ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 1, 2025

AI BASED SECRET IMAGE GENERATION USING WATER MARKING TECHNOLOGY

J.Preethi[1], Sana Tabussum[2], Boddupalli Mohanbabu[3], Ganji Dheeraj Kumar[4], Shaik Asif[5]

[1] Assistant Professor, Dept. of AI-ML, Sri Indu College of Engineering and Technology, Hyderabad,
[2 3 4] Research Student, Dept. of AI-ML, Sri Indu College of Engineering and Technology, Hyderabad

Abstract: AI-Generated Content (AIGC) has made significant advancements in both popularity and realism. While the development of generative large models offers immense potential to enhance creativity and operational efficiency, it also introduces a range of risks and challenges, particularly concerning issues such as copyright infringement due to model misuse and the authenticity of generated content. In response to the need for standardized management and application of AIGC models, researchers are increasingly focusing on exploring effective strategies for managing and protecting the authentication of AIGC models, as well as ensuring the traceability of generated images through digital watermarking technologies. This survey provides a comprehensive review of three core areas: the evolution of image generation technologies, traditional and state-of-the-art digital image watermarking algorithms, and watermarking methods specific to AIGC. Additionally, we examine common performance evaluation metrics used in this field. Finally, we discuss the unresolved issues and propose several potential directions for future research.

Keywords: digital image watemarking; image security; AIGC watermarking; deep learning

1. Introduction

The swift advancement of artificial intelligence (AI) is profoundly transforming society and reshaping people's lifestyles. In this progression, Artificial Intelligence Generated Content (AIGC), at the cutting edge of AI technology, has the remarkable advantage of being able to create a massive amount of high-quality content with far greater efficiency, reduced cost, and superior quality than traditional previous techniques. The pervasive application of this technology has seamlessly integrated into daily life, such as literature review [1], medical diagnosis [2], and painting [3]. However, with the proliferation and application of AIGC technology, a myriad of new security challenges have emerged, including concerns about content authenticity, intellectual property rights protection and the potential abuse risks [4–6]. Therefore, comprehensive research is imperative to address these issues and foster the healthy evolution of AIGC technology.

Digital watermarking is a potent copyright protection technology that embeds watermark information into digital carriers (text, images, audio, etc.) to obtain watermarked versions, allowing for the extraction of this information when necessary to confirm the copyright ownership [7,8]. With the aim of continuously improving performance, traditional digital image watermarking technologies are being developed primarily from two perspectives: the spatial domain and the transform domain. With the flourishing of deep learning, digital image watermarking schemes have started using the encoder-decoder structure for watermark embedding and extraction, which can effectively enhance its robustness in a variety of application scenarios while ensuring high invisibility. Subsequently, watermarking technology has also been adopted to safeguard the copyright of deep neural network models, verifying their ownership through a similar embedding and extraction concept. Therefore, in the context of AIGC, digital watermarking technology is also used to accomplish the purpose of the AIGC image traceability and the associated model ownership verification, which can effectively address these existing security threats [9].

This paper focuses on the image part and provides an overview of the existing image watermarking techniques. Firstly, the development history of four image generation technologies is reviewed, followed by the corresponding classification descriptions of digital image watermarking algorithms and AIGC model watermarking algorithms. Finally, the performance indicators of watermarking algorithms are briefly mentioned.

The organizational structure of this survey is as follows: Section 2 presents four types of image generation techniques, including Generative Adversarial Networks (GANs), Transformer models, Variational Autoencoders (VAEs), and diffusion models. Section 3 represents two kinds of digital image watermarking schemes, including the traditional-based and the deep learning-based. Watermarking methods of the AI model and AI-Generated images are involved in Section 4. Section 5 includes some various performance evaluation metrics for image watermarking algorithms. Section 6 discusses some promising research directions. Finally, the survey is concluded in Section 7.

2. Image Generation Technologies

In recent years, the field of image generation has experienced remarkable advancement as a crucial application of artificial intelligence technology. Generative models rooted in deep learning have been capable of producing spectacular visual content, finding wide application across a spectrum of disciplines, such as art creation, entertainment, scientific research, and commercial applications. With the improvement of hardware algorithm capabilities, large-scale generative models have rapidly evolved, emerging as a substantial driving force at the forefront of artificial intelligence, distinguished by their extraordinary creativity and intelligent expression. In terms of quality and diversity of image generation, these models have reached unprecedented levels, indicating the immense potential of large-scale models in the realm of image generation. This section mostly concentrates on the generative adversarial networks, the Transformer models, the variational autoencoders, and diffusion models to introduce the current development status of image generation technologies.

2.1. Generative Adversarial Networks (GANs) Based Techniques for Image Generation

In 2014, Goodfellow et al. [10] published a pivotal academic paper introducing Generative Adversarial Network (GAN) and marking its inception. The core idea of GAN is grounded in a zero-sum game strategy, with a network structure typically comprising a generator G and a discriminator D. G mainly learns to generate data that appears to be authentic, while D primarily focuses on distinguishing whether the input data originates from a real dataset or is generated. During the training phase, G and D engage in competition to progressively enhance the generator's performance. The ultimate goal is to make the data generated by G challenging to be discerned by D. Considering a latent vector z with a probability distribution of p_z, while the probability distribution of the real sample x is p_{data}. The corresponding objective loss function of GAN is:

$$L_{GAN} = \mathrm{E}_{x\sim p_{data}}[\log D(x)] + \mathrm{E}_{z\sim p_z}[\log(1 - D(G(z)))], \quad (1)$$

where the probability values $D(x)$ and $D(G(z))$ represent the corresponding outputs of the discriminator for original and generated samples, respectively. Throughout the entire training process, the performance of the generator enhances as it learns to generate more realistic data, while the capabilities of the discriminator improve as it better distinguishes between real data and fake data.

GAN is mainly used for generating image data, and through ongoing optimization by researchers, numerous distinctive GAN architectures have emerged to improve the quality of generated images. DCGAN [11] substituted multiple-layers perceptrons with convolutional neural networks and introduced the batch normalization techniques, resulting in greatly improved image quality and stability. The emergence of DCGAN has sparked significant attention among researchers, showcasing the feasibility of incorporating the neural networks into GAN. Reed et al. [12] first introduced the GAN-INT-CLS

method for text condition-driven image generation, which lies in using the text embedding technique to encode descriptive text into vectors and integrate them with a generator network. Therefore, the generator can generate the corresponding images based on specific text descriptions. To further align image details with text descriptions, Xu et al. [13] developed AttnGAN, which was unique in integrating the attention mechanisms, allowing for a more accurate reflection of text details in varying image regions, leading to higher quality and resolution. Progressive Generative Adversarial Network (ProGAN) was proposed by NVIDIA in 2017 [14], which gradually increases image resolution during training, starting from a low resolution and adding more layers progressively. Although ProGAN solves the difficulty in generating high-resolution images, its ability to control specific features of generated images is still limited. Style Generative Adversarial Network (StyleGAN) was devised by Karras et al. [15] to tackle this issue, incorporating the notion of "Style", and allowing the network to independently regulate the appearance elements of generated images. This technique can not only unsupervise these distinct advanced attributes like expressions, poses, and identities, but also learn randomly transformed image details by adding noise like hair and freckles, greatly improving the image quality and user control over the scale of particular image attributes. Subsequently, Kang et al. [16] proposed an open-source library called StudioGAN, which was one of the latest developments in the application of GAN for real-world image synthesis. StudioGAN supports a wide array of GAN architectures, tuning methods, adversarial losses, regularization modules, differentiable enhancements, evaluation metrics, and evaluation baselines, which can facilitate the development of GAN research and simplify the reproduction and comparison of GAN models for researchers.

Due to the challenge of the disparity between text and image domains, relying solely on the discriminator is inefficient and struggles to ensure semantic alignment between the generated images and the input text. Consequently, Qiao et al. [17] utilized the concept of the re-description to learn text and image generation, and designed MirrorGAN based on the semantic text embedding module, the global local collaborative attention module, and the semantic text regeneration alignment module to mirror and regenerate text descriptions from generated images, achieving higher quality image generation aligned with semantics. Pan et al. [18] explored DragGAN to enhance the controllability of GAN, synthesizing visual content that aligns with user's requirements, which allowed for flexible and precise control over the pose, shape, expression, and layout of generated objects. Based on feature-based motion supervision and novel point tracking methods, users can use DragGAN to accurately manipulate the poses, shapes, expressions, and layouts of various categories, ensuring realistic output results even in the face of some challenging scenes such as content illusion masking and shape distortion. With the expansion of model parameters and the rapid construction of cross-modal pre-trained models, GAN-based image generation models have made swift advancements. The GALIP (Generative Advanced CLIPs) proposed by Tao et al. [19] leveraged the CLIP model to design loss functions for generators and discriminators, aligning the GAN feature space with the CLIP to guide the generation of visual concepts. Along with increasing training efficiency, it also reduces computational costs while achieving performance comparable to large pre-trained models.

The GAN series of networks have achieved tremendous success over the past ten years, with its key advantage lying in constructing network models by drawing on the adversarial training networks to reach Nash equilibrium, thereby generating sufficiently realistic data. However, GANs also face some challenges such as the unstable training process, pattern collapse, and a lack of diversity in the generation samples.

2.2. Transformer-Based Techniques for Image Generation

The Transformer model [20], characterized by its unique self-attention mechanism and powerful representation learning capability, has been widely adopted in natural language processing and computer vision, achieving numerous breakthroughs. In cross-modal image generation tasks, Transformer-based autoregressive models have also demonstrated superior performance. The fun-

damental strategy of the Transformer model involves using a specially trained decoding network to convert the feature sequences into a complete image after an encoder predicts the feature sequence of an image based on other input conditions such as text and sketches. The self-attention mechanism enables parallel data processing, significantly boosting computational efficiency and successfully capturing distant connections in feature sequences. Esser et al. [21] proposed an image generation model VQGAN that included a Transformer. The model first trained a convolutional neural network-based GAN to produce the compressed images, and then selected a Transformer network-based image compression model to restore these compressed images to the original real images, thereby generating high-resolution images. Employing solely the Transformer architecture and no convolutional layers, Jiang et al. [22] constructed a TransGAN network, which can generate various image contents with high fidelity and reasonable texture details. Using Vision Transformers instead of Convolutional Neural Networks, Lee et al. [23] designed VitGAN, which not only guaranteed stability and convergence of the model training phase but also significantly improved the quality of image generation. Peebles et al. [24] introduced the Transformer architecture into diffusion models, and proposed DiTs, which replaced the conventional U-net backbone used in diffusion models with Transformers and demonstrated significant advantages in improving generated image quality.

