are extremely tight.

Application Scenarios and Future Directions

Lightweight watermarking is not an abstract research topic but a practical necessity. In modern digital ecosystems, content is often created, transmitted, and consumed in real-time, which imposes stringent requirements on latency, performance, and power consumption. Figure 1 illustrates the main areas where lightweight watermarking is crucial: streaming platforms, video conferencing, gaming and VR environments, and IoT multimedia. Each of these scenarios highlights different constraints - from latency to energy efficiency - that traditional AI-based watermarking methods cannot easily address.

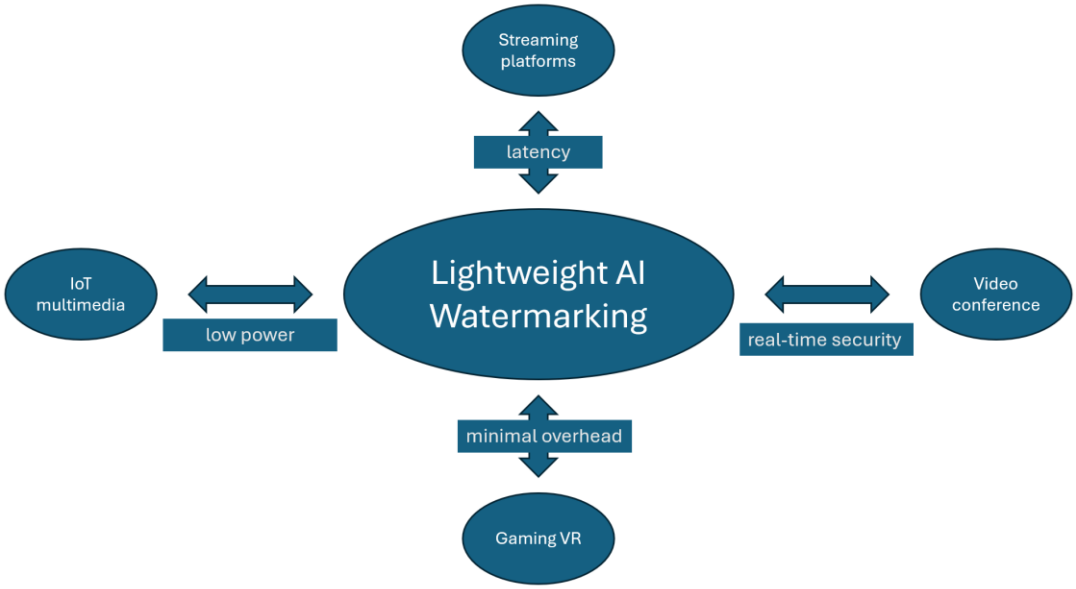

Figure 1. Application scenarios for lightweight AI watermarking.

Lightweight watermarking is particularly important for streaming platforms, where even minor processing delays can disrupt the continuity of the audiovisual transmission. Therefore, embedding or verifying signals must be performed with minimal latency, without degrading the stream quality.

In video conferencing, the challenge extends beyond latency to include privacy and trust. Participants must be assured that the content being shared is authentic, and watermarking must not introduce delays that would impair real-time communication.

In gaming and virtual reality (VR), watermarking must coexist with highly dynamic and interactive environments. In this case, the computational load directly impacts responsiveness and user experience, making performance a prerequisite for practical implementation.

IoT multimedia systems, such as smart cameras and surveillance networks, operate under strict resource and power constraints. Watermarking methods for these devices must be both lightweight and energy-efficient, ensuring security without burdening limited hardware capabilities.

Summary

The evolution of watermarking methods has highlighted the constant trade-off between robustness and performance. Although modern AI-based solutions, such as transformers and diffusion models, offer high levels of robustness, their high computational cost limits their application in real-time environments. This paper argues that the future of watermarking lies in lightweight approaches that balance robustness with latency, scalability, and energy constraints.

Strategies such as model compression, knowledge distillation, and lightweight architectures can pave the way for practical watermarking solutions suitable for streaming, video conferencing, gaming, and the IoT. These domains require not only secure and reliable watermarking but also methods that can be implemented on constrained hardware without compromising the user experience.

Future research should therefore focus on bridging the gap between robustness and performance, ensuring that watermarking can meet the demands of real-time systems while adapting to emerging standards and regulatory frameworks.

Acknowledgment

This research work was funded by the project of the Military University of Technology titled: "*New Neural Network Architectures for Signal and Data Processing in Radiocommunications and Multimedia*." Project No. UGB 22-054/2025.

References

- Aberna, P., Agilandeeswari, L. and Aashich, B. (2023), Vision Transformer-Based Watermark Generation for Authentication and Tamper Detection Using Schur Decomposition and Hybrid Transforms, International Journal of Computer Information Systems and Industrial Management Applications, vol. 15, pp. 107–121.
- Cheng, H., Zhang, M. and Shi, J.Q. (2024), *A Survey on Deep Neural Network Pruning – Taxonomy, Comparison, Analysis, and Recommendations,* arXiv preprint arXiv:2308.06767. https://doi.org/10.48550/arXiv.2308.06767.
- Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. and Adam, H. (2017), *MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,* arXiv preprint arXiv:1704.04861. htt Kandi, H., Mishra, D. and Gorthi, S.R.K.S. (2017)ps://doi.org/10.48550/arXiv.1704.04861.
- Kandi, H.; Mishra, D.; Gorthi, S.R.K.S. (2017), Exploring the Learning Capabilities of Convolutional Neural Networks for Robust Image Watermarking, Computers & Security, vol. 65, pp. 247–268. https://doi.org/10.1016/j.cose.2016.11.016.
- Nagel, M., Fournarakis, M., Amjad, R.A., Bondarenko, Y., van Baalen, M. and Blankevoort, T. (2021), *A White Paper on Neural Network Quantization,* arXiv preprint arXiv:2106.08295. https://doi.org/10.48550/arXiv.2106.08295.
- Zhang, L., Liu, X., Martin, A.V., Bearfield, C.X., Brun, Y. and Guan, H. (2024), *Attack-Resilient Image Watermarking Using Stable Diffusion,* arXiv preprint arXiv:2401.04247. https://doi.org/10.48550/arXiv.2401.04247.
- Zhu, J., Kaplan, R., Johnson, J. and Fei-Fei, L. (2018), *HiDDeN: Hiding Data With Deep Networks,* arXiv preprint arXiv:1807.09937. https://doi.org/10.48550/arXiv.1807.09937.