A TAXONOMY OF WATERMARKING METHODS FOR AI-GENERATED CONTENT

Pierre Fernandez, Hady Elsahar, Tomáš Souček, Sylvestre-Alvise Rebuffi,
Valeriu Lacatusu, Tuan Tran, Alexandre Mourachko
FAIR, Meta
Correspondance to: pfz@meta.com

ABSTRACT

As AI-generated content features more prominently in our lives, it becomes important to develop methods for tracing their origin. Watermarking is a promising approach, but a clear categorization of existing techniques is lacking. We propose a simple taxonomy of watermarking methods for generative AI based on where they are applied in the deployment of the models: (1) *post-hoc watermarking*, adding watermarks after content generation; (2) *out-of-model watermarking*, embedding watermarks during generation without modifying the model; (3) *in-model watermarking*, integrating watermarks directly into the model's parameters. By providing a structured overview of existing techniques across image, audio, and text domains, this taxonomy aims to help researchers, policymakers, and regulators make informed decisions about which approach best fits their needs, acknowledging that no single method is universally superior and that different approaches may be suited to specific use cases and requirements.

1 INTRODUCTION

The internet has become a playground for generative AI models, which are developed and adopted at an unprecedented pace. In text generation, ChatGPT reached 100 million users in just two months, and has had a significant impact on the way people interact with AI since then. It has counterparts in image generation with DALL·E (Ramesh et al., 2021) and Midjourney; video with Sora and Kling; and audio with Suno and Udio. Closed models are generally followed by open-weights alternatives of the same quality in less than a year, with *e.g.*, Llama (Touvron et al., 2023) or Stable Diffusion (Rombach et al., 2022).

The use of generative models to create online content has increased rapidly in recent years, increasing volumes of content produced overall and, in some instances, replacing traditional content creation methods. By late 2023, they had already produced as many images as photographers had taken in 150 years of photography (Valyaeva, 2023). The emergence of generative models, and their outputs has prompted concern from parts of the media and creative industries, as well as governments due to concerns about the potential for misinformation, fraud, and impersonation. There have been instances of generative models allegedly being used to create scam books sold on Amazon (Knibbs, 2024) to influence campaigns (Goldstein et al., 2023) or to create deepfakes and impersonate public figures (Harris, 2018; Shen et al., 2019). These risks linked to AI-generated content are heightened because it is difficult to detect and attribute such content to models that generated them. A study by Nightingale & Farid (2022) notably found that human participants exhibited a relatively

Figure 1: Methods for tracing AI-generated content and when they intervene in models' deployment.

low accuracy rate in distinguishing between authentic and artificially generated facial images, with an average accuracy of 48.2% (*i.e.*, AI-generated faces were deemed most trustworthy). The same finding applies to text (Spitale et al., 2023).

As shown in Figure 1, there are several approaches to trace AI-generated content. Digital forensics detects AI-generated content by analyzing artifacts, but becomes less reliable as models improve. Fingerprinting enables content retrieval but requires storing signatures in databases, which is hard at scale and may lead to privacy concerns. Metadata provides cryptographic proof of origin but is often stripped during sharing. Watermarking is an important technique for tracing content from generative models since providers have control over the model's outputs. In this context, it is put forward by most of the regulations on AI (USA, 2023; Chi, 2023; Eur, 2023) to allow for a better transparency. The idea is to watermark the content during or after the generation to embed a proof that it is AI-generated, or an identifier of the specific model that generated it. It can be applied *post-hoc*, after generation; at *generation time* outside the model (*out-of-model*) by changing the way it generates the content, which usually offers better robustness or imperceptibility; or by modifying the model's weights (*in-model*), which gives an additional layer of protection in case of open-weights models or leaks, at the cost of additional implementation complexity.

Recent surveys on watermarking usually focus only on single domains, *e.g.*, image or text watermarking (Wan et al., 2022; Liu et al., 2024), or categorize watermarking methods based on the watermark properties, *e.g.*, the number of bits of information stored in the watermark (Boenisch, 2021). More recently, Zhao et al. (2024) gave a good overview of the terminology and threat models for watermarking AI-generated content. However, their taxonomy does not provide clear definitions of terms such as *semantic watermarking* or *in-processing watermarking*. It also does not capture all nuances needed to understand the level of protection offered by these types of watermarking methods. A case in point is watermarking for large language models (LLMs), which Zhao *et al*. call both *semantic* and *in-processing*, although it is unsuitable for open-weights models since the watermark is only applied at decoding time. This lack of clarity is problematic, especially for policymakers and regulators who need to understand the different methods and their implications.

Therefore, this paper aims to introduce: (1) A comprehensive cross-modal taxonomy of watermarking approaches for AI-generated content that clearly delineates methods based on when and how they intervene in the generation pipeline; (2) A detailed analysis of the trade-offs between different watermarking approaches to guide policymakers and regulators in understanding the technical capabilities and limitations of different watermarking methods.

2 TAXONOMY

This section provides an overview of watermarking techniques developed across the image, audio and text domains. It categorizes them into: **(2.1) post-hoc watermarking** that applies the content after generation; **(2.2) out-of-model watermarking**, which embeds the watermark during the generation process but requires no change to the original model weights; and **(2.3) in-model watermarking**, which embeds the watermark in the model itself by modifying the model weights.

2.1 POST-HOC WATERMARKING

Image. Traditional image watermarking methods are usually classified into two categories depending on the space on which the watermark is embedded. In *spatial domain*, the watermark is encoded by directly modifying pixels, such as flipping low-order bits of selected pixels (Van Schyndel et al., 1994). For example, Nikolaidis & Pitas (1998) slightly modify the intensity of randomly selected image pixels while taking into account properties of the human visual system, robustly to JPEG compression and low-pass filtering. Bas et al. (2002) create content descriptors defined by salient points and embeds the watermark by adding a pattern on triangles formed by the tessellation of these points. Ni et al. (2006) use the zero or the minimum points of the histogram of an image and slightly modifies the pixel grayscale values to embed data into the image. The second category is *frequency domain* watermarking, which usually spreads a pseudo-random noise sequence across the entire frequency spectrum of the host signal, and provides better robustness (Cox et al., 1997). The first step is a transformation that computes the frequency coefficients. The watermark is then added to these coefficients taking into account the human visual system. The coefficients are

mapped back onto the original pixel space through the inverse transformation to generate the watermarked image. The transform domains include Discrete Fourier Transform (DFT) (Urvoy et al., 2014), Quaternion Fourier Transform (QFT) (Bas et al., 2003; Ouyang et al., 2015), Discrete Cosine Transform (DCT) (Bors & Pitas, 1996; Piva et al., 1997; Barni et al., 1998; Li et al., 2011), Discrete Wavelet Transform (DWT) (Xia et al., 1998; Barni et al., 2001; Furon & Bas, 2008), both DWT and DCT (Feng et al., 2010; Zear et al., 2018), etc.

Deep learning-based methods have recently emerged as alternatives to traditional ones. They are often built as encoder/decoder networks: the encoder embeds the watermark in the image and the decoder tries to extract it. They are trained end-to-end to invisibly encode information while being resilient to transformations applied during training. This makes it easier to build robust systems and avoids algorithms hand-crafted for specific transformations. HiDDeN (Zhu et al., 2018) is the best example of this approach, and has been extended in several ways. Luo et al. (2020) add adversarial training in the attack simulation, to bring robustness to unknown transformations. Zhang et al. (2019b; 2020); Yu (2020) use an attention filter further improving imperceptibility. Ahmadi et al. (2020) adds a circular convolutional layer that helps diffusing the watermark signal over the image. Wen & Aydore (2019) use robust optimization with worst-case attack as if an adversary were trying to remove the mark. Another line of works focus on steganography (Baluja, 2017; Wengrowski & Dana, 2019; Zhang et al., 2019a; Tancik et al., 2020; Jing et al., 2021; Ma et al., 2022b), where the goal is to hide a message in the image without being detected, rather than to robustly extract it (*e.g.*, against crops). Many other approaches focused on improving robustness, imperceptibility, speed, etc. (Jia et al., 2021; Bui et al., 2023b;a; Huang et al., 2023b; Evennou et al., 2024; Pan et al., 2024b). In parallel to the encoder/decoder architectures, Vukotić et al. (2018; 2020), followed by Kishore et al. (2022) introduce an approach that is closer to traditional watermarking methods and uses neural networks as a fixed transform into a latent space. Since there is no inverse transform, the embedding is done iteratively by gradient descent over the pixels.