2.3. Variational AutoEncoder (VAE)-Based Techniques for Image Generation

The goal of the Variational Autoencoder (VAE) [25] is to generate novel samples by learning the latent distribution of data, marking the beginning of significant progress in deep learning in the field of image generation. VAE maps the input data to a latent space via an encoder, then reconstructs these latent variables back to the data space using a decoder, thereby generating high-quality samples akin to the training data.

To achieve more precise control over generated images, Sohn et al. proposed the Conditional Variational Autoencoder (CVAE) [26], whose basic idea is to take the conditional information in the sampling process of latent variables into account. This allows the model to generate samples with specific features based on additional information, making it more suitable for tasks requiring specific conditions, such as image restoration and style transformation. Using a cyclic variational autoencoder with an attention mechanism, Mansimov et al. [27] designed alignDraw from text to image, which can generate the corresponding images using text that is not in the training dataset. In order to solve potential representation ambiguity problems in the traditional autoencoders, Van et al. [28] put forward the vector-quantized variant autoencoder (VQVAE), which discretized the encoding in the latent space to obtain the nearest discrete code vector approximating the latent representation. During the generation stage, this nearest vector was mapped back to a continuous latent representation, effectively improving the visual quality of the generated image. Huang et al. [29] developed IntroVAE, which integrated the concept of GAN to enable self-evaluation of the quality of generated samples and self-improvement to enhance generation performance. Based on the IntolVAE, Daniel et al. [30] introduced soft IntroVAE, replacing the hinge loss term of generated samples with a smooth exponential loss, significantly improving the stability of training phase.

Despite their notable achievements in tasks such as image generation and latent representation learning, VAEs, and their variations still have drawbacks. Firstly, due to the use of variational inference and reparameterization, VAEs often struggle to capture fine details in the data distribution. Secondly, VAEs typically assume that the latent space follows the Gaussian distribution, which limits their ability to perform on complex distributions. These limitations have resulted in the tendency for VAEs to produce blurry results in the data reconstruction process, making their performance in generating high-resolution images generally inferior to other generative models.

2.4. Diffusion-Based Techniques for Image Generation

Ho et al. [31] introduced the Denoising Diffusion Probability Model (DDPM), which simulates diffusion and recovery processes in physical systems to generate high-quality data samples. It achieves realistic image generation by adding noise into the existing images and progressively eliminating it. This approach, based on random processes, yields remarkable results in terms of generation quality and diversity, and it also contributes to the model having a more stable training process. The release of DDPM has shifted the focus from GANs to diffusion models, establishing them as the mainstream in image generation. However, one drawback of DDPM is its relatively slow sampling speed. To address this issue, Nichol et al. [32] proposed an improved DDPM with learnable variance in the reverse process and cosine noise in the forward process, thereby achieving faster sampling and better likelihood estimation. By replacing the original Markov process with a non-Markov process, Song et al. [33] presented the Denoising Diffusion Probabilistic Model (DDIM), which utilized an effective sampling strategy that greatly improved the sampling speed while minimizing the impact on the quality of the generated samples.

In order to implement the conditional control during random sampling, Dhariwal et al. [34] first developed a classifier-guided diffusion process that steered the diffusion process towards specific classification labels, thereby exerting a degree of control over the generated images. Building upon this foundation, Ho et al. [35] introduced a classifier-free guided diffusion process, enabling the joint training of both conditional and unconditional diffusion processes. The sampling procedure is adjusted according to the scaling factor, a method that relies solely on the diffusion process without requiring a classifier model. To further improve the performance of text-to-image generation, GLIDE [36] incorporated randomly sampled Gaussian noise along with text embeddings, obtained through the CLIP model, into the diffusion model for training. Additionally, a CLIP feature-based guided diffusion loss was used to compute the similarity between the noisy image and the input text at each specific diffusion time step. Despite these advancements, however, the diversity of generated images is insufficient. Imagen [37] released by Google showed an outstanding performance in generating image details and understanding complex textual input. In addition to optimizing the inference time, the convergence speed and memory efficiency within the U-Net architecture, noise-level conditioning techniques and cascaded diffusion models were also applied to improve the generation quality. DALL·E-2 [38] consisted of a prior model and a decoder model. The former was responsible for producing the CLIP image embeddings based on text prompts, while the latter generated the corresponding images based on the image embeddings. This architecture facilitates a variety of innovative generation tasks, including image editing and image generation.

In order to enhance the efficiency and stability of the training process for diffusion probabilistic models, the Latent Diffusion Model (LDM) [39] compressed these high-frequency details of the original image into a latent space that could fully capture the perceptual content of the image, thereby reducing computational complexity while maintaining the generative capabilities of the diffusion model. By leveraging the extensive LAION-2B dataset [40] with 2 billion images for training, the research team released a series of LDM-based pre-trained models, collectively known as "Stable Diffusion" [41]. At present, Stable Diffusion stands as the most widely discussed and applied diffusion model in the field of image generation.

3. Digital Image Watermarking

As illustrated in Figure 1, the watermark is embedded into the original image to protect it. After some possible attacks, the watermark can still be extracted from the watermarked image, where both the watermark embedding and extraction processes share a common key. Based on the difference in the embedding domain of the watermark information, digital image watermarking technologies can be divided into two types: spatial domain watermarking [42] and transform domain watermarking

ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 1, 2025

[43]. Spatial domain image watermarking algorithms embed the watermark information directly into the pixel values of the host image, while transform domain watermarking algorithms embed the watermark into the coefficients in the transformed domain and then perform an inverse transform on the modified coefficients back to the spatial domain to obtain the watermarked host image.

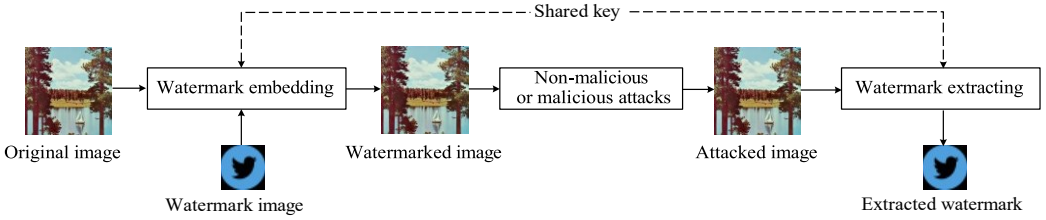

Figure 1. The basic process of digital image watermarking scheme.

3.1. Traditional Image Watermarking

The most commonly spatial domain watermarking algorithm [44] is based on the Least Significant Bit (LSB) technique, which alters only the least significant bits, minimizing the impact on the visual quality of the image. However, it is easily susceptible to image attacks. Patchwork [45] is another classical spatial domain algorithm, which embeds watermark information into the statistical features of the image, fully leveraging the human visual system's insensitivity to slight changes. **Spatial domain-based image watermarking** algorithms [46,47] are straightforward to implement and have low computational complexity, but they suffer from poor robustness and are easily compromised by image attacks or cracked.

Transform domain watermarking algorithms, on the other hand, exhibit relatively stronger robustness and can withstand some common image processing attacks, such as scaling, rotation, and filtering of the watermarked image. Notable transform domain watermarking algorithms include Discrete Cosine Transform (DCT) [48], Discrete Wavelet Transform (DWT) [49], and Fractional Fourier Transform (DFT) [50]. Although single transform domain schemes provide some improvement in effectiveness, they are still inadequate to meet the increasing demands, prompting the emergence of hybrid transform domain image watermarking methods. Wang et al. [51] performed multi-level DWT and singular value decomposition on the host image to embed the watermark information. Compared to a single transform domain, this method showed improved robustness against geometric attacks and some composite attacks. Similarly, Mohammed et al. [52] employed a continuous DCT-DWT transformation, adaptively selecting the RGB components of the color image for watermark embedding. According to experiments, this method was resilient to geometric and filtering attacks and has good invisibility. The performance improvement of multi-transform domain watermarking algorithms stems from their ability to combine the benefits of various transform techniques, performing multi-level transformations on the host image. This disperses and conceals the watermark information more effectively, thereby enhancing the robustness of the watermarked image. At the same time, multi-transform domain algorithms [53,54] usually also consider the adaptability and invisibility of the image to achieve an optimal balance between watermark robustness and image quality.

3.2. Deep Learning-Based Image Watermarking

Conventional image watermarking schemes often exhibit limited generalization capabilities and struggle to defend against novel types of attacks. These manually designed methods, which heavily rely on designers' domain knowledge and specific application scenarios, tend to lack robustness. However, with the rapid advancement of deep learning, utilizing neural networks for watermark embedding has emerged as the dominant trend in image watermarking research for image copyright protection.

Based on convolutional neural networks (CNN), Haribabu et al. [55] firstly proposed a robust image watermarking model, where watermark information was embedded by learning image features through an autoencoder. In the embedding phase, Kandi et al. [56] employed two types of autoencoder structures for feature extraction: one for embedding and the other for extraction. The same autoencoder network was then used in the receiving phase to retrieve the watermark information. HiDDeN [57] achieved the generation of visually indistinguishable watermarked images and the extraction of watermark information through joint training of encoder and decoder networks. Ahmadi et al. [58] adopted two fully convolutional networks with residual structures for watermark information embedding and extraction, while also designing a differentiable model for JPEG compression, which can effectively enhance the robustness against JPEG compression attacks. Mellimi et al. [59] proposed a deep neural network model capable of incorporating the watermark information into the high-frequency information of the wavelet domain. While this approach improved watermark invisibility, the robustness of the watermarked images was weakened due to the limited features provided by high-frequency information. Building on HiDDeN, Hao et al. [60] added a high-pass filter before the discriminator and gave higher weight to the central region of the image. This model was more robust to noise interference, but it concentrated the watermark information in the central area, resulting in degraded visual quality of the watermarked image. Tancik et al. [61] put forward the StegaStamp watermarking scheme, which embedded bit-string watermarks into photo images with exceptionally high perceptual invisibility. The encoder-decoder architecture, trained through deep learning, effectively enhances the adaptability of this watermarking scheme against various types of attacks.