Audio. Given the similar nature of the signals, audio watermarking techniques are very similar to image watermarking ones (although they lag a bit behind). Traditional methods relied on embedding watermarks either in the time or frequency domains (Lie & Chang, 2006; Kalantari et al., 2009; Natgunanathan et al., 2012; Xiang et al., 2018; Su et al., 2018; Liu et al., 2019; Tai & Mansour, 2019), usually including domain specific features to design the watermark and its corresponding decoding function. To accurately extract audio watermarks, synchronization between the encoder and decoder is crucial. However, this can be disrupted by desynchronization attacks such as time and pitch scaling. To address this issue, various techniques have been developed. One approach is block repetition, which repeats the watermark signal along both the time and frequency domains (Kirovski & Malvar, 2003; Kirovski & Attias, 2003). Another method involves implanting synchronization bits into the watermarked signal (Xiang et al., 2014). During decoding, these synchronization bits serve to improve synchronization and mitigate the effects of de-synchronization attacks.

Most deep learning-based audio watermarking methods follow a HiDDeN-like encoder/decoder framework (Qu et al., 2023; Pavlović et al., 2022; Liu et al., 2023b; Ren et al., 2023; Chen et al., 2023; O'Reilly et al., 2024). The approach presented in AudioSeal (San Roman et al., 2024) is similar, but is zero-bit and allows for detection at the time-step level. Similar to the approach of Vukotić et al. (2018) in the image domain, Wu et al. (2023); Kong & Zhang (2020) embed the watermark by iteratively modifying the audio such that its representation lies within a certain region of the feature space of a pre-trained network.

Text. Watermarking text is commonly thought as more challenging than images or audio, since its discrete nature makes it harder to modify without altering its meaning.

The earliest works address watermarking for documents by altering text characteristics such as characters or spacing (Brassil et al., 1995), which is not very robust since this may be changed directly on a text editor. Text watermarking methods traditionally modify the grammatical or syntactical structure of the text with pre-established rules that embed watermarks without significantly altering its meaning (Topkara et al., 2005). For instance, Topkara et al. (2006c) embed information through synonym substitution, while Topkara et al. (2006b;a); Meral et al. (2009) use word reordering through passivization, preposing, topicalization, etc. Steganography methods have also been developed for text, working on the same principles (Winstein, 1998; Chapman et al., 2001; Bolshakov, 2004; Shirali-Shahreza & Shirali-Shahreza, 2008; Chang & Clark, 2014; Xiang et al., 2017). These edit-

based systems suffer from low robustness to text modifications, and low payload, *e.g.*, 1 or 2 bits per sentence as in CoverTweet (Wilson & Ker, 2016). Similar to other media, deep learning-based methods have been developed more recently. These methods either use pre-trained masked language models (Ueoka et al., 2021) or end-to-end encoder/decoder networks (Abdelnabi & Fritz, 2021).

2.2 Out-of-model watermarking

Out-of-model watermarking modifies the generation process but does not alter the model's parameters. This method is easier to implement because it does not require retraining or fine-tuning the model. However, once the model is released, users may disable the watermarking step, and the watermarking technique becomes known which would lead to potential vulnerabilities or exploitation.

Image. A prominent approach in diffusion models involves modifying the initial noise to embed a watermark. For instance, Tree-Ring (Wen et al., 2023) adds tree-ring-shaped patterns to the initial noise and later extracts the watermark by inverting the diffusion process. Subsequent works improve this technique: Hong et al. (2024) refine the inversion process, Ci et al. (2024b) extend the method for multi-bit watermarking, and Lei et al. (2024) propose an encoder-decoder framework to embed and extract watermarks within the initial noise.

Another approach is to use adapters, as done by Ci et al. (2024a) and Rezaei et al. (2024), taking the secret message as input. These methods operate out-of-model, since a user may choose to remove the adapters or change the message before generating the content.

Audio. To the best of our knowledge, contrary to image watermarking, no work has explored out-of-model watermarking for audio generation, except for Zhou et al. (2024), which embeds watermarks in the latent space of a speech synthesis model.

Text. For text generated by Large Language Models (LLMs), early watermarking techniques emerged shortly after the release of ChatGPT. These methods typically use a secret key and a hash of previous tokens to modify token generation. Kirchenbauer et al. (2023) modify the probability distribution by biasing a subset of vocabulary tokens, while Aaronson & Kirchner (2023) use the Gumbel trick to alter the sampling process. Follow-up research improves these techniques: Fernandez et al. (2023a) refine statistical detection methods, while Yoo et al. (2023b;a); Qu et al. (2024) develop multi-bit watermarking techniques. Christ et al. (2023); Kuditipudi et al. (2023) introduce position-based pseudo-randomness to enhance detectability. Other works (Huang et al., 2023a; Giboulot & Furon, 2024) improve detection algorithms by embedding watermarks in high-entropy text segments. Further methods incorporate semantic information from previous tokens to increase robustness to text modifications (Liu et al., 2023a; Liu & Bu, 2024; Fu et al., 2024; Hou et al., 2023; 2024). To facilitate benchmarking, Piet et al. (2023); Pan et al. (2024a) develop toolkits for evaluating watermark robustness.

2.3 In-model watermarking

In contrast to out-of-model watermarking, in-model watermarking embeds the watermark within the model's parameters. This approach enables open-sourcing models without revealing the watermarking method. However, it usually requires training or fine-tuning the model, making implementation more complex.

Image. Early techniques embedded watermarks by modifying the training set (Wu et al., 2020; Yu et al., 2021; Zhao et al., 2023), though these methods are computationally expensive and difficult to scale. Alternative methods integrate watermarking objectives into the training process. Fei et al. (2022; 2024) incorporate additional loss terms in Generative Adversarial Networks (GANs) so that generated images inherently contain watermarks. For diffusion models, Stable Signature (Fernandez et al., 2023b) fine-tunes the latent decoder to embed a watermark, while Feng et al. (2024) fine-tune the U-Net responsible for noise prediction. Yu et al. (2022); Fei et al. (2023) propose hypernetwork-based watermarking techniques, eliminating the need to fine-tune models for individual users. Similarly, Kim et al. (2024) adapt Stable Signature to generate the watermarked LDM decoder weights on-the-fly which is faster than with fine-tuning.

Table 1: Watermarking approaches, their pros and cons, and suitable use cases.

Approach	Method description	Pros and cons	Suitable for
Post-hoc	Separate watermarking models from generative AI models	+ Flexible, model-agnostic, easy to implement + Allows smooth improvements − Limited in open-source scenarios − Can be bypassed easily	Online-hosted models, APIs, protecting content after creation
Generation-time Out-of-model	Alters sampling or inference process without changing the underlying model	+ Improved runtime performance + Easy to implement − Easily bypassed in open-source	Scenarios for which the generation speed matters, image when detection speed is not important, LLM watermarking
Generation-time In-model	Embeds watermark within generative model's weights	+ Better protection + Suitable for open-source − Requires model modification − Computationally expensive, hinders security updates	Models deployed on device or open-sourced, requiring strong protection

Audio. The literature on audio generation-based watermarking is also still emerging. San Roman et al. (2025) introduce a method that embeds watermarks robust to audio tokenization. Juvela & Wang (2023) propose a GAN-like approach where a generator and a collaborator co-train, ensuring that watermarks appear in generated speech but not in real speech.

Text. Fewer methods exist for in-model text watermarking compared to out-of-model approaches. Gu et al. (2023) train language models to generate watermarked text by distilling existing watermarking methods into the model's parameters. Xu et al. (2024) use reinforcement learning to embed watermarks, leveraging techniques similar to instruction fine-tuning (Ouyang et al., 2022) with a paired language model detector as the reward model.

2.4 Summary of the pros and cons

Each watermarking approach presents a trade-off between flexibility, robustness, security, and ease of implementation.

Post-hoc watermarking, allows for the separated development of watermarking models from generative AI models, enabling continuous improvement of watermarking techniques (*e.g.*, add security patches against known attacks) without hindering the progress of generative model research. Furthermore, post-hoc watermarking is model-agnostic, meaning any watermarking model can be paired with any generative model, providing flexibility and versatility. This approach is particularly well-suited for protecting online-hosted models and APIs. Additionally, post-hoc watermarking enables the post-processing of watermarks before integration with generated content, allowing for enhancements such as imperceptibility, localized watermarking (San Roman et al., 2024; Sander et al., 2025), and tamper localization (Zhang et al., 2024) to be designed independently of content generation, thereby strengthening the overall robustness of the watermarking scheme.

Having said that, post-hoc watermarking has its limitations particularly in open-source or open-weights scenarios. In such cases, the watermark can sometimes be easily bypassed as simply as by commenting out one single line of code, as exemplified by the Stable Diffusion repository[1]. To mitigate this, code obfuscation techniques can be employed (Zhou et al., 2023). These methods conceal critical model details – such as structure, parameters, and attributes – through techniques like renaming, parameter encapsulation, and neural structure obfuscation.

[1] https://github.com/CompVis/stable-diffusion

Table 2: Categorization of image, audio and text watermarking techniques from the above taxonomy. In particular, post-hoc watermarking methods are very popular, especially for image and audio, while generation-time out-of-model is more popular for text with the branch of LLM watermarking at decoding time. In comparison, in-model watermarking is still at an early stage.

Category	Domain	References
Post-hoc	Image	(Van Schyndel et al., 1994; Nikolaidis & Pitas, 1998; Bas et al., 2002; Ni et al., 2006; Cox et al., 1997; Zhu et al., 2018; Luo et al., 2020; Zhang et al., 2019b; 2020; Yu, 2020; Ahmadi et al., 2020; Wen & Aydore, 2019; Baluja, 2017; Wengrowski & Dana, 2019; Zhang et al., 2019a; Tancik et al., 2020; Jing et al., 2021; Ma et al., 2022b; Jia et al., 2021; Bui et al., 2023b;a; Huang et al., 2023b; Evennou et al., 2024; Pan et al., 2024b; Vukotić et al., 2018; 2020; Kishore et al., 2022)
	Audio	(Qu et al., 2023; Pavlović et al., 2022; Liu et al., 2023b; Ren et al., 2023; Chen et al., 2023; O'Reilly et al., 2024; San Roman et al., 2024)
	Text	(Brassil et al., 1995; Topkara et al., 2005; 2006c;b; Meral et al., 2009; Winstein, 1998; Chapman et al., 2001; Bolshakov, 2004; Shirali-Shahreza & Shirali-Shahreza, 2008; Chang & Clark, 2014; Xiang et al., 2017; Ueoka et al., 2021; Abdelnabi & Fritz, 2021)
Generation-time Out-of-Model	Image	(Wen et al., 2023; Hong et al., 2024; Ci et al., 2024b; Lei et al., 2024; Yu et al., 2022; Ci et al., 2024a; Rezaei et al., 2024)
	Audio	(Zhou et al., 2024)
	Text	(Venugopal et al., 2011; Kirchenbauer et al., 2023; Aaronson & Kirchner, 2023; Fernandez et al., 2023a; Yoo et al., 2023b;a; Qu et al., 2024; Christ et al., 2023; Kuditipudi et al., 2023; Huang et al., 2023a; Lee et al., 2023; Liu et al., 2023a; Liu & Bu, 2024; Fu et al., 2024; Hou et al., 2023; 2024; Giboulot & Furon, 2024; Piet et al., 2023; Pan et al., 2024a)
Generation-time In-Model	Image	(Wu et al., 2020; Yu et al., 2021; Zhao et al., 2023; Fei et al., 2022; 2024; Fernandez et al., 2023b; Fei et al., 2023; Kim et al., 2024; Feng et al., 2024)
	Audio	(San Roman et al., 2025; Juvela & Wang, 2023)
	Text	(Gu et al., 2023; Xu et al., 2024)

Generation-time watermarking seamlessly integrates a watermark into AI-generated content during its creation process. Generation-time methods are further divided into out-of-model and in-model approaches as follows.

- **Out-of-model watermarking** alters the sampling or inference process without changing the underlying model. This approach offers several advantages, including improved runtime performance (which is particularly relevant for video watermarking), ease of implementation (particularly for text), and better robustness or imperceptibility. Despite the advantages of this approach, it does not provide better safety for open-sourcing code and weights than post-hoc watermarking. The inference process can be easily replaced by the standard inference process, rendering the watermark ineffective. For example, in LLM watermarking (Kirchenbauer et al., 2023; Aaronson & Kirchner, 2023), the watermarking technique can be replaced by a standard decoding (like top-p sampling) if the model's weights are released. Similarly, nothing enforces the users of any image diffusion model to sample from it using the Tree-Ring (Wen et al., 2023) or Gaussian Shading (Yang et al., 2024) watermarking methods. For the specific case of these last two methods, another drawback is that the watermark detection process is slow.

- **In-model watermarking**, on the other hand, embeds the watermark within some of the generative model's weights. This approach offers stronger protection against watermark removal or tampering, particularly for open-weight models. However, it comes with certain limitations. Implementing in-model watermarking typically requires full or partial fine-tuning, which can be computationally expensive, time-consuming, and challenging for version control. Moreover, many methods, such as Stable Signature (Fernandez et al., 2023b), apply watermarking within the decoders of generative models – components responsible for transforming latent or quantized representa-

tions back into pixel or waveform space. While image model decoders are rarely fine-tuned or altered by users, the same does not hold true in the audio domain, where vocoders are frequently fine-tuned, open-sourced, and repurposed across various projects. As a result, embedding watermarking techniques in these models may not be sufficient to ensure protection in open-source environments, as these components can be easily swapped out for alternatives.

3 Related work on content tracing

Tracing the origin of digital content is a problem that is traditionally approached passively, through copy detection or digital forensics. We provide an overview of the literature on these two approaches, and across image, audio and text modalities, with a particular focus on AI-generated content. We also summarize their pros and cons in Tab. 3.

3.1 Digital forensics (or passive detection)

More specific to the context of AI-generated content, digital forensics methods aim to detect if a piece of content has been generated or altered by an AI model. These methods are commonly referred to as passive detection in recent literature, as they enable identification of AI-generated content without requiring any prior intervention or modification, such as watermarking, or pre-registration in a database. Instead, passive detection allows for post-hoc analysis, where investigators can examine the content itself to determine whether it has been generated or manipulated by an AI model. Most methods spot imperceptible hidden traces of generated content, such as variation in words probabilities (Mitchell et al., 2023), odd frequencies in images (Corvi et al., 2023) or voice synthesizer artifacts (Le et al., 2023). Relying on these traces makes the detectors very brittle to shifts in the distribution of content, and makes them fall short in effectiveness compared to watermarking techniques (Sadasivan et al., 2023; Saberi et al., 2024). As a key example, one state-of-the-art detection method (Wang et al., 2023b) is fooled to random chance levels simply by compressing generated images with JPEG (Grommelt et al., 2024), because all natural images in their training dataset were in the JPEG format. Besides, these detectors are likely to get worse as generative models get better and as their artifacts disappear.

Image. Detection of synthetic/manipulated images has a long history (Farid, 2009; Barni et al., 2023). It is now very active in the context of deep-fake detection (Guarnera et al., 2020; Zhao et al., 2021). Many works focus on the detection of GAN-generated images (Chai et al., 2020; Gragnaniello et al., 2021; Wang et al., 2020; Zhang et al., 2019c). One approach is to detect inconsistencies in generated images via lights (Farid, 2022a), perspective (Farid, 2022b; Sarkar et al., 2024), physical objects (Ma et al., 2022a) or faces (Li & Lyu, 2018; Wang et al., 2019; Boháček & Farid, 2023). These approaches are restricted to photo-realistic images or faces, artworks not intended to be physically correct are not covered. Other approaches track traces left by the generators in the spatial (Marra et al., 2019; Yu et al., 2019) or frequency (Frank et al., 2020; Zhang et al., 2019c) domains. There are extensions to diffusion model in recent works (Corvi et al., 2022; Sha et al., 2022; Epstein et al., 2023) that show encouraging results.

Speech. In the forensics community, the detection of synthetic speech is traditionally done by building features and exploiting statistical differences between fake and real. These features can be hand-crafted from the analysis of waveforms, spectrograms or formants (Sahidullah et al., 2015; Janicki, 2015; AlBadawy et al., 2019; Borrelli et al., 2021; Cuccovillo et al., 2024) and/or learned (Müller et al., 2022; Barrington et al., 2023). The approach of most audio generation papers (Borsos et al., 2022; Kharitonov et al., 2023; Borsos et al., 2023; Le et al., 2023) is to train end-to-end deep learning classifiers on what their models generate, similarly as Zhang et al. (2017); Tak et al. (2021b;a); Jung et al. (2022). These networks primarily focus on non-vocal spectrogram regions (Salvi et al., 2023; 2024), which explains why they are sensitive to the addition or removal of audio artifacts.

Text. Detection of LLM-generated text is a relatively new field. Similarly to images and audio, it either relies on hand-crafted textual features or on models trained for detection. However, contrary to images and audio, the former approach is more popular since training and running LLMs is

computationally expensive and cumbersome. The features are for instance based on n-gram analysis (Yang et al., 2023), on the probability and rank of the observed tokens (Gehrmann et al., 2019; Ippolito et al., 2019), on the perplexity of the text observed by the LLM under scrutiny or a surrogate model (Vasilatos et al., 2023; Wang et al., 2023a), on several of them (Hans et al., 2024), or on its curvature (Mitchell et al., 2023). The other class of methods trains classifiers (Bhattacharjee et al., 2024) or other language models, often by fine-tuning the model to detect itself (Solaiman et al., 2019; Zellers et al., 2019). Similarly to images and audios, these detection methods are often brittle to shifts in the text distribution and not very reliable (Sadasivan et al., 2023).