In practical applications, digital image watermarking must also account for physical attacks in real-world scenarios, such as distortions caused by printing and screen capturing. To strengthen the robustness of digital image watermarking technology against such physical attacks, Liu et al. [62] designed an anti-printing watermarking scheme, which involved the end-to-end training on a large dataset of film overlay layer images to improve resilience against film overlays on printed photos. The DeNoL watermarking method, put out by Fang et al. [63], could enhance the robustness of watermarking systems across channels (such as screen capture scenarios) by using small sample data. Compared to traditional technology-based watermarking methods, these deep learning-based watermarking algorithms achieve superior invisibility and stronger robustness, providing effective solutions for the performance optimization of watermarking algorithms when they face different application environments.

4. AIGC Model Watermarking

With the advent of the era of large models, the commercialization of AIGC models has given rise to a series of security issues. For instance, generative image models like Stable Diffusion require vast quantities of high-quality image data and continuous computational resources for training, which not only entails substantial financial costs but also raises concerns about the commercial value of training data and the protection of intellectual property. In addition, the problems related to false dissemination and traceability of AIGC-generated content further exacerbate security challenges, including privacy breaches, ethical violations and the prevention of inappropriate content generation. These issues such as unauthorized content synthesis, copyright conflicts, and the spreading of fake images are particularly prominent in the realm of image generation. To address these security threats, digital watermarking technology is gradually being integrated into AIGC models and their generated content, forming two main aspects:

- **Model right authentication watermarking**: aimed at protecting the intellectual property rights of AIGC models, which can primarily be divided into white-box watermarking and black-box watermarking, and diffusion model authentication watermarking is also included.

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 1, 2025

- **Generation image traceability watermarking**: used to track the copyright of generated images of the diffusion model, including two types: adding to latent space and adding to initial noise.

The application of these watermarking technologies plays a crucial role in addressing content security issues in the field of image generation and in protecting the intellectual property rights of model owners, contributing to the construction of a safer and more trustworthy ecosystem for AI-generated content .

4.1. Watermarking of AI Model

The initial goal of model right authentication watermarking is to protect the copyright of deep neural networks (DNNs). Depending on whether the watermark extractor needs to understand the internal details of DNNs, these methods can be roughly classified into white-box model watermarking and black-box model watermarking. Furthermore, for both white-box watermarking and black-box watermarking, their classical processes of watermark embedding and copyright verification are depicted in Figure 2, respectively. With the burgeon of text-to-image diffusion models, such as Stable Diffusion, the watermarking algorithms for ownership verification of diffusion models have also emerged one after another to address copyright protection concerns.

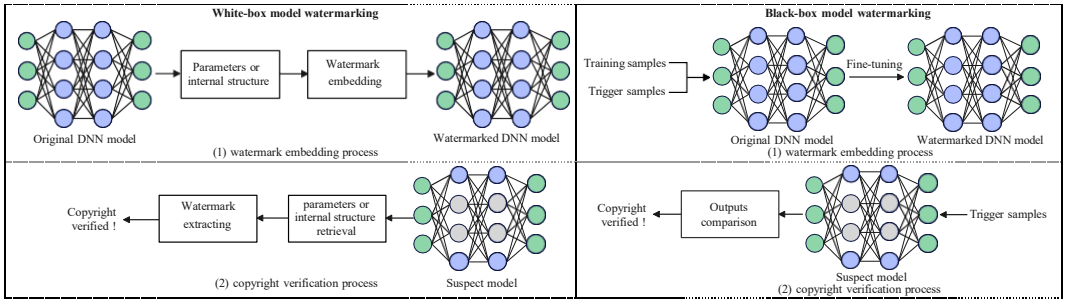

Figure 2. White-box model watermarking and black-box model watermarking.

4.1.1. White-Box Model Watermarking

White-box model watermarking algorithms [64–66] emphasize the accessibility of the internal details of the target DNN models during the watermark extraction process. By obtaining the concrete information like the parameters and internal structure of the DNN model, they accomplish the copyright verification of the target model. The design concept mainly involves implementing white-box model watermarking technology by adjusting the model's parameters or network structure to embed the watermark. Uchida et al. [67] first proposed a white-box scheme for the copyright protection of neural network models with the digital watermarking technology, which embedded the watermark by selecting the model's weights and designed activation functions to constrain the product of a custom spreading matrix and the weights within a specific range, thereby reducing the impact on the model's original classification accuracy. To enhance flexibility, Rouhani et al. [68] adopted the layer activation outputs directly related to the input content as the watermark embedding location, establishing a clever mapping relationship with the inputs. However, the watermark capacity in this scheme is relatively small. To improve security, Chen et al. [69] exploited the collusion-resistant codes to generate watermark information, achieving a watermarking scheme capable of resisting collusion attacks and tracing model users. Wang et al. [70] created an adversarial training mechanism to constrain the distribution of weights before and after watermark embedding, enhancing the invisibility of the watermark. To further strengthen the robustness and mitigate the impact of embedding capacity, Tartaglione et al. [71] proposed a zero-watermark white-box scheme, which randomly selected and marked the parts of the model's weights. They constrained the weight updates during the training process via a loss function, and subsequently verifying the copyright. The above-described white-box watermarking approaches regard the parameters as the carrier.

Then Lou et al. [72] proposed a watermarking scheme that uses the model structure as the carrier, employing a neural network search strategy to determine the optimal network structure for watermark embedding. During the verification phase, the side-channel information is captured to obtain the model structure and extract the watermark information. However, the requirement for suspicious model content information, which is challenging to get in real-world application scenarios, limits the practicality of the white-box watermarking approach in the verification stage. As a result, white-box watermarking algorithms will not have a particularly wide variety of applications.

4.1.2. Black-Box Model Watermarking

Black-box model watermarking schemes [73–75] can overcome this limitation. In these schemes, the model's copyright verifier only needs to query the model's API interface, without accessing the model's internal structure. The general framework of the black-box model watermarking scheme [76] is as follows: during the watermark embedding stage, the model owner uses both the normal training samples and the selected trigger samples to train the model that needs to be protected, resulting in a watermarked model; in the copyright verification stage, trigger samples are input into the target model to obtain the predicted results. If the predicted outputs are consistent with the pre-set ones, the copyright verification is successful, confirming that the verifier holds the model's copyright. The black-box model watermarking schemes take full advantage of the redundant characteristics in deep neural network neurons and represent the unique copyright identification of the model owner through the specific input-output pairs. In order to better fit practical application scenarios, researchers mainly focus on the advancement of the black-box model watermarking.

The concept of backdoor triggering is a typical technique for verification. Zhang et al. [77] studied the corresponding performance of different triggering backdoor forms, including images independent of the training set, random noise and the addition of particular strings to selected parts of the training dataset. Guo et al. [78] used the user information to guide the generation of embedded noise as a trigger backdoor, establishing a link between the noise and the model users, thereby clarifying the specific ownership. Based on the genetic algorithm, Guo et al. [79] continued to search for the optimal watermark embedding position for trigger images in their previous scheme [78]. Sun et al. [80] employed an additional category for the triggers but selected the images unrelated to the training set as the trigger set, resulting in weaker concealment. Li et al. [81] concealed the copyright information in the adopted training samples in an invisible manner to generate trigger samples and constrained the distribution between trigger samples and training ones, effectively improving the secrecy of trigger images.

On the basis of the black-box verification framework, various model watermarking schemes commit to optimizing the performance from distinct perspectives. In response to the challenge of labeling trigger images, Zhang et al. [82] suggested using the chaos algorithm to label trigger images to defend against possible forgery attacks. In order to resist escape attacks, where autoencoder reconstruction strategies are used to modify trigger images, Li et al. [83] added a watermark embedding stage in the autoencoder reconstruction process, thereby improving the robustness against such attacks, albeit at the cost of some reduction in the original task accuracy. Based on the irreversibility of one-way hash functions, Zhu et al. [84] designed a series of continuous one-way hash functions to scramble trigger images and assign the same pre-set labels, ensuring the security of triggers during the verification process and thus guarding against obfuscation attacks. In order to counter model-stealing attacks, Szyller et al. [85] created a defense mechanism that modified the returned results in response to the accessing data from the attackers. Similarly, Charette et al. [86] injected a specific signal distribution into the output, allowing the stolen model to learn this distribution and withstand this attack. Li et al. [87] proposed an initial model training strategy that tightly coupled the model's original performance with the watermark pattern. If an attacker wants to forge the watermark pattern, they must spend a huge computational cost to jump out of the local optimal solution. In a condition when an attacker

knows the triggering mode and the corresponding target label, they can trigger a backdoor attack. To address this issue, Xu et al. [88] applied the confidence scores from each category of the triggered image output as watermark information, thereby enhancing the security of the watermark. Li et al. [89] used a backdoor-free method for watermark embedding, weakening the connection between specific triggering patterns and the target labels. Therefore, attackers cannot trigger backdoor attacks based on the obtained triggering information.

4.1.3. Diffusion Model Watermarking

The prevailing copyright authentication watermarking scheme for the text-to-image diffusion model mainly hinges on constructing the pairs of trigger words and verification images. These pairs are adopted to fine-tune the model on the dataset and use input trigger words and output verification images for model ownership authentication. Zhao et al. [90] established a pre-defined text-image pair as the trigger input-output and selected some uncommon words, such as "[V]," as the trigger text. The corresponding output image of the trigger word is a QR code containing the copyright information. Without requiring the original training data and internal details of the diffusion model, Yuan et al. [91] crafted a pre-defined prompt and watermark to fine-tune the pre-trained diffusion model, ultimately achieving the authentication of model copyright. Ma et al. [92] fine-tuned the replication of the image decoder, where the original image decoder is used for normal image generation. Once a trigger word is input, a validation image can be generated by combining the finetuned diffusion model with the fine-tuned image decoder. To allow commonly used words to function as trigger words, Liu et al. [93] placed a set of commonly used words at a fixed position of the input prompt and matched this prompt with the validation image. What's more, the constructed prompts were used to fine-tune the diffusion model to obtain a watermarked version. Peng et al. [94] learned the watermark diffusion process by training or fine-tuning the diffusion model and used its shared reverse noise for sampling to extract embedded watermarks without compromising the performance of the original generation task. In order to resist forgery attacks, Yuan et al. [95] selected the hash functions and keys to irreversibly generate trigger prompts, ensuring that attackers could not reverse the construction of the specific prompts through internal associations, effectively safeguarding the copyright of diffusion models. The above schemes are primarily aimed at the copyright protection for model owners and are not suitable for application scenarios that require tracing a large number of model users and their generated images.

4.2. Watermarking of AI-Generated images

In the context of distributing and deploying the text-to-image diffusion model, concerns have arisen regarding the potential misuse of generative models, particularly due to the high realism of the generated images. Accordingly, it is necessary to use traceability watermarking methods to supervise both the models and the generated images. Generation image traceability watermarking refers to incorporating watermarks into the process of diffusion models from the texts and images. Its mechanism mainly involves injecting watermarks into the generated images by fine-tuning the structure of the model or adding watermark information in crucial stages of model generation. This traceability watermarking mechanism contains specific identifiers on the basis of model generation, providing traceable features for the generated content. Then the watermarks can not only maintain the intellectual property of diffusion models but also trace the responsible parties of the generated images [96,97]. According to the different embedding positions of watermarks, diffusion traceability watermarking can be divided into two categories: fine-tuning-based watermarking and initial-noise-modification-based watermarking. As presented in Figure 3, the watermark can be injected into various positions of the generation process.

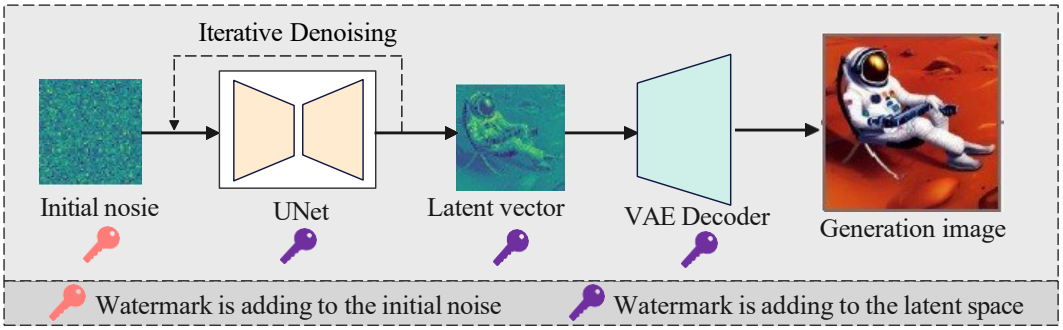

Figure 3. Illustration of different watermark placement with the Stable Diffusion model, including adding to the latent space and the initial noise.

4.2.1. Fine-Tuning-Based Watermarking

Fine-tuning-based watermarking involves embedding watermarks by fine-tuning part of the diffusion model, such as VAE or UNet [98], or fine-tuning an extra pre-trained watermark embedder, aiming to embed watermarks while generating the image, and the embedded watermark could be accurately extracted or detected by the watermark extractor, which is trained or fine-tuned together with the watermark embedder.

Fine-tuning the decoder of VAE. The schemes [99–102] involve watermark embedding from the perspective of fine-tuning the decoder of VAE, enabling the VAE decoder to simultaneously embed watermarks and decode latent vectors into images. Based on a pre-trained watermark embedding encoder-decoder framework, Fernandez et al. [99] fine-tuned the pre-trained watermark decoder together with the image decoder of the diffusion model to generate images containing specific binary sequence watermarks, thereby achieving copyright protection of the model and user traceability. But this method could only embed a fixed watermark. If the user needs to modify the watermark, the image decoder of the diffusion model must be fine-tuned again. Xiong et al. [100] first designed an information encoder to convert the watermark information into an information matrix and then embedded this information matrix into the intermediate output of the image decoder to obtain a watermarked image. In order to improve the security, it is necessary to dynamically adjust the polarity of the loss value to control the use of the information matrix. If the user does not use the information matrix, the quality of the generated image will significantly decrease. Kim et al. [101] proposed a mapping network to convert watermark information into intermediate fingerprint images within the diffusion model dimension and weight-modulated them with a decoder to embed watermark information. This method only required one forward transmission process for different watermark information, greatly saving computational cost. Ci et al. [102] employed a pre-trained watermark encoder to perform secondary processing on the intermediate output of the image decoder to embed watermark information, ultimately obtaining a watermarked image that can be extracted by the pre-trained watermark decoder for copyright authentication and user traceability.

Fine-tuning the extra watermark encoder-decoder. Methods [103–105] embed watermarks into the latent vector before VAE. Bui et al. [103] put forward a lightweight secret information encoder to map secret information into the latent space and embedded the secret information by making small offsets to the latent space. Due to the use of a pre-trained autoencoder as the base model, there is no need to learn image distribution, resulting in a simple training process and good performance in embedding and extracting secret information. Considering the dilemma that the watermark embedding and extracting schemes in pixel space cannot balance image quality and watermark robustness, Meng et al. [104] chose to embed and detect in latent space and proposed a progressive training strategy that could effectively resist different watermark attacks, maintaining the quality of the generated image containing watermarks. Zhang et al. [105] designed a plug-and-play watermarking framework that embedded watermarks without modifying the components of the diffusion model while using the

watermark embedding strength factor to balance the contradiction between watermarked images and watermark extraction quality. The watermark scheme performed good generalization and could be transferred to different versions of diffusion models.

Fine-tuning the U-Net. Methods [106,107] embed the watermark into Unet space. Min et al. [106] fine-tuned the first layer network of the diffusion model based on the U-Net structure and embedded watermarks within a certain number of denoising iterations. Specifically, the watermark information is first converted into an intermediate output containing watermark information through a pre-trained linear layer. Then in the final few sampling steps of image generation, this intermediate output was input into a fine-tuned diffusion model to generate a watermarked image. In the white-box scene, Feng et al. [107] first trained a set of encoder and decoder modules of watermark information in the latent space and added the watermark into the generated image with a watermark low-rank adaptation module. Furthermore, prior preserving fine-tuning was designed to protect the original generation performance.

Most fine-tuning-based methods do not alter the layout of the generated image. However, when the model undergoes fine-tuning or compression by an attacker, the embedded watermark can be easily removed.

4.2.2. Initial-Noise-Modification-Based Watermarking

The denoising sampling process based on DDIM is approximately reversible when the random seed is fixed. Therefore, the approximate reversibility of DDIM can be used to achieve more robust watermark embedding and extraction. The method based on this characteristic embeds the watermark pattern into the initial noise of the diffusion model and then detects the presence of the watermark from the inversed initial noise reconstructed using the DDIM inversion process from the watermarked generated image.

Robustness-enhanced methods. Wen et al. [108] proposed a diffusion watermarking scheme, namely Tree-Ring, which embedded watermarks in the frequency domain of initial noise, and the watermark embedding area is circular, which enhances the robustness of watermarks against rotation attacks. The L1 distance between the extracted watermark information and the embedded one is calculated and compared with a pre-set threshold to achieve traceability. However, this cannot accomplish identifying various keys. Ci et al. [109] extended the Tree Ring scheme to multiple-keys watermarking, achieving stronger watermark extraction robustness through discretization and lossless watermark embedding strategies, which could effectively resist rotation attacks. Arabi et al. [110] embedded watermark information into the initial latent vector using generated Fourier patterns and designed a two-stage detection framework for extracting watermark information, which could defend against forgery attacks and removal attacks.

Fidelity-enhanced methods. The above methods modified the distribution of the initial noise, hence the layout of the generated images will be modified. To address this problem, some researchers focus on the lossless watermarking based on the reversibility of the diffusion process, which could be achieved by distributing and maintaining watermark embedding. Yang et al. [111] proposed a lossless watermarking method, which first encrypted the watermark message to uniform distribution by a cryptography algorithm and then utilized the inverse transform sampling method to transform the watermark from uniform distribution to a Gaussian distribution, thereby reducing the impact on the generated image. Many deep-learning-based watermarking schemes have been proposed based on noise encoder and decoder frameworks. The method studied by Lei et al. [112] was mainly driven by watermark information, using an encoder to directly generate initial noise containing watermarks and then using DDIM-based forward and backward processes combined with an information decoder to generate watermarked images and extract watermark information. Yu et al. [113] considered a diffusion model as hiding secret images, which combined approximately the same initial noise with a different prompt and constructed an image-hiding method using public and private keys to get

a watermarked generation image. Zhang et al. [114] followed the embedding method of Tree-Ring but performed DDIM inversion on the existing image and then added the watermark to the initial noise vector after inversion. And an image enhancement module was introduced after generating the watermarked image to improve the visual quality. This method needs to optimize the initial Gaussian noise of each image to ensure that the denoised results closely match the original results.

Embedding the watermark into the initial noise can achieve high robustness, and by maintaining the Gaussian distribution of the initial noise, the impact on the quality of the generated image is minimized. However, it is important to note that watermarking schemes based on the approximate invertibility of DDIM require the use of a consistent diffusion model for both watermark embedding and extraction. This means that the extraction process necessitates the use of the diffusion model. In practical applications, using the diffusion model itself as the watermark extractor is neither convenient for storage nor for deployment.

5. Performance Evaluation

It is crucial to comprehensively evaluate watermarking schemes from different perspectives. This paper mainly considers four key aspects: watermark capacity, watermark detection accuracy, fidelity, and robustness. The watermark capacity refers to the effective watermark payload contained in the watermark algorithm. Watermark detection accuracy measures the performance of watermark extraction. Fidelity assesses whether the original performance of the watermarked model will decrease after adding the watermarks. Robustness evaluates whether watermarked models or images can resist malicious or non-malicious attacks.