3.2 Fingerprinting and retrieval (or copy detection)

Fingerprinting involves creating a unique identifier for a piece of digital content, called hash or fingerprint in reference to the uniqueness of human fingerprints. This fingerprint can then be used to identify the content even if it has been modified or compressed, but does not allow to reconstruct the original content. Copy detection, on the other hand, involves comparing two pieces of digital content to determine if they are identical or similar. This is often done using indexing algorithms that store the fingerprints of all the content in a database, and then compare the hash of the queried content to the hashes in the database to determine if it is a copy.

These hashes are vector representations that can be binary or real-valued ($\in \{0,1\}^k$ or $\mathbb{R}^k$). They were traditionally hand-crafted with color histograms, GIST descriptors, constellation maps, etc. (Swain & Ballard, 1991; Oliva & Torralba, 2001; Wang et al., 2003; Perronnin et al., 2010), but are now usually generated from self-supervised feature extractors (Chen et al., 2020; Oquab et al., 2024; Devlin et al., 2018; Hsu et al., 2021; Raffel et al., 2020). The feature extractors are not perfectly robust to content modifications. In other words, the hashes are not perfectly invariant to transformations, *e.g.*, an audio and its $\times 1.25$ speed-up version may have different ones. Besides, storing the hashes is cumbersome and reverse search, *i.e.*, finding the content that has a given hash, must be approximate to be tractable at scale. Therefore, the hashes are often stored using methods like locality-sensitive hashing (LSH) (Charikar, 2002; Datar et al., 2004) or product quantization (PQ) (Jegou et al., 2010). These indexing structures have a dual role of compressing the hashes and enabling fast approximate search. See FAISS (Douze et al., 2024) for a review and efficient implementations of these algorithms. The two above factors result in errors especially in an adversarial setting (Douze et al., 2021; Papakipos et al., 2022; Wang et al., 2022). More recently Active indexing (Fernandez et al., 2023c) aims to reduce these errors by actively modifying images before their release, in a similar way to watermarking, Krishna et al. (2023) demonstrate the use of retrieval in the context of AI-generated text and Défossez et al. (2024) in the context of a speech LLM production environment. Another downside is the need of storing the hashes in a database, which makes it harder to share and impossible for open content moderation systems.

3.3 Cryptographic metadata

In the context of origin tracing, cryptographic metadata is digital information associated with a piece of content to provide evidence of its authenticity and/or provenance. The Coalition for Content Provenance and Authenticity (C2PA) and the International Press Telecommunications Council (IPTC) have recently proposed two standards. The upside is that forging fake cryptographic signatures is extremely hard, however the metadata are often removed during re-uploads or screenshots. For instance, a study by Imatag (2018) shows that only 3% of images on the internet come with copyright metadata. It is therefore particularly suited for authenticating real content, for which the creators want the content to be traced back to them, but less for tracing origin of AI-generated content in the wild. Besides, they are more a subject of standardization bodies than research, because all actors of the content production chain must adhere to the same protocol for it to be effective.

This metadata includes various types of cryptographic information, such as timestamps, used to record the date and time when the content was created or modified, or provenance information about the origin of the content like its creator. All this information is encrypted with a private key, using algorithms like RSA (Rivest et al., 1978) or ECDSA (Johnson et al., 2001), which makes it impossible to forge without the private key but possible to verify with the public key. It can also include content bindings, used to check the authenticity of the content and ensure that it has not been tampered with. They are encrypted hash values representing the content which are attached to it as a

Table 3: Content tracing approaches and their properties with respect to flexibility, robustness, and security.

Approach		Flexibility	Robustness	Security
Passive detection		High (works on any content without prior preparation)	Medium (vulnerable to distribution shifts and model improvements)	Medium (cannot be trivially bypassed but detectors may become obsolete)
Fingerprinting		Medium (requires storing in database, hard to scale)	Medium (somewhat robust to modifications)	Low (privacy concerns with centralized databases)
Metadata		Medium (still requires adherence to standards)	Low (metadata easily removed during sharing)	High (cryptographically secure)
Visible WM		High (easy to apply to any content)	Low (easily removed or tampered with)	Low (vulnerable to removal)
Invisible WM	Post-hoc	High (can be applied after generation)	High (but can be still be removed by strong attackers)	Medium (can be bypassed if open-sourced)
	Generation time Out-of-model	Medium (requires special encoder/decoder)	High (but can be still be removed by strong attackers)	Medium (can be bypassed if open-sourced)
	Generation time In-model	Low (needs training the generative model specifically)	High (but can be still be removed by strong attackers)	High (tied to the model's weights)

digital signature. The bindings are categorized into two types. Hard bindings (a.k.a., , cryptographic bindings), are computed directly from the raw bits of the content and can be used to ensure that the manifest belongs with the asset and that the asset has not been modified. Soft bindings, on the other hand, are computed from the digital content of an asset (as in fingerprinting) and can be used to identify derived assets (C2PA, 2024).

3.4 Visible watermarking

Visible watermarks are straightforward and widely recognized. However, in addition to degrading the quality of the content, they are also easy to remove or tamper with, making them less reliable. For instance, a visible watermark on the left side of an image can be removed by cropping the image or through inpainting techniques (Dekel et al., 2017). *Invisible watermarks* (the focus of the paper) are imperceptibility embedded within the content itself. It makes them riskier to remove because there is no certainty that it has been removed without access to the watermark detector or extractor.

4 Conclusion

This paper proposes a taxonomy of watermarking methods for generative AI, categorizing them into post-hoc watermarking (adding watermarks after generation), out-of-model watermarking (embedding during generation), and in-model watermarking (integrating into model parameters). Each approach presents trade-offs between flexibility, robustness, security, and implementation complexity. Our discussion also highlights alternative methods like digital forensics and fingerprinting, which complement watermarking techniques but have their own limitations. This taxonomy spans image, audio, and text domains, providing a structured overview of existing techniques. It helps stakeholders understand that no single method is universally superior, as each approach's suitability depends on specific use cases and requirements. These methods can be viewed as complementary tools for addressing various challenges in AI-generated content protection.

While this taxonomy provides distinctions not present in previous work and helps understand trends in the literature, it still has limitations. In particular, the distinction between in-model and out-of-model watermarking does not fully capture the fundamental similarities between methods that operate at the signal level versus those that manipulate high-level semantic features. For instance, watermarking approaches from Zhou et al. (2024); Juvela & Wang (2023); San Roman et al. (2025) share underlying principles in how they modify signal-level attributes, despite being categorized differently in our taxonomy. Future work could refine this categorization to better distinguish between methods that apply watermarks through low-level signal manipulation versus those that meaningfully alter the output distribution along semantic dimensions.

REFERENCES

Chinese ai governance rules, 2023. URL http://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm. Accessed on August 29, 2023.

European ai act, 2023. URL https://artificialintelligenceact.eu/. Accessed on August 29, 2023.

Scott Aaronson and Hendrik Kirchner. Watermarking gpt outputs, 2023. URL https://www.scottaaronson.com/talks/watermark.ppt.

Sahar Abdelnabi and Mario Fritz. Adversarial watermarking transformer: Towards tracing text provenance with data hiding. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 121–140. IEEE, 2021.

Mahdi Ahmadi, Alireza Norouzi, Nader Karimi, Shadrokh Samavi, and Ali Emami. Redmark: Framework for residual diffusion watermarking based on deep networks. *Expert Systems with Applications*, 146:113157, 2020. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2019.113157. URL https://www.sciencedirect.com/science/article/pii/S0957417419308759.

Ehab A AlBadawy, Siwei Lyu, and Hany Farid. Detecting ai-synthesized speech using bispectral analysis. In *CVPR workshops*, pp. 104–109, 2019.

Shumeet Baluja. Hiding images in plain sight: Deep steganography. *NeurIPS*, 2017.

Mauro Barni, Franco Bartolini, Vito Cappellini, and Alessandro Piva. A dct-domain system for robust image watermarking. *Signal processing*, 66(3):357–372, 1998.

Mauro Barni, Franco Bartolini, and Alessandro Piva. Improved wavelet-based watermarking through pixel-wise masking. *IEEE transactions on image processing*, 10(5):783–791, 2001.

Mauro Barni, Patrizio Campisi, Edward J Delp, Gwenaël Doërr, Jessica Fridrich, Nasir Memon, Fernando Pérez-González, Anderson Rocha, Luisa Verdoliva, and Min Wu. Information forensics and security: A quarter-century-long journey. *IEEE Signal Processing Magazine*, 40(5):67–79, 2023.

Sarah Barrington, Romit Barua, Gautham Koorma, and Hany Farid. Single and multi-speaker cloned voice detection: From perceptual to learned features. *arXiv preprint arXiv:2307.07683*, 2023.

Patrick Bas, J-M Chassery, and Benoit Macq. Geometrically invariant watermarking using feature points. *IEEE transactions on image Processing*, 11(9):1014–1028, 2002.