5.1. Watermark Capacity

According to the exact content of the embedded watermark information, watermarking algorithms can be divided into zero-bit watermarking and multi-bit watermarking. In the case of zero-bit watermarking, the watermark detection algorithm is used solely to analyze whether there exists a watermark in the carrier. A multi-bit watermarking corresponds to an n-bit string, which needs to determine not only whether the detected carrier contains a watermark but also identify the content of the retrieved string. Zero-bit watermarking has higher robustness due to its simple task and is often used for copyright protection. While multi-bit watermarking has greater flexibility and can be used for fingerprint recognition and content traceability.

5.2. Watermark Detection Accuracy

In zero-bit watermarking schemes, watermark detection is essentially a binary classification problem. Therefore, its detection success rate is mainly evaluated by classification metrics such as True Positive Rate (TPR) and False Positive Rate (FPR). FPR represents the probability of negative samples being mistaken for positive ones, while TPR refers to the proportion of actual positive samples correctly predicted as positive ones. Overall, for the practical application of proposed watermarking algorithms, high TPR under low FPR is the goal that watermarking algorithms need to pursue. True Positive Rate at 0.01 False Positive Rate [115], which is a typical indicator in the generation image traceability watermarking methods.

In multi-bit watermarking schemes, it is necessary not only to detect the presence of watermarks but also to ensure the integrity of watermark information extraction. Information extraction integrity is accomplished through bit error rate (BER) [116] or bit accuracy [117]. Through the detection process, BER denotes the probability of each bit being incorrect, while bit accuracy represents the probability of each bit being correct. Increasing the effective payload of the watermark often causes a shift in the distribution of the training data, thereby reducing the quality of the generated image. However, this can be alleviated by increasing the resolution of the generated image.

5.3. Fidelity

When embedding watermarks into the generative model, it often leads to changes in the model structure, which in turn affects its performance, including the model and the generated image quality. For the model, the CLIP [118] metric evaluates the correlation between watermarked images generated by the watermarked model and text prompts. The lower the CLIP value, the less impact of the watermark embedding on the semantics of the generation image for the generation model. The generation image quality mainly evaluates the quality of watermarked images, indicating the degree of impact on the image after watermark embedding. Mainly through two aspects, including reference indicators and non-reference indicators. Reference indicators include the Peak Signal to Noise Ratio (PSNR) [119], the Structural Similarity (SSIM) [120], and the Fre'chet Inception Distance (FID) [121]. No-reference indicators include the Natural Image Quality Assessment Score (NIQE) [122], and the Perceived Image Quality Assessment Score (PIQE) [123]. The quality of watermarked images improves with increasing SSIM and PSNR and decreasing FID, NIQE, and PIQE.

5.4. Robustness

5.4.1. Robustness Against Image Processing

Robustness means that the ability of watermarked images can still detect and extract the watermark information after various non-malicious or malicious image processing operations. Non-malicious image precessing operation refers to the inevitable post-processing of an image, including compression, filtering, or printing. For example, when an image is uploaded to a social media platform, it may undergo the compression operation during transmission, potentially degrading the hidden watermark. In contrast, malicious image precessing operation means the transmission of images to malicious tampering individuals with ulterior motives, with the aim of destroying and erasing the watermark signals, and even altering the identity of copyright owners. For non-malicious attacks, attack technologies typically add random perturbations, JPEG compression, random cropping, scaling, rotation, blurring, color enhancement, and other attack methods and evaluate the accuracy of watermark recovery, thereby determining whether the watermarking scheme is robust.

In the research field of AIGC watermarking, the focus should be more on the scheme's resistance to malicious attacks, especially against watermark removal attacks and watermark forgery attacks. In a white-box scenario, attackers can access the decoder of the watermark extractor, but cannot obtain the true watermark or encoder of the watermark extractor. Jiang et al. [124] proposed the WEvade method, which added the adversarial perturbation to the generated image to make the assumed watermark output by the decoder different from the original watermark, thereby achieving the watermark forgery. Li et al. [125] used pre-trained diffusion models for content processing and generated GANs for watermark removal or watermark forgery and only needed to be watermarked-generated images to remove watermarks under black-box access.

5.4.2. Robustness Against Model Processing

After obtaining a watermarked model in public, the attacker's goal is to apply various attack technologies to invalidate the verification process of the watermark, and the attack does not degrade the model's performance while minimizing the computational cost. Model processing attacks involve modifications to the model's parameters or structure in order to erase the watermark, with common methods including model fine-tuning, model compression, and functionality-equivalent attack. **Model fine-tuning**: the attacker adopts a small number of samples to fine-tune the watermarked model, attempting to remove the watermark while maintaining the model's overall performance [126]. Four kinds of commonly used fine-tuning methods consist of Fine-Tuning Last Layer (FTLL), Fine-Tuning All Layers (FTAL), Re-Training Last Layer (RTLL), and Re-Training All Layers (RTAL). **Model compression**: weight pruning and weight quantization are commonly used techniques to compress the model size, facilitating efficient inference and deployment in the resource-constrained scenario. Weight

pruning is to set the parameters with smaller absolute values to zero, as these parameters with smaller absolute values have less impact on the performance of the model performance [127]. Weight quantization refers to representing model parameters in a low precision format (e.g., 8-bit integers or lower) to reduce storage requirements [128]. **Functionality-equivalent attack**: this is an attack targeting the white-box model watermarking, which can achieve watermark removal without compromising model performance by adjusting model parameters or structure appropriately. This is straightforward and efficient, requiring no access to the dataset, no additional training of the model, and no prior knowledge about watermarking algorithms [129]. The DNN model watermarking scheme always considers these model processing attacks during the design angle, but the robustness of the AIGC watermarking scheme to these attacks is not adequately addressed.

5.4.3. Robustness Against Adversarial Attacks

Adversarial attacks are always malicious in nature and aim to disrupt the effectiveness of the copyright verification process in the watermarked models. **Evasion attack**: this type of attack targets watermarking schemes based on trigger-based backdoor models. Due to the different distribution of trigger samples compared to normal test samples, the attacker can exploit this characteristic to detect trigger samples and escape the watermark verification process [130]. **Model extraction attack**: it refers to the process of obtaining a new functional approximation model by simulating the input-output mapping of the victim model [131], including multiple model extraction methods, such as retraining, cross-structure retraining, distillation, etc. While these attacks require a certain amount of computing resources and data, making the cost higher, they can effectively remove various types of watermarks from the model. **Forgery attack**: also known as ambiguity attack, it refers to an attacker forging a new watermark without altering the model, thus causing ambiguity during model ownership verification [132]. The model owner may struggle to verify the copyright when the watermark is not unique, making the verification process unclear. **Collusion attack**: multiple attackers, each with the same host neural network but distinct implanted fingerprints, collaborate to generate an unmarked model [133]. In order to better adapt to real-world application scenarios, more consideration on the robustness against the adversarial attacks, including but not limited to the adversarial attacks mentioned above, is crucial and needed for the AIGC watermarking algorithms.

Table 1 presents a comparison of AIGC watermarking methods in three aspects, including watermark capacity, fidelity, and robustness, wihch is able to offer a referential idea to think about the performance of the designed AIGC watermarking schemes.

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 1, 2025

Table 1. A qualitative comparison of some AIGC watermarking algoritms in performace indictors.

Types	Methods	Watermark capacity	Fidelity		Robustness	
			generation images	generation model	non-malicious operation	malicious operation
Adding to the latent sapce	Fernandez et al. [99]	n	C	X	C	C
	Xiong et al. [100]	n	C	X	C	C
	Kim et al. [101]	0	C	C	C	C
	Ci et al. [102]	n	C	X	C	C
	Bui et al. [103]	n	C	X	C	X
	Meng et al. [104]	n	C	X	C	C
	Zhang et al. [105]	n	C	X	C	X
	Min et al. [106]	n	C	X	C	C
	Feng et al. [107]	n	C	X	C	C
Adding to the initial noise	Wen et al. [108]	0	C	C	C	X
	Ci et al. [109]	0	C	C	C	X
	Arabi et al. [110]	0	C	X	C	C
	[illegible] [111]	n	C	C	C	C
	Lei et al. [112]	n	C	C	C	C
	Yu et al. [113]	n	C	X	C	C
	Zhang et al. [114]	0	C	X	C	X

* 0 means zero-bit watermarking, and n means multi-bits watermarking. C denotes this metric is considered and X denotes this is not considered.

6. Discussion

Currently, researchers have presented several AIGC watermarking algorithms, greatly advancing the development of copyright protection and user traceability technologies for image generation models. Nonetheless, several challenges remain in the watermarking for generative models.

- **Trade-off between watermark capacity, fidelity, and robustness.** Watermarking in the latent domain, either for modifying the initial noise or the latent vector, offers enhanced robustness and has a smaller effect on the generated content's quality. However, due to the typically lower dimensionality of the latent representations, the amount of watermark information that can be embedded is considerably smaller. In contrast, the pixel domain allows for the embedding of more watermark information, but embedding watermarks in the pixel domain can degrade the quality of the generated image, and its robustness is generally lower. Future research needs to focus on how to better balance these competing factors.
- **Robustness of watermarks against neural network-based attacks.** Most existing AIGC watermarking methods evaluate the robustness performance limited to image post-processing attacks but do not have strong defense ability against the attacks of generative models. More watermark attack patterns based on neural networks should be considered in the design process of watermarking technologies to ensure greater robustness.

- **Joint generative scalable watermarking.** Thinking that the current generation model has a large number of parameters that are difficult to train and fine-tune, the research about watermarking schemes can focus on lightweight additional modules. By training watermark encoder-decoder models with low computational overhead and strong scalability, watermarking schemes can be incorporated into any image generation models.

7. Conclusions

Recent years have witnessed the rapid growth of AIGC technology, which accelerated its applicability across a wide range of fields, driven by the evolution of GANs, Transformers, VAEs, and diffusion models. As AIGC continues to grow, both diverse opportunities and difficulties have emerged currently, with model ownership protection and generation image traceability becoming prominent research areas. In this context, watermarking is considered a key and effective solution strategy for addressing these issues. Consequently, this paper starts with digital image watermarking, introduces DNN authentication watermarking and diffusion model authentication watermarking, and lists advanced AIGC watermarking technology from two perspectives of scenario applications. Finally, the crucial performance evaluation metrics for watermarking algorithms and some potential research prospects of AIGC watermarking are provided.