Patrick Bas, Nicolas Le Bihan, and J-M Chassery. Color image watermarking using quaternion fourier transform. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 3, pp. III–521. IEEE, 2003.

Amrita Bhattacharjee, Raha Moraffah, Joshua Garland, and Huan Liu. Eagle: A domain generalization framework for ai-generated text detection. *arXiv preprint arXiv:2403.15690*, 2024.

Franziska Boenisch. A systematic review on model watermarking for neural networks. *Frontiers in big Data*, 4:729663, 2021.

Matyáš Boháček and Hany Farid. A geometric and photometric exploration of gan and diffusion synthesized faces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 874–883, 2023.

Igor A Bolshakov. A method of linguistic steganography based on collocationally-verified synonymy. In *International Workshop on Information Hiding*, pp. 180–191. Springer, 2004.

Clara Borrelli, Paolo Bestagini, Fabio Antonacci, Augusto Sarti, and Stefano Tubaro. Synthetic speech detection through short-term and long-term prediction traces. *EURASIP Journal on Information Security*, 2021 (1):1–14, 2021.

Adrian G Bors and Ioannis Pitas. Image watermarking using dct domain constraints. In *ICIP*, 1996.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matthew Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. Audiolm: A language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2523–2533, 2022.

Zalán Borsos, Matt Sharifi, Damien Vincent, Eugene Kharitonov, Neil Zeghidour, and Marco Tagliasacchi. Soundstorm: Efficient parallel audio generation. *arXiv preprint arXiv:2305.09636*, 2023.

Jack T Brassil, Steven Low, Nicholas F Maxemchuk, and Lawrence O'Gorman. Electronic marking and identification techniques to discourage document copying. *IEEE Journal on Selected Areas in Communications*, 13(8):1495–1504, 1995.

Tu Bui, Shruti Agarwal, and John Collomosse. Trustmark: Universal watermarking for arbitrary resolution images. *arXiv preprint arXiv:2311.18297*, 2023a.

Tu Bui, Shruti Agarwal, Ning Yu, and John Collomosse. Rosteals: Robust steganography using autoencoder latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 933–942, 2023b.

C2PA. C2pa specification, 2024. URL https://c2pa.org/specifications/. Accessed on August 29, 2024.

Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *European conference on computer vision*, pp. 103–120. Springer, 2020.

Ching-Yun Chang and Stephen Clark. Practical linguistic steganography using contextual synonym substitution and a novel vertex coding method. *Computational linguistics*, 40(2):403–448, 2014.

Mark Chapman, George I Davida, and Marc Rennhard. A practical and effective approach to large-scale automated linguistic steganography. In *International Conference on Information Security*, pp. 156–165. Springer, 2001.

Moses S Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pp. 380–388, 2002.

Guangyu Chen, Yu Wu, Shujie Liu, Tao Liu, Xiaoyong Du, and Furu Wei. Wavmark: Watermarking for audio generation. *arXiv preprint arXiv:2308.12770*, 2023.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

Miranda Christ, Sam Gunn, and Or Zamir. Undetectable watermarks for language models. *Cryptology ePrint Archive*, 2023.

Hai Ci, Yiren Song, Pei Yang, Jinheng Xie, and Mike Zheng Shou. Wmadapter: Adding watermark control to latent diffusion models. *arXiv preprint arXiv:2406.08337*, 2024a.

Hai Ci, Pei Yang, Yiren Song, and Mike Zheng Shou. Ringid: Rethinking tree-ring watermarking for enhanced multi-key identification. *arXiv preprint arXiv:2404.14055*, 2024b.

Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. *arXiv preprint arXiv:2211.00680*, 2022.

Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Ingemar J Cox, Joe Kilian, F Thomson Leighton, and Talal Shamoon. Secure spread spectrum watermarking for multimedia. *IEEE transactions on image processing*, 1997.

Luca Cuccovillo, Milica Gerhardt, and Patrick Aichroth. Audio transformer for synthetic speech detection via multi-formant analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4409–4417, 2024.

Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the twentieth annual symposium on Computational geometry*, pp. 253–262, 2004.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*, 2024.

Tali Dekel, Michael Rubinstein, Ce Liu, and William T Freeman. On the effectiveness of visible watermarks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2146–2154, 2017.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Matthijs Douze, Giorgos Tolias, Ed Pizzi, Zoë Papakipos, Lowik Chanussot, Filip Radenovic, Tomas Jenicek, Maxim Maximov, Laura Leal-Taixé, Ismail Elezi, et al. The 2021 image similarity dataset and challenge. *arXiv preprint arXiv:2106.09672*, 2021.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. *arXiv preprint arXiv:2401.08281*, 2024.

David C Epstein, Ishan Jain, Oliver Wang, and Richard Zhang. Online detection of ai-generated images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 382–392, 2023.

Gautier Evennou, Vivien Chappelier, Ewa Kijak, and Teddy Furon. Swift: Semantic watermarking for image forgery thwarting. *arXiv preprint arXiv:2407.18995*, 2024.

Hany Farid. Image forgery detection. *IEEE Signal processing magazine*, 26(2):16–25, 2009.

Hany Farid. Lighting (in) consistency of paint by text. *arXiv preprint arXiv:2207.13744*, 2022a.

Hany Farid. Perspective (in) consistency of paint by text. *arXiv preprint arXiv:2206.14617*, 2022b.

Jianwei Fei, Zhihua Xia, Benedetta Tondi, and Mauro Barni. Supervised gan watermarking for intellectual property protection. In *2022 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–6. IEEE, 2022.

Jianwei Fei, Zhihua Xia, Benedetta Tondi, and Mauro Barni. Robust retraining-free gan fingerprinting via personalized normalization. In *2023 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–6. IEEE, 2023.

Jianwei Fei, Zhihua Xia, Benedetta Tondi, and Mauro Barni. Wide flat minimum watermarking for robust ownership verification of gans. *IEEE Transactions on Information Forensics and Security*, 2024.

Liu Ping Feng, Liang Bin Zheng, and Peng Cao. A dwt-dct based blind watermarking algorithm for copyright protection. In *ICCSIT*. IEEE, 2010.

Weitao Feng, Wenbo Zhou, Jiyan He, Jie Zhang, Tianyi Wei, Guanlin Li, Tianwei Zhang, Weiming Zhang, and Nenghai Yu. Aqualora: Toward white-box protection for customized stable diffusion models via watermark lora. *arXiv preprint arXiv:2405.11135*, 2024.

Pierre Fernandez, Antoine Chaffin, Karim Tit, Vivien Chappelier, and Teddy Furon. Three bricks to consolidate watermarks for large language models. In *International Workshop on Information Forensics and Security (WIFS)*, 2023a.

Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. In *International Conference on Computer Vision*, pp. 22466–22477, 2023b.

Pierre Fernandez, Matthijs Douze, Hervé Jégou, and Teddy Furon. Active image indexing. In *International Conference on Learning Representations (ICLR)*, 2023c.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deepfake image recognition. In *International conference on machine learning*, pp. 3247–3258. PMLR, 2020.

Yu Fu, Deyi Xiong, and Yue Dong. Watermarking conditional text generation for ai detection: Unveiling challenges and a semantic-aware watermark remedy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18003–18011, 2024.

Teddy Furon and Patrick Bas. Broken arrows. *EURASIP Journal on Information Security*, 2008:1–13, 2008.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. Gltr: Statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043*, 2019.

Eva Giboulot and Teddy Furon. Watermax: breaking the llm watermark detectability-robustness-quality tradeoff. *arXiv preprint arXiv:2403.04808*, 2024.

Josh A Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. Generative language models and automated influence operations: Emerging threats and potential mitigations. *arXiv preprint arXiv:2301.04246*, 2023.

Diego Gragnaniello, Davide Cozzolino, Francesco Marra, Giovanni Poggi, and Luisa Verdoliva. Are gan generated images easy to detect? a critical analysis of the state-of-the-art. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2021.

Patrick Grommelt, Louis Weiss, Franz-Josef Pfreundt, and Janis Keuper. Fake or jpeg? revealing common biases in generated image detection datasets. *arXiv preprint arXiv:2403.17608*, 2024.

Chenchen Gu, Xiang Lisa Li, Percy Liang, and Tatsunori Hashimoto. On the learnability of watermarks for language models. *arXiv preprint arXiv:2312.04469*, 2023.

Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. Deepfake detection by analyzing convolutional traces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 666–667, 2020.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Spotting llms with binoculars: Zero-shot detection of machine-generated text. *arXiv preprint arXiv:2401.12070*, 2024.

Douglas Harris. Deepfakes: False pornography is here and the law cannot protect you. *Duke L. & Tech. Rev.*, 17:99, 2018.

Seongmin Hong, Kyeonghyun Lee, Suh Yoon Jeon, Hyewon Bae, and Se Young Chun. On exact inversion of dpm-solvers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7069–7078, 2024.