Author Contributions: Conceptualization, H.L., L.L. and J.L.; investigation, H.L. and L.L.; resources, H.L. and L.L.; writing—original draft preparation, H.L.; writing—review and editing, H.L., L.L. and J.L.; visualization, H.L. and L.L.; supervision, L.L. and J.L.; funding acquisition, L.L. and J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported in part by the National Natural Science Foundation of China under Grant 62302286.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DCGAN	Convolutional Generative Adversarial Networks
GAN-INT-CLS	Matching-aware discriminator & Learning with manifold interpolation
AttenGAN	Attentional Generative Adversarial Networks
VQGAN	Vector Quantised Generative Adversarial Networks
TransGAN	Transformer-based Generative Adversarial Networks
VitGAN	Vision Transformer-based Generative Adversarial Networks
DiTs	Diffusion Transformers
IntroVAE	Introspective Variational Autoencoder
DeNoL	Decoupling Noise Layer

References

1. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M. A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; Lample, G. LLaMA: open and efficient foundation language models. *arXiv* **2023**, arXiv:2302.13971.
2. Thirunavukarasu, A.J.; Ting, D.S.J.; Elangovan, K.; Gutierrez, L.; Tan, T.F.; Ting, D.S.W. Large language models in medicine. *Nat. Med.* **2023**, *29*, 1930-1940.
3. Xiao, S.T.; Wang, Y.Z.; Zhou, J.J.; Yuan, H.Y.; Xing, X.R.; Yan, R.R.; Wang, S.T.; Huang, T.J.; Liu, Z. Omnigen: unified image generation. *arXiv* **2024**, arXiv:2409.11340.
4. Zhang, H.L.; Edelman, B.L.; Francati, D.; Venturi, D.; Ateniese, G.; Barak, B. Watermarks in the sand: impossibility of strong watermarking for generative models. *arXiv* **2023**, arXiv:2311.04378.

5. Hu, Y.P.; Jiang, Z.Y.; Guo, M.Y.; Gong, N.Z.Q. Stable signature is unstable: removing image watermark from diffusion models. *arXiv* **2024**, arXiv:2405.07145.
6. Zhang, X.Y.; Xu, Y.M.; Li, R.Y.; Yu, J.W.; Li, W.Q.; Xu, Z.P.; Zhang, J. V2a-mark: versatile deep visual-audio watermarking for manipulation localization and copyright protection. *arXiv* **2024**, arXiv:2404.16824.
7. Hosny, K.M.; Magdi, A.; ElKomy, O.; Hamza, H.M. Digital image watermarking using deep learning: a survey. *Comput. Sci. Rev.* **2024**, *53*, 100662.
8. Kui, R.; Yang, Z.P.; Lu, L.; Liu, J.; Li, Y.M.; Wan, J.; Zhao, X.D.; Feng, X.H.; Shao, S. SoK: on the role and future of AIGC watermarking in the era of Gen-AI. *arXiv* **2024**, arXiv:2411.11478.
9. Pang, Q.; Hu, S.Y.; Zheng, W.T.; Smith, V. Attacking LLM watermarks by exploiting their strengths. *arXiv* **2024**, arXiv:2402.16187.
10. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, Canada, 8-13 December 2014; pp. 2672-2680.
11. Radford, A.; Metz, L.; Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv* **2015**, arXiv:1511.06434.
12. Reed, S.; Akata, Z.; Yan, X.C.; Logeswaran, L.; Schiele, B.; Lee, H. Generative adversarial text to image synthesis. In Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA, 19-24 June 2016; pp. 1060-1069.
13. Xu, T.; Zhang, P.C.; Huang, Q.Y.; Zhang, H.; Gan, Z.; Huang, X.L.; He, X.D. AttnGAN: fine-grained text to image generation with attentional generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18-23 June 2018; pp. 1316-1324.
14. Karras, T.; Aila, T.; Laine, S. Progressive growing of GANs for improved quality 'stability' and variation. *arXiv* **2017**, arXiv:1710.10196.
15. Karras, T.; Laine, S.; Aila, T. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15-20 June 2019; pp. 4401-4410.
16. Kang, M.; Shin, J.; Park J. StudioGAN: a taxonomy and benchmark of GANs for image synthesis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 15725-15742.
17. Qiao, T.T.; Zhang, J.; Xu, D.Q.; Tao D.C. MirrorGAN: learning text-to-image generation by redescription. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15-20 June 2019; pp. 1505-1514.
18. Pan, X.G.; Tewari, A.; Leimkuhler, T.; Liu, L.J.; Meka, A.; Theobalt, C. Drag your GAN: interactive point-based manipulation on the generative image manifold. In Proceedings of the ACM SIGGRAPH 2023 Conference, Los Angeles, CA, USA, 6-10 August 2023; pp. 1-11.
19. Tao, M.; Bao, B.K.; Tang, H.; Xu C.S. GALIP: generative adversarial CLIPs for text-to-image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17-24 June 2023; pp. 14214-14223.
20. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, 4-9 December 2017; pp. 6000-6010.
21. Esser, P.; Rombach, R.; Ommer, B. Taming transformers for high-resolution image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20-25 June 2021; pp. 12868–12878.
22. Jiang, Y.F.; Chang, S.Y.; Wang, Z.Y. TransGAN: two pure transformers can make one strong gan, and that can scale up. In Proceedings of the 35th International Conference on Neural Information Processing Systems, Virtual, 2021; pp. 14745–14758.
23. Lee, K.J.; Chang, H.W.; Jiang, L.; Zhang, H.; Tu, Z.W.; Liu, C. VitGAN: training gans with vision transformers. *arXiv* **2021**, arXiv:2107.04589.
24. Peebles, W.; Xie, S. Scalable diffusion models with transformers. In Proceedings of the International Conference on Computer Vision, Paris, France, 1-6 October 2023; pp. 4172–4182.
25. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.

26. Sohn, K.; Lee, H.; Yan, X. Learning structured output representation using deep conditional generative models. In Proceedings of the 28th International Conference on Neural Information Processing Systems, Montréal, Canada, 7-12 December 2015; pp. 3483-3491.
27. Mansimov, E.; Parisotto, E.; Ba, J.L. Generating images from captions with attention. *arXiv* **2015**, arXiv:1511.02793.
28. Van, D.O.A.; Vinyals, O. Neural discrete representation learning. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, 4-9 December 2017; pp. 6309-6318.
29. Huang, H.B.; Li, Z.H.; He, R.; Sun, Z.N.; Tan, T.N. IntroVAE: introspective variational autoencoders for photographic image synthesis. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montréal, Canada, 3-8 December 2018; pp. 52-63.
30. Daniel, T.; Tamar A. Soft-IntroVAE: analyzing and improving the introspective variational autoencoder. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20-25 June 2021; pp. 4391-4400.
31. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. In Proceedings of the 34th Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 6-12 December 2020; pp. 6840-6851.
32. Nichol, A.Q.; Dhariwal, P. Improved denoising diffusion probabilistic models. In Proceedings of the 38th International Conference on Machine Learning, Virtual, 2021; pp. 8162-8171.
33. Song, J.; Meng, C.; Ermon, S. Denoising diffusion implicit models. *arXiv* **2015**, arXiv:2010.02502.
34. Dhariwal, P.; Nichol, A. Diffusion models beat GANs on image synthesis. In Proceedings of the 34th Conference on Neural Information Processing Systems, Virtual, 2021; pp. 8780-8794.
35. Ho, J.; Salimans, T. Classifier-free diffusion guidance. *arXiv* **2022**, arXiv:2207.12598.
36. Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; Chen M. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, 17-23 July 2022; pp. 16784-16804.
37. Saharia, C.; Chan, W.; Saxena, S.; Li, L.L.; Whang, J.; Denton, E.; Ghasemipour, S.K.S.; Ayan, B.K.; Mahdavi, S.S.; Lopes, R.G.; Salimans, T.; Ho, J.; Fleet, D.J.; Norouzi, M. Photorealistic text-to-image diffusion models with deep language understanding. In Proceedings of the 36th Conference on Neural Information Processing Systems, New Orleans, LA, USA, 28 November-9 December 2022; pp. 36479-36494.
38. Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; Chen, M. Hierarchical text-conditional image generation with CLIP latents. *arXiv* **2022**, arXiv: 22204.06125.
39. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18-24 June 2022; pp. 10674-10685.
40. Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.W.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; Schramowski, P.; Kundurthy, S.R.; Crowson, K.; Schmidt, L.; Kaczmarczyk, R.; Jitsev, J. LAION-5B: an open large-scale dataset for training next generation image-text models. In Proceedings of the 36th Conference on Neural Information Processing Systems, New Orleans, LA, USA, 28 November-9 December 2022; pp. 25278-25294.
41. Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; Rombach, R. Sdxl: improving latent diffusion models for high-resolution image synthesis. *arXiv* **2023**, arXiv: 2307.01952.
42. Arya, A.; Soni, S. Performance evaluation of secrete image steganography techniques using least significant bit (LSB) method. *Int. J. Comput. Trends technol.* **2018**, *6(2)*, 160-165.
43. Singh, D.; Singh, S.K. DWT-SVD and DCT based robust and blind watermarking scheme for copyright protection. *Multimed. Tools Appl.* **2017**, *76(11)*, 13001-13024.
44. Li, X.L.; Yang, B.; Cheng, D.F.; Zeng, T.Y. A generalization of LSB matching. *IEEE Signal Process. Lett.* **2009**, *16(2)*, 69-72.
45. Bender, W.; Gruhl, D.; Morimoto, N.; Lu, A. Techniques for data hiding. *IBM Syst. J.* **1996**, *35(3.4)*, 313-336.
46. Kahlessenane, F.; Khaldi, A.; Kafi, M.R.; Eushci, S. A color value differentiation scheme for blind digital image watermarking. *Multimed. Tools Appl.* **2021**, *80(13)*, 19827-19844.
47. Zhang, X.T.; Su, Q.T.; Sun, Y.H.; Chen, S.Y. A robust and high-efficiency blind watermarking method for color images in the spatial domain. *Multimed. Tools Appl.* **2023**, *82(18)*, 27217-27243.

48. Liu, S.; Pan, Z.; Song, H. Digital image watermarking method based on DCT and fractal encoding. *IET Image Process.* **2017**, *11(10)*, 815-821.
49. Kashyap, N.; Sinha, G.R. Image watermarking using 3-level discrete wavelet transform (DWT). *Int. J. Mod. Educ. Comput. Sci.* **2012**, *4(3)*, 50.
50. Kumar, M.; Rewani, R. Digital image watermarking using fractional Fourier transform via image compression. In Proceedings of the 2013 IEEE International Conference on Computational Intelligence and Computing Research, Enathi, India, 26-28 December 2013; pp. 1-4.
51. Wang, K.S.; Gao, T.G.; You, D.T.; Xu, X.J.; Kan, H.B. A secure dual-color image watermarking scheme based 2D DWT, SVD and chaotic map. *Multimed. Tools Appl.* **2022**, *81(5)*, 6159-6190.
52. Mohammed, A.O.; Hussein, H.I.; Mstafa, R.J.; Abdulazeez, A.M. A blind and robust color image watermarking scheme based on DCT and DWT domains. *Multimed. Tools Appl.* **2023**, *82(21)*, 32855-32881.
53. Naffouti, S.E.; Kricha, A.; Sakly, A. A sophisticated and provably grayscale image watermarking system using DWT-SVD domain. *Visual Comput.* **2023**, *39(9)*, 4227-4247.
54. Begum, M.; Uddin, M.S. Digital image watermarking techniques: a review. *Information* **2020**, *11(2)*, 110.
55. Haribabu, K.; Subrahmanyam, G.; Mishra, D. A robust digital image watermarking technique using autoencoder based convolutional neural networks. In Proceedings of the 2015 IEEE Workshop on Computational Intelligence: Theories, Applications and Future Directions, Kanpur, India, 14-17 December 2015; pp. 1-6.
56. Kandi, H.; Mishra, D.; Gorthi, S.R.K.S. Exploring the learning capabilities of convolutional neural networks for robust image watermarking. *Comput. Secur.* **2017**, *65*, 247-268.
57. Zhu, J.R.; Kaplan, R.; Johnson, J.; Li, F.F. HiDDeN: hiding data with deep networks. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8-14 September 2018; pp. 657-672.
58. Ahmadi, M.; Norouzi, A.; Karimi, N.; Samavi, S.; Emami, A. ReDMark: framework for residual diffusion watermarking based on deep networks. *Expert Syst. Appl.* **2020**, *146*, 113157.
59. Mellimi, S.; Rajput, V.; Ansari, I.A.; Ahn, C.W. A fast and efficient image watermarking scheme based on deep neural network. *Pattern Recognit. Lett.* **2021**, *151*, 222-228.
60. Hao, K.L.; Feng, G.R.; Zhang, X.P. Robust image watermarking based on generative adversarial network. *China Commun.* **2020**, *17(11)*, 131-140.
61. Tancik, M.; Mildenhall, B.; Ng, R. StegaStamp: invisible hyperlinks in physical photographs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13-19 June 2020; pp. 2114-2123.
62. Liu, G.Z.; Si, Y.C.; Qian, Z.X.; Zhang, X.P.; Li, S.; Peng, W.L. WRAP: watermarking approach robust against film-coating upon printed photographs. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, Canada, 29 October-3 November 2023; pp. 7274-7282.
63. Fang, H.; Chen, K.J.; Qiu, Y.P.; Liu, L.Y.; Xu, K.; Fang, C.F.; Zhang, W.M.; Chang, E. DeNoL: a few-shot-sample-based decoupling noise layer for cross-channel watermarking robustness. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, Canada, 29 October-3 November 2023; pp. 7345-7353.
64. Cortiñas-Lorenzo, B.; Pérez-González, F. Adam and the ants: on the influence of the optimization algorithm on the detectability of DNN watermarks. *Entropy* **2020**, *22(12)*, 1379.
65. Kuribayashi, M.; Tanaka, T.; Funabiki, N. DeepWatermark: embedding watermark into DNN model. In Proceedings of 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, Auckland, New Zealand, 7-10 December 2020; pp. 1340-1346.
66. Ong, D.S.; Chan, C.S.; Ng, K.W.; Fan, L.X.; Yang, Q. Protecting intellectual property of generative adversarial networks from ambiguity attacks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20-25 June 2021; pp. 3629-3638.
67. Uchida, Y.; Nagai, Y.; Sakazawa, S.; Satoh, S. Embedding watermarks into deep neural networks. In Proceedings of 2017 ACM on International Conference on Multimedia Retrieval, Bucharest, Romania, 6-9 June 2017; pp. 269-277.
68. Rouhani, B.D.; Chen, H.L.; Koushanfar, F. DeepSigns: an end-to-end watermarking framework for ownership protection of deep neural networks. In Proceedings of the 24th International Conference on Architectural Support for Programming Languages and Operating Systems, Providence, RI, USA, 13-17 April 2019; pp. 485-497.
69. Chen, H.L.; Rouhani, B.D.; Fu, C.; Zhao, J.S.; Koushanfar, F. DeepMarks: a secure fingerprinting framework for digital rights management of deep learning models. In Proceedings of 2019 on International Conference on Multimedia Retrieval, Ottawa, Canada, 10-13 June 2019; pp. 105-113.

70. Wang, T.H.; Kerschbaum, F. RIGA: covert and robust white-box watermarking of deep neural networks. In Proceedings of 2021 Web Conference, Ljubljana, Slovenia, 19-23 April 2021; pp. 993-1004.
71. Tartaglione, E.; Grangetto, M.; Cavagnino, D.; Botta, M. Delving in the loss landscape to embed robust watermarks into neural networks. In Proceedings of the 25th International Conference on Pattern Recognition, Milan, Italy, 10-15 January 2021; pp. 1615-1631.
72. Lou, X.X.; Guo, S.W.; Li, J.W.; Zhang, T.W. Ownership verification of DNN architectures via hardware cache side channels. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32(11)*, 8078-8093.
73. Le Merrer, E.; Perez, P.; Trédan, G. Adversarial frontier stitching for remote neural network watermarking. *Neural Comput. Appl.* **2020**, *32(13)*, 9233-9244.
74. Li, L.; Zhang, W.M.; Barni, M. Universal blackmarks: key-image-free blackbox multi-bit watermarking of deep neural networks. *IEEE Signal Process. Lett.* **2023**, *30*, 36-40.
75. Namba, R.; Sakuma, J. Robust watermarking of neural network with exponential weighting. In Proceedings of 2019 ACM Asia Conference on Computer and Communications Security, Auckland, New Zealand, 9-12 July 2019; pp. 228-240.
76. Adi. Y.; Baum, C.; Cissé, M.; Pinkas, B.; Keshet, J. Turning your weakness into a strength: watermarking deep neural networks by backdooring. In Proceedings of the 27th USENIX Security Symposium, Baltimore, MD, USA, 15-17 August 2018; pp. 1615-1631.
77. Zhang, J.L.; Gu, Z.S.; Jang, J.; Wu, H.; Stoecklin, M.P.; Huang, H.Q.; Molloy, I. Protecting intellectual property of deep neural networks with watermarking. In Proceedings of 2018 on Asia Conference on Computer and Communications Security, Incheon, Republic of Korea, 4 June 2018; pp. 159-172.
78. Guo, J.; Potkonjak, M. Watermarking deep neural networks for embedded systems. In Proceedings of 2018 IEEE/ACM International Conference on Computer-Aided Design, San Diego, CA, USA, 5-8 November 2018; pp. 1-8.
79. Guo, J.; Potkonjak, M. Volutionary trigger set generation for DNN black-box watermarking. *arXiv* **2021**, arXiv: 1906.04411.
80. Sun, S.C.; Xue, M.F.; Wang, J.; Liu, W.Q. Protecting the intellectual properties of deep neural networks with an additional class and steganographic images. *arXiv* **2021**, arXiv: 2104.09203.
81. Li, Z.; Hu, C.Y.; Zhang, Y.; Guo, S.Q. How to prove your model belongs to you: a blind-watermark based framework to protect intellectual property of DNN. In Proceedings of the 35th Annual Computer Security Applications Conference, New York, NY, USA, 9-13 December 2019; pp. 126-137.
82. Zhang, Y.Q.; Jia, Y.R.; Wang, X.Y.; Niu, Q.; Chen, N.D. DeepTrigger: a watermarking scheme of deep learning models based on chaotic automatic data annotation. *IEEE Access* **2020**, *8*, 213296-213305.
83. Li, F.Q.; Wang, S.L. Persistent watermark for image classification neural networks by penetrating the autoencoder. In Proceedings of 2021 IEEE International Conference on Image Processing, Anchorage, AK, USA, 19-22 September 2021; pp. 3063-3067.
84. Zhu, R.J.; Zhang, X.P.; Shi, M.T.; Tang, Z.J. Secure neural network watermarking protocol against forging attack. *EURASIP J. Image Video Process.* **2020**, *2020*, 37.
85. Szyller, S.; Atli, B.G.; Marchal, S.; Asokan, N. DAWN: dynamic adversarial watermarking of neural networks. In Proceedings of the 29th ACM International Conference on Multimedia, Chengdu, China, 20-24 October 2021; pp. 4417-4425.
86. Charette, L.; Chu, L.Y.; Chen, Y.Z.; Pei, J.; Wang, L.J.; Zhang, Y. Cosine model watermarking against ensemble distillation. In Proceedings of the 36th AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 22 February-1 March 2022; pp. 9512-9520.
87. Li, H.Y.; Wenger, E.; Shan, S.; Zhao, B.Y.; Zheng, H.T. Piracy resistant watermarks for deep neural networks. *arXiv* **2020**, arXiv: 1901.01226.
88. Xu, X.R.; Li, Y.Q.; Yuan, C. "Identity bracelets" for deep neural networks. *IEEE Access* **2020**, *8*, 102065-102074.
89. Li, Y.M.; Bai, Y.; Jiang, Y.; Yang, Y.; Xia, S.T.; Li, B. Untargeted backdoor watermark: towards harmless and stealthy dataset copyright protection. *arXiv* **2022**, arXiv: 2210.00875.
90. Zhao, Y.Q.; Pang, T.Y.; Du, C.; Yang, X.; Cheung, N.M.; Lin, M. Recipe for watermarking diffusion models. *arXiv* **2023**, arXiv: 2303.10137.
91. Yuan, Z.H.; Li, L.; Wang, Z.C.; Zhang, X.P. Watermarking for stable diffusion models. *IEEE Internet Things J.* **2025**, *11(21)*, 35238-35249.