Abe Bohan Hou, Jingyu Zhang, Tianxing He, Yichen Wang, Yung-Sung Chuang, Hongwei Wang, Lingfeng Shen, Benjamin Van Durme, Daniel Khashabi, and Yulia Tsvetkov. Semstamp: A semantic watermark with paraphrastic robustness for text generation. *arXiv preprint arXiv:2310.03991*, 2023.

Abe Bohan Hou, Jingyu Zhang, Yichen Wang, Daniel Khashabi, and Tianxing He. k-semstamp: A clustering-based semantic watermark for detection of machine-generated text. *arXiv preprint arXiv:2402.11399*, 2024.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

Baihe Huang, Banghua Zhu, Hanlin Zhu, Jason D. Lee, Jiantao Jiao, and Michael I. Jordan. Towards optimal statistical watermarking, 2023a.

Jiangtao Huang, Ting Luo, Li Li, Gaobo Yang, Haiyong Xu, and Chin-Chen Chang. Arwgan: Attention-guided robust image watermarking model based on gan. *IEEE Transactions on Instrumentation and Measurement*, 72:1–17, 2023b.

Imatag. State of image metadata, 2018. URL https://www.imatag.com/blog/state-of-image-metadata-in-2018.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. Automatic detection of generated text is easiest when humans are fooled. *arXiv preprint arXiv:1911.00650*, 2019.

Artur Janicki. Spoofing countermeasure based on analysis of linear prediction error. In *Sixteenth annual conference of the international speech communication association*, 2015.

Herve Jegou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *Transactions on Pattern Analysis and Machine Intelligence*, 33(1):117–128, 2010.

Zhaoyang Jia, Han Fang, and Weiming Zhang. Mbrs: Enhancing robustness of dnn-based watermarking by mini-batch of real and simulated jpeg compression. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 41–49, 2021.

Junpeng Jing, Xin Deng, Mai Xu, Jianyi Wang, and Zhenyu Guan. Hinet: Deep image hiding by invertible network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4733–4742, 2021.

Don Johnson, Alfred Menezes, and Scott Vanstone. The elliptic curve digital signature algorithm (ecdsa). *International journal of information security*, 1:36–63, 2001.

Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 6367–6371. IEEE, 2022.

Lauri Juvela and Xin Wang. Collaborative watermarking for adversarial speech synthesis. *arXiv preprint arXiv:2309.15224*, 2023.

Nima Khademi Kalantari, Mohammad Ali Akhaee, Seyed Mohammad Ahadi, and Hamidreza Amindavar. Robust multiplicative patchwork method for audio watermarking. *IEEE Trans. Speech Audio Process.*, 17 (6):1133–1141, 2009. doi: 10.1109/TASL.2009.2019259. URL https://doi.org/10.1109/TASL.2009.2019259.

Eugene Kharitonov, Damien Vincent, Zalán Borsos, Raphaël Marinier, Sertan Girgin, Olivier Pietquin, Matthew Sharifi, Marco Tagliasacchi, and Neil Zeghidour. Speak, read and prompt: High-fidelity text-to-speech with minimal supervision. *ArXiv*, abs/2302.03540, 2023.

Changhoon Kim, Kyle Min, Maitreya Patel, Sheng Cheng, and Yezhou Yang. 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*, pp. 8974–8983, 2024.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. *arXiv preprint arXiv:2301.10226*, 2023.

Darko Kirovski and Hagai Attias. Audio watermark robustness to desynchronization via beat detection. In Fabien A. P. Petitcolas (ed.), *Information Hiding*, pp. 160–176, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg. ISBN 978-3-540-36415-3.

Darko Kirovski and Henrique S. Malvar. Spread-spectrum watermarking of audio signals. *IEEE Trans. Signal Process.*, 51(4):1020–1033, 2003. doi: 10.1109/TSP.2003.809384. URL https://doi.org/10.1109/TSP.2003.809384.

Varsha Kishore, Xiangyu Chen, Yan Wang, Boyi Li, and Kilian Q Weinberger. Fixed neural network steganography: Train the images, not the network. In *International Conference on Learning Representations*, 2022.

Kate Knibbs. Scammy ai-generated book rewrites are flooding amazon, 2024. URL https://www.wired.com/story/scammy-ai-generated-books-flooding-amazon/. Accessed on Jul. 29, 2024.

Yehao Kong and Jiliang Zhang. Adversarial audio: A new information hiding method. In *INTERSPEECH*, pp. 2287–2291, 2020.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *Advances in Neural Information Processing Systems*, 36:27469–27500, 2023.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust distortion-free watermarks for language models. *arXiv preprint arXiv:2307.15593*, 2023.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al. Voicebox: Text-guided multilingual universal speech generation at scale. *arXiv preprint arXiv:2306.15687*, 2023.

Taehyun Lee, Seokhee Hong, Jaewoo Ahn, Ilgee Hong, Hwaran Lee, Sangdoo Yun, Jamin Shin, and Gunhee Kim. Who wrote this code? watermarking for code generation. *arXiv preprint arXiv:2305.15060*, 2023.

Liangqi Lei, Keke Gai, Jing Yu, and Liehuang Zhu. Diffusetrace: A transparent and flexible watermarking scheme for latent diffusion model. *arXiv preprint arXiv:2405.02696*, 2024.

Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*, 2018.

Zhen Li, Kim-Hui Yap, and Bai-Ying Lei. A new blind robust image watermarking scheme in svd-dct composite domain. In *ICIP*, 2011.

Wen-Nung Lie and Li-Chun Chang. Robust and high-quality time-domain audio watermarking based on low-frequency amplitude modification. *IEEE Trans. Multim.*, 8(1):46–59, 2006. doi: 10.1109/TMM.2005.861292. URL https://doi.org/10.1109/TMM.2005.861292.

Aiwei Liu, Leyi Pan, Xuming Hu, Shiao Meng, and Lijie Wen. A semantic invariant robust watermark for large language models. *arXiv preprint arXiv:2310.06356*, 2023a.

Aiwei Liu, Leyi Pan, Yijian Lu, Jingjing Li, Xuming Hu, Xi Zhang, Lijie Wen, Irwin King, Hui Xiong, and Philip Yu. A survey of text watermarking in the era of large language models. *ACM Computing Surveys*, 57 (2):1–36, 2024.

Chang Liu, Jie Zhang, Han Fang, Zehua Ma, Weiming Zhang, and Nenghai Yu. Dear: A deep learning-based audio re-recording resilient watermarking. In Brian Williams, Yiling Chen, and Jennifer Neville (eds.), *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pp. 13201–13209. AAAI Press, 2023b. doi: 10.1609/aaai.v37i11.26550.

Yepeng Liu and Yuheng Bu. Adaptive text watermark for large language models. *arXiv preprint arXiv:2401.13927*, 2024.

Zhenghui Liu, Yuankun Huang, and Jiwu Huang. Patchwork-based audio watermarking robust against desynchronization and recapturing attacks. *IEEE Trans. Inf. Forensics Secur.*, 14(5):1171–1180, 2019. doi: 10.1109/TIFS.2018.2871748. URL https://doi.org/10.1109/TIFS.2018.2871748.

Xiyang Luo, Ruohan Zhan, Huiwen Chang, Feng Yang, and Peyman Milanfar. Distortion agnostic deep watermarking. In *CVPR*, 2020.

Jingwei Ma, Lucy Chai, Minyoung Huh, Tongzhou Wang, Ser-Nam Lim, Phillip Isola, and Antonio Torralba. Totems: Physical objects for verifying visual integrity. In *European Conference on Computer Vision*, pp. 164–180. Springer, 2022a.

Rui Ma, Mengxi Guo, Yi Hou, Fan Yang, Yuan Li, Huizhu Jia, and Xiaodong Xie. Towards blind watermarking: Combining invertible and non-invertible mechanisms. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 1532–1542, 2022b.

Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do gans leave artificial fingerprints? In *2019 IEEE conference on multimedia information processing and retrieval (MIPR)*, pp. 506–511. IEEE, 2019.

Hasan Mesut Meral, Bülent Sankur, A Sumru Özsoy, Tunga Güngör, and Emre Sevinç. Natural language watermarking via morphosyntactic alterations. *Computer Speech & Language*, 23(1):107–125, 2009.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *arXiv*, 2023.

Nicolas M Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger. Does audio deepfake detection generalize? *arXiv preprint arXiv:2203.16263*, 2022.

Iynkaran Natgunanathan, Yong Xiang, Yue Rong, Wanlei Zhou, and Song Guo. Robust patchwork-based embedding and decoding scheme for digital audio watermarking. *IEEE Trans. Speech Audio Process.*, 20 (8):2232–2239, 2012. doi: 10.1109/TASL.2012.2199111. URL https://doi.org/10.1109/TASL.2012.2199111.

Zhicheng Ni, Yun-Qing Shi, Nirwan Ansari, and Wei Su. Reversible data hiding. *IEEE Transactions on circuits and systems for video technology*, 2006.

Sophie J Nightingale and Hany Farid. Ai-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, 119(8):e2120481119, 2022.

Nikos Nikolaidis and Ioannis Pitas. Robust image watermarking in the spatial domain. *Signal processing*, 1998.

Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

Junlin Ouyang, Gouenou Coatrieux, Beijing Chen, and Huazhong Shu. Color image watermarking based on quaternion fourier transform and improved uniform log-polar mapping. *Computers & Electrical Engineering*, 2015.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Patrick O'Reilly, Zeyu Jin, Jiaqi Su, and Bryan Pardo. Maskmark: Robust neuralwatermarking for real and synthetic speech. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4650–4654. IEEE, 2024.

Leyi Pan, Aiwei Liu, Zhiwei He, Zitian Gao, Xuandong Zhao, Yijian Lu, Binglin Zhou, Shuliang Liu, Xuming Hu, Lijie Wen, et al. Markllm: An open-source toolkit for llm watermarking. *arXiv preprint arXiv:2405.10051*, 2024a.

Minzhou Pan, Yi Zeng, Xue Lin, Ning Yu, Cho-Jui Hsieh, Peter Henderson, and Ruoxi Jia. Jigmark: A black-box approach for enhancing image watermarks against diffusion model edits. *arXiv preprint arXiv:2406.03720*, 2024b.

Zoë Papakipos, Giorgos Tolias, Tomas Jenicek, Ed Pizzi, Shuhei Yokoo, Wenhao Wang, Yifan Sun, Weipu Zhang, Yi Yang, Sanjay Addicam, et al. Results and findings of the 2021 image similarity challenge. In *NeurIPS 2021 Competitions and Demonstrations Track*, pp. 1–12. PMLR, 2022.

Kosta Pavlović, Slavko Kovačević, Igor Djurović, and Adam Wojciechowski. Robust speech watermarking by a jointly trained embedder and detector using a dnn. *Digital Signal Processing*, 122:103381, 2022.

Florent Perronnin, Yan Liu, Jorge Sánchez, and Hervé Poirier. Large-scale image retrieval with compressed fisher vectors. In *CVPR*, pp. 3384–3391. IEEE, 2010.

Julien Piet, Chawin Sitawarin, Vivian Fang, Norman Mu, and David Wagner. Mark my words: Analyzing and evaluating language model watermarks. *arXiv preprint arXiv:2312.00273*, 2023.

Alessandro Piva, Mauro Barni, Franco Bartolini, and Vito Cappellini. Dct-based watermark recovering without resorting to the uncorrupted original image. In *Proceedings of international conference on image processing*, volume 1, pp. 520–523. IEEE, 1997.

Wenjie Qu, Dong Yin, Zixin He, Wei Zou, Tianyang Tao, Jinyuan Jia, and Jiaheng Zhang. Provably robust multi-bit watermarking for ai-generated text via error correction code. *arXiv preprint arXiv:2401.16820*, 2024.

Xinghua Qu, Xiang Yin, Pengfei Wei, Lu Lu, and Zejun Ma. Audioqr: Deep neural audio watermarks for qr code. *IJCAI*, 2023.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.

Yanzhen Ren, Hongcheng Zhu, Liming Zhai, Zongkun Sun, Rubing Shen, and Lina Wang. Who is speaking actually? robust and versatile speaker traceability for voice conversion. *arXiv preprint arXiv:2305.05152*, 2023.

Ahmad Rezaei, Mohammad Akbari, Saeed Ranjbar Alvar, Arezou Fatemi, and Yong Zhang. Lawa: Using latent space for in-generation image watermarking. *arXiv preprint arXiv:2408.05868*, 2024.

Ronald L Rivest, Adi Shamir, and Leonard Adleman. A method for obtaining digital signatures and public-key cryptosystems. *Communications of the ACM*, 21(2):120–126, 1978.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Mehrdad Saberi, Vinu Sankar Sadasivan, Keivan Rezaei, Aounon Kumar, Atoosa Chegini, Wenxiao Wang, and Soheil Feizi. Robustness of ai-image detectors: Fundamental limits and practical attacks. *ICLR*, 2024.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*, 2023.

Md Sahidullah, Tomi Kinnunen, and Cemal Hanilçi. A comparison of features for synthetic speech detection. *ISCA (the International Speech Communication Association)*, 2015.

Davide Salvi, Paolo Bestagini, and Stefano Tubaro. Towards frequency band explainability in synthetic speech detection. In *2023 31st European Signal Processing Conference (EUSIPCO)*, pp. 620–624. IEEE, 2023.

Davide Salvi, Temesgen Semu Balcha, Paolo Bestagini, and Stefano Tubaro. Listening between the lines: Synthetic speech detection disregarding verbal content. *arXiv preprint arXiv:2402.05567*, 2024.

Robin San Roman, Pierre Fernandez, Hady Elsahar, Alexandre D´efossez, Teddy Furon, and Tuan Tran. Proactive detection of voice cloning with localized watermarking. In *International Conference on Machine Learning*, 2024.

Robin San Roman, Pierre Fernandez, Antoine Deleforge, Yossi Adi, and Romain Serizel. Latent watermarking of audio generative models. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

Tom Sander, Pierre Fernandez, Alain Durmus, Teddy Furon, and Matthijs Douze. Watermark anything with localized messages. In *International Conference on Learning Representations*, 2025.

Ayush Sarkar, Hanlin Mai, Amitabh Mahapatra, Svetlana Lazebnik, David A Forsyth, and Anand Bhattad. Shadows don't lie and lines can't bend! generative models don't know projective geometry... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28140–28149, 2024.

Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. De-fake: Detection and attribution of fake images generated by text-to-image diffusion models. *arXiv preprint arXiv:2210.06998*, 2022.

Cuihua Shen, Mona Kasra, Wenjing Pan, Grace A Bassett, Yining Malloch, and James F O'Brien. Fake images: The effects of source, intermediary, and digital media literacy on contextual assessment of image credibility online. *New media & society*, 21(2):438–463, 2019.

M Hassan Shirali-Shahreza and Mohammad Shirali-Shahreza. A new synonym text steganography. In *2008 international conference on intelligent information hiding and multimedia signal processing*, pp. 1524–1526. IEEE, 2008.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*, 2019.

Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. Ai model gpt-3 (dis) informs us better than humans. *Science Advances*, 9(26):eadh1850, 2023.

Zhaopin Su, Guofu Zhang, Feng Yue, Lejie Chang, Jianguo Jiang, and Xin Yao. Snr-constrained heuristics for optimizing the scaling parameter of robust audio watermarking. *IEEE Trans. Multim.*, 20(10):2631–2644, 2018. doi: 10.1109/TMM.2018.2812599. URL https://doi.org/10.1109/TMM.2018.2812599.

Michael J Swain and Dana H Ballard. Color indexing. *International journal of computer vision*, 7(1):11–32, 1991.

Yuan-Yen Tai and Mohamed F Mansour. Audio watermarking over the air with modulated self-correlation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2452–2456. IEEE, 2019.

Hemlata Tak, Jee-weon Jung, Jose Patino, Madhu Kamble, Massimiliano Todisco, and Nicholas Evans. End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection. *arXiv preprint arXiv:2107.12710*, 2021a.

Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. End-to-end anti-spoofing with rawnet2. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6369–6373. IEEE, 2021b.

Matthew Tancik, Ben Mildenhall, and Ren Ng. Stegastamp: Invisible hyperlinks in physical photographs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2117–2126, 2020.

Mercan Topkara, Cuneyt M Taskiran, and Edward J Delp III. Natural language watermarking. In *Security, Steganography, and Watermarking of Multimedia Contents VII*, volume 5681, pp. 441–452. SPIE, 2005.

Mercan Topkara, Giuseppe Riccardi, Dilek Hakkani-Tür, and Mikhail J Atallah. Natural language watermarking: Challenges in building a practical system. In *Security, Steganography, and Watermarking of Multimedia Contents VIII*, volume 6072, pp. 106–117. SPIE, 2006a.

Mercan Topkara, Umut Topkara, and Mikhail J Atallah. Words are not enough: sentence level natural language watermarking. In *Proceedings of the 4th ACM international workshop on Contents protection and security*, pp. 37–46, 2006b.

Umut Topkara, Mercan Topkara, and Mikhail J Atallah. The hiding virtues of ambiguity: quantifiably resilient watermarking of natural language text through synonym substitutions. In *Proceedings of the 8th workshop on Multimedia and security*, pp. 164–174, 2006c.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Honai Ueoka, Yugo Murawaki, and Sadao Kurohashi. Frustratingly easy edit-based linguistic steganography with a masked language model. *arXiv preprint arXiv:2104.09833*, 2021.