92. Ma, Z.Y.; Jia, G.L.; Qi, B.Q.; Zhou, B.W. Safe-SD: safe and traceable stable diffusion with text prompt trigger for invisible generative watermarking. *arXiv* **2024**, arXiv: 2407.13188.
93. Liu, Y.G.; Li, Z.; Backes, M.; Shen, Y.; Zhang, Y. Watermarking diffusion model. *arXiv* **2023**, arXiv: 2305.12502.
94. Peng, S.; Chen, Y.F.; Wang, C.; Jia, X.H. Protecting the intellectual property of diffusion models by the watermark diffusion process. *arXiv* **2023**, arXiv: 2306.03436.
95. Yuan, Z.H.; Li, L.; Wang, Z.C.; Zhang, X.P. Protecting copyright of stable diffusion models from ambiguity attacks. *Signal Process.* **2025**, *227*, 109722.
96. Guo, Z.J.; Li, M.L.; Zhou, M.Y.,; Peng, W.L.,; Li, S.; Qian, Z.X.; Zhang, X.P. Survey on digital watermarking technology for artificial intelligence generated content models. *J. Cybersecurity* **2024**, *2(1)*, 13-39.
97. Liu, G.H.; Chen, T.; Theodorou, E.A.; Tao, M. Mirror diffusion models for constrained and watermarked generation. In Proceedings of the 37th Conference on Neural Information Processing Systems, New Orleans, LA, USA, 10-16 December 2023; pp. 42898-42917.
98. Ronneberger, O.; Fischer, P.; Brox, T. U-net: convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 18–22 September 2015; pp. 234-241.
99. Fernandez, P.; Couairon, G.; Jégou, H.; Douze, M.; Furon, T. The stable signature: rooting watermarks in latent diffusion models. In Proceedings of the International Conference on Computer Vision, Paris, France, 1-6 Oct 2023; pp. 22409-22420.
100. Xiong, C.; Qin, C.; Feng, G.R.; Zhang, X.P. Flexible and secure watermarking for latent diffusion model. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, Canada, 29 October-3 November 2023; pp. 1668-1676.
101. Kim, T.; Min, K.; Patel, M.; Cheng, S.; Yang, Y.Z. WOUAF: weight modulation for user attribution and fingerprinting in text-to-image diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 16-22 June 2024; pp. 8974-8983.
102. Ci, H.; Song, Y.R.; Yang, P.; Xie, J.H.; Shou, M.Z. WMAdapter: adding waterMark control to latent diffusion models. *arXiv* **2024**, arXiv: 2406.08337.
103. Bui, T.; Agarwal, S.; Yu, N.; Collomosse, J. RoSteALS: robust steganography using autoencoder latent space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17-24 June 2023; pp. 933-942.
104. Meng, Z.L.; Peng, B.; Dong, J. Latent watermark: inject and detect watermarks in latent diffusion space. *arXiv* **2024**, arXiv: 2404.00230.
105. Zhang, G.K.; Wang, L.G.; Su, Y.T.; Liu, A.A. A training-free plug-and-play watermark framework for stable diffusion. *arXiv* **2024**, arXiv: 2404.05607.
106. Min, R.; Li, S.; Chen, H.Y.; Cheng, M.H. A watermark-conditioned diffusion model for IP protection. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September-4 October 2024; pp. 104-120.
107. Feng, W.T.; Zhou, W.B.; He, J.Y.; Zhang, J.; Wei, T.Y.; Li, G.L.; Zhang, T.W.; Zhang, W.M.; Yu, N.H. AquaLoRA: toward white-box protection for customized stable diffusion models via watermark LoRA. *arXiv* **2024**, arXiv: 2405.11135.
108. Wen, Y.X.; Kirchenbauer, J.; Geiping, J.; Goldstein, T. Tree-ring watermarks: fingerprints for diffusion images that are invisible and robust. *arXiv* **2023**, arXiv: 2305.20030.
109. Ci, H.; Yang, P.; Song, Y.R.; Shou, M.Z. Ringid: rethinking tree-ring watermarking for enhanced multi-key identification. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September-4 October 2024; pp. 338-354.
110. Arabi, K.; Feuer, B.; Witter, T.; Hegde, C.; Cohen, N. Hidden in the noise: two-stage robust watermarking for images. *arXiv* **2024**, arXiv: 2412.04653.
111. Yang, Z.J.; Zeng, K.; Chen, K.J.; Fang, H.; Zhang, W.M.; Yu, N.H. Gaussian shading: provable performance-lossless image watermarking for diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16-22 June 2024; pp. 12162-12171.
112. Lei, L.Q.; Gai, K.K.; Yu, J.; Zhu, L.H. DiffuseTrace: a transparent and flexible watermarking scheme for latent diffusion model. *arXiv* **2024**, arXiv: 2405.02696.
113. Yu, J.W.; Zhang, X.Y.; Xu, Y.M.; Zhang, J. CRoSS: diffusion model makes controllable, robust and secure image steganography. In Proceedings of the 37th Conference on Neural Information Processing Systems, New Orleans, LA, USA, 10-16 December 2023; pp. 80730-80743.

114. Zhang, L.J.; Liu, X.; Martin, A.V.; Bearfield, C.X.; Brun, Y.; Guan, H. Robust image watermarking using stable diffusion. *arXiv* **2024**, arXiv: 2401.04247.
115. Zhao, X.D.; Zhang, K.X.; Wang, Y.X.; Li, L. Generative autoencoders as watermark attackers: analyses of vulnerabilities and threats. *arXiv* **2023**, arXiv: 2306.01953.
116. Feng, W.T.; He, J.Y.; Zhang, J.; Zhang, T.W.; Zhou, W.B; Zhang, W.M.; Yu, N.H. Catch You everything everywhere: guarding textual inversion via concept watermarking. *arXiv* **2023**, arXiv: 2309.05940.
117. Lei, L.Q.; Gai, K.K.; Yu, J.; Zhu, L.H.; Wu, Q. Conceptwm: a diffusion model watermark for concept protection. *arXiv* **2024**, arXiv: 2411.11688.
118. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; Sutskever, I. Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning, Long Beach, CA, USA, 18-24 July 2021; pp. 8878-8763.
119. Setiadi, D.R.I.M. PSNR vs SSIM: imperceptibility quality assessment for image steganography. *Multimed. Tools Appl.* **2021**, *80(6)*, 8423-8444.
120. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13(4)*, 600-612.
121. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Proceedings of the 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4-9 December 2017; pp. 6626-6637.
122. Mittal, A.; Soundararajan, R.; Bovik, A.C. Making a "completely blind" image quality analyzer. *IEEE Signal Process. Lett.* **2013**, *20*, 209-212.
123. Venkatanath, N.; Praneeth, D.; Maruthi Chandrasekhar, B.H.; Sumohana, S.C.; Swarup S, M. Blind image quality evaluation using perception based features. In Proceedings of the 2015 Twenty First National Conference on Communications, Mumbai, India, 27 February-1 March 2015; pp. 1-6.
124. Jiang, Z.Y.; Zhang, J.H.; Gong, N.Z. Evading watermark based detection of AI-generated content. In Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, Copenhagen, Denmark, 26-30 November 2023; pp. 1168-1181.
125. Li, G.L.; Chen, Y.F.; Zhang, J.; Li, J.W.; Guo, S.W.; Zhang, T.W. Towards the vulnerability of watermarking artificial intelligence generated content. *arXiv* **2023**, arXiv: 2310.07726.
126. Chen, X.Y.; Wang, W.X.; Ding, Y.M.; Bender, C.; Jia, R.; Li, B.; Song, D. Leveraging unlabeled data for watermark removal of deep neural networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, USA, 9-15 June 2019; pp. 1-6.
127. Blalock, D.; Gonzalez Ortiz, J.J.; Frankle, J.; Guttag, J. What is the state of neural network pruning? In Proceedings of Machine Learning and Systems, Austin, TX, USA, 2-4 March 2020; pp. 129-146.
128. Liu, G.Y.; Xu, T.L.; Ma, X.Q.; Wang, C. Your model trains on my data? Protecting intellectual property of training data via membership fingerprint authentication. *IEEE Trans. Inf. Forensics Secur.* **2022**, *17*, 1024-1037.
129. Lv, P.Z.; Li, P.; Zhang, S.Z.; Chen, K.; Liang, R.G.; Ma, H.L. A robustness-assured white-box watermark in neural networks. *IEEE Trans. Dependable Secur. Comput.* **2023**, *20(6)*, 5214-5229.
130. Hitaj, D.; Mancini, L.V. Have you stolen my model? evasion attacks against deep neural network watermarking techniques. *arXiv* **2018**, arXiv: 1809.00615.
131. Lukas, N.; Jiang, E.; Li, X.D.; Kerschbaum, F. SoK: how robust is image classification deep neural network watermarking? In Proceedings of the 2022 IEEE Symposium on Security and Privacy, San Francisco, CA, USA, 22-26 May 2022; pp. 787-804.
132. Yuan, Z.H.; Li, L.; Wang, Z.C.; Zhang, X.P. Ambiguity attack against text-to-image diffusion model watermarking. *Signal Process.* **2024**, *221*, 109509.
133. Luo, Y.L.; Li, Y.Z.; Qin, S.; Fu, Q.; Liu, J.X. Copyright protection framework for federated learning models against collusion attacks. *Inf. Sci.* **2024**, *680*, 121161.