Matthieu Urvoy, Dalila Goudia, and Florent Autrusseau. Perceptual dft watermarking with improved detection and robustness to geometrical distortions. *IEEE Transactions on Information Forensics and Security*, 2014.

USA. Ensuring safe, secure, and trustworthy ai. https://www.whitehouse.gov/wp-content/uploads/2023/07/Ensuring-Safe-Secure-and-Trustworthy-AI.pdf, July 2023. Accessed: [july 2023].

Alina Valyaeva. Ai has already created as many images as photographers have taken in 150 years, 2023. URL https://journal.everypixel.com/ai-image-statistics. Accessed on July 18, 2024.

Ron G Van Schyndel, Andrew Z Tirkel, and Charles F Osborne. A digital watermark. In *Proceedings of 1st international conference on image processing*, volume 2, pp. 86–90. IEEE, 1994.

Christoforos Vasilatos, Manaar Alam, Talal Rahwan, Yasir Zaki, and Michail Maniatakos. Howkgpt: Investigating the detection of chatgpt-generated university student homework through context-aware perplexity analysis. *arXiv preprint arXiv:2305.18226*, 2023.

Ashish Venugopal, Jakob Uszkoreit, David Talbot, Franz Josef Och, and Juri Ganitkevitch. Watermarking the outputs of structured prediction with an application in statistical machine translation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 1363–1372, 2011.

Vedran Vukotić, Vivien Chappelier, and Teddy Furon. Are deep neural networks good for blind image watermarking? In *WIFS*, 2018.

Vedran Vukotić, Vivien Chappelier, and Teddy Furon. Are classification deep neural networks good for blind image watermarking? *Entropy*, 2020.

Wenbo Wan, Jun Wang, Yunming Zhang, Jing Li, Hui Yu, and Jiande Sun. A comprehensive survey on robust image watermarking. *Neurocomputing*, 488:226–247, 2022.

Avery Wang et al. An industrial strength audio search algorithm. In *Ismir*, volume 2003, pp. 7–13. Washington, DC, 2003.

Sheng-Yu Wang, Oliver Wang, Andrew Owens, Richard Zhang, and Alexei A Efros. Detecting photoshopped faces by scripting photoshop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10072–10081, 2019.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8695–8704, 2020.

Wenhao Wang, Yifan Sun, and Yi Yang. A benchmark and asymmetrical-similarity learning for practical image copy detection. *arXiv preprint arXiv:2205.12358*, 2022.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, et al. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. *arXiv preprint arXiv:2305.14902*, 2023a.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22445–22455, 2023b.

Bingyang Wen and Sergul Aydore. Romark: A robust watermarking system using adversarial training. *arXiv preprint arXiv:1910.01221*, 2019.

Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. *arXiv preprint arXiv:2305.20030*, 2023.

Eric Wengrowski and Kristin Dana. Light field messaging with deep photographic steganography. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1515–1524, 2019.

Alex Wilson and Andrew D Ker. Avoiding detection on twitter: embedding strategies for linguistic steganography. *Electronic Imaging*, 28:1–9, 2016.

Keith Winstein. Lexical steganography through adaptive modulation of the word choice hash. *Unpublished. http://www.imsa.edu/keithw/tlex*, 1998.

Hanzhou Wu, Gen Liu, Yuwei Yao, and Xinpeng Zhang. Watermarking neural networks with watermarked images. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(7):2591–2601, 2020.

Shiqiang Wu, Jie Liu, Ying Huang, Hu Guan, and Shuwu Zhang. Adversarial audio watermarking: Embedding watermark into deep feature. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 61–66. IEEE, 2023.

Xiang-Gen Xia, Charles G Boncelet, and Gonzalo R Arce. Wavelet transform based watermark for digital images. *Optics Express*, 1998.

Lingyun Xiang, Xinhui Wang, Chunfang Yang, and Peng Liu. A novel linguistic steganography based on synonym run-length encoding. *IEICE transactions on Information and Systems*, 100(2):313–322, 2017.

Yong Xiang, Iynkaran Natgunanathan, Song Guo, Wanlei Zhou, and Saeid Nahavandi. Patchwork-based audio watermarking method robust to de-synchronization attacks. *IEEE ACM Trans. Audio Speech Lang. Process.*, 22(9):1413–1423, 2014. doi: 10.1109/TASLP.2014.2328175. URL https://doi.org/10.1109/TASLP.2014.2328175.

Yong Xiang, Iynkaran Natgunanathan, Dezhong Peng, Guang Hua, and Bo Liu. Spread spectrum audio watermarking using multiple orthogonal PN sequences and variable embedding strengths and polarities. *IEEE ACM Trans. Audio Speech Lang. Process.*, 26(3):529–539, 2018. doi: 10.1109/TASLP.2017.2782487. URL https://doi.org/10.1109/TASLP.2017.2782487.

Xiaojun Xu, Yuanshun Yao, and Yang Liu. Learning to watermark llm-generated text via reinforcement learning. *arXiv preprint arXiv:2403.10553*, 2024.

Xianjun Yang, Wei Cheng, Yue Wu, Linda Petzold, William Yang Wang, and Haifeng Chen. Dna-gpt: Divergent n-gram analysis for training-free detection of gpt-generated text. *arXiv preprint arXiv:2305.17359*, 2023.

Zijin Yang, Kai Zeng, Kejiang Chen, Han Fang, Weiming Zhang, and Nenghai Yu. Gaussian shading: Provable performance-lossless image watermarking for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12162–12171, 2024.

KiYoon Yoo, Wonhyuk Ahn, Jiho Jang, and Nojun Kwak. Robust multi-bit natural language watermarking through invariant features. *arXiv preprint arXiv:2305.01904*, 2023a.

KiYoon Yoo, Wonhyuk Ahn, and Nojun Kwak. Advancing beyond identification: Multi-bit watermark for language models. *arXiv preprint arXiv:2308.00221*, 2023b.

Chong Yu. Attention based data hiding with generative adversarial networks. In *AAAI*, 2020.

Ning Yu, Larry S Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7556–7566, 2019.

Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14448–14457, 2021.

Ning Yu, Vladislav Skripniuk, Dingfan Chen, Larry Davis, and Mario Fritz. Responsible disclosure of generative models using scalable fingerprinting. In *International Conference on Learning Representations (ICLR)*, 2022.

Aditi Zear, Amit Kumar Singh, and Pardeep Kumar. A proposed secure multiple watermarking technique based on dwt, dct and svd for application in medicine. *Multimedia tools and applications*, 77:4863–4882, 2018.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. *NeurIPS*, 2019.

Chunlei Zhang, Chengzhu Yu, and John HL Hansen. An investigation of deep learning frameworks for speaker verification antispoofing. *IEEE Journal of Selected Topics in Signal Processing*, 11(4):684–694, 2017.

Honglei Zhang, Hu Wang, Yuanzhouhan Cao, Chunhua Shen, and Yidong Li. Robust watermarking using inverse gradient attention. *arXiv preprint arXiv:2011.10850*, 2020.

Kevin Alex Zhang, Alfredo Cuesta-Infante, Lei Xu, and Kalyan Veeramachaneni. Steganogan: High capacity image steganography with gans. *arXiv preprint arXiv:1901.03892*, 2019a.

Kevin Alex Zhang, Lei Xu, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Robust invisible video watermarking with attention. *arXiv preprint arXiv:1909.01285*, 2019b.

Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *2019 IEEE international workshop on information forensics and security (WIFS)*, pp. 1–6. IEEE, 2019c.

Xuanyu Zhang, Runyi Li, Jiwen Yu, Youmin Xu, Weiqi Li, and Jian Zhang. Editguard: Versatile image watermarking for tamper localization and copyright protection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11964–11974, 2024.

Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2185–2194, 2021.

Xuandong Zhao, Sam Gunn, Miranda Christ, Jaiden Fairoze, Andres Fabrega, Nicholas Carlini, Sanjam Garg, Sanghyun Hong, Milad Nasr, Florian Tramer, et al. Sok: Watermarking for ai-generated content. *arXiv preprint arXiv:2411.18479*, 2024.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Ngai-Man Cheung, and Min Lin. A recipe for watermarking diffusion models. *arXiv preprint arXiv:2303.10137*, 2023.

Junzuo Zhou, Jiangyan Yi, Tao Wang, Jianhua Tao, Ye Bai, Chu Yuan Zhang, Yong Ren, and Zhengqi Wen. Traceablespeech: Towards proactively traceable text-to-speech with watermarking. *arXiv preprint arXiv:2406.04840*, 2024.

Mingyi Zhou, Xiang Gao, Jing Wu, John Grundy, Xiao Chen, Chunyang Chen, and Li Li. Modelobfuscator: Obfuscating model information to protect deployed ml-based systems. In *Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis*, pp. 1005–1017, 2023.

Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. Hidden: Hiding data with deep networks. In *ECCV*, 2018.