КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 4 (32), 2026
ISSN 2663 – 4023
DOI 10.28925/2663-4023.2026.32.1082
УДК 004.8

Mykhailo Marchuk
Phd student at the Department of Informational Security
Vinnytsia National Technical University
ORCID: 0009-0003-4773-6541
smoke222catches@gmail.com

Vitalii Lukichov
Associate Professor at the Department of Informational Security
Vinnytsia National Technical University
ORCID: 0000-0002-3423-5436
lukichov.vitalyi@vntu.edu.ua

THE SURVEY ON WATERMARKING METHODS FOR PROACTIVE DEFENSE AGAINST DEEPFAKE

Abstract. As generative models advance, deepfake content is becoming indistinguishable from reality and passive forensic detection methods are becoming increasingly ineffective. The misuse of generative tools provide for adversaries opportunities for social engineering, disinformation campaigns and fraud. This requires a new class of forensics tools based on the preemptive marking of authentic content in order to defend it from being used for deepfake media generation or disinformation campaigns. In this survey we provide a comprehensive analysis of watermarking solutions for the purpose of proactive defense from deepfake. We identified most of the existing deepfake watermarking solutions in literature and provided taxonomy for them. Also we identified core metrics and datasets for training deep learning models for proactive defense watermarking. We make quantitative and qualitative comparisons of existing solutions, their methods, metrics and purposes. In the end we provide a summary of open problems and challenges in the field. This survey lays a foundation for future development of proactive deepfake defense methods and policies for generative AI compliance.

Keywords: deepfake; watermarking; image forensics; deep learning; steganography; information security

INTRODUCTION

The rapid development of generative artificial intelligence has transformed how digital content is created, distributed, and consumed. Foundation models capable of producing photorealistic images, videos, and audio have enabled a wide range of beneficial applications for creativity, accessibility and science. However, the same capabilities have introduced significant challenges for privacy, security, and digital trust. In particular, advances in synthetic media generation - commonly referred to as deepfakes - have lowered the cost and technical barrier for producing highly realistic facial manipulations that are difficult for humans to detect.

Synthetic media has been leveraged to orchestrate large-scale disinformation campaigns [30], conduct social engineering and business-email compromise attacks [31], [32], impersonate public officials, and manipulate biometric authentication systems. Recent incidents across the world demonstrate the severity of the threat. In 2024, a multinational firm in Hong Kong was defrauded of approximately 25 million USD after an employee was tricked by a deepfake video call impersonating senior management [54]. Similar cases have been reported in Singapore [55], where attackers used deepfaked voices and videos to authorize

fraudulent financial transfers or obtain sensitive information. These incidents illustrate a growing class of information-security attacks that rely not on software vulnerabilities, but on realistic synthetic media capable of undermining human and organizational trust.

Problem formulation. Despite ongoing progress in forensic analysis, passive detection methods are becoming increasingly ineffective as standalone defenses [33], [34]. These techniques rely on identifying statistical artifacts, generative fingerprints, or inconsistencies in synthetic content. Yet rapid advances in generative models - higher-resolution diffusion architectures, adversarial training, and model fine-tuning - continually reduce these detectable traces. As a result, detectors degrade quickly as new model families emerge. Moreover, passive detectors frequently suffer from poor cross-model and cross-domain generalization: a detector trained on one class of generators or manipulation types often fails when confronted with unseen architectures, compression pipelines, or low-quality real-world videos. This growing gap between evolving generative capabilities and static forensic models underscores the need for proactive security mechanisms that embed verifiable provenance information at creation time rather than attempting to infer it after the fact.

In response to these limitations, proactive detection approaches have emerged as a promising alternative. Rather than attempting to infer authenticity from the statistical remnants of a generative process, proactive methods embed verifiable signals into synthetic content at the moment of creation. These signals enable provenance verification, generator source tracing and manipulations detection even when the underlying generative techniques evolve. Because proactive mechanisms do not depend on fragile visual artifacts, they are more resilient to cross-model variation, compression pipelines, and distribution shifts that typically undermine passive detectors. Among these approaches, watermarking has gained particular attention due to its compatibility with both centralized and decentralized generation workflows, its ability to operate at scale, and its suitability for integration into modern generative architectures.

Analysis of recent studies and publications. Mirsky and Lee provided a broad review of creation and detection of deepfakes in which they also established classification of deepfake generation methods [52]. Nguyen-Le et al. propose a comprehensive survey on proactive defense methods, specifically on disruption of generation pipeline and watermarking [53]. Yu et al. focus on both passive and proactive approaches but specifically in video modality [54]. Ben Jabra and Ben Farah made a comprehensive survey on deep learning based watermarking methods [55].

The goal of the paper. This survey examines design principles, security guarantees, threat models, and current limitations of watermarking-based proactive defenses in the context of rapidly improving deepfake generation methods.

Our contributions are:

- A literature review of proactive watermarking methods against deepfake malicious usage;
- The taxonomy of deepfake watermarking solutions;
- The definition of current challenges and limitations for proactive deepfake watermarking.

PRELIMINARIES

We use “deepfake” as an umbrella term that covers both forged media and the computational tools used to generate them. Deepfake media typically results from altering facial features, expressions, identity attributes, or entire scenes in images, audio, or video to produce content that appears authentic. Deepfake generation tools are software systems – often released as standalone applications, open-source frameworks, or integrated model

pipelines – that automate this manipulation process. These tools rely on artificial neural networks, particularly deep generative architectures such as GANs, autoencoders, and diffusion models, because such networks are capable of learning complex, high-dimensional distributions of human appearance and motion. By training on a large corpora of real facial or audiovisual data, these models learn to synthesize highly realistic outputs and to transfer identity, expression, or style from one subject to another. This neural-network-driven capacity for precise, photorealistic synthesis is what enables modern deepfakes to closely approximate real human behavior and thereby pose significant challenges for detection and verification.

Deepfakes can be broadly grouped into four types that reflect the main goals of contemporary face-generation systems. The first is face swapping, in which the identity of one person is transferred onto the facial structure of another. The second is face reenactment, which alters a person's facial movements or expressions to imitate the performance of a different actor. A third category is talking-face generation, where a synthetic speaking video is produced from audio or text, enabling realistic lip motion and head dynamics without requiring the subject's actual footage. The fourth category is facial attribute editing, which selectively modifies specific properties of a face – such as age, makeup, hairstyle, or emotional expression – while keeping the individual recognizable. These four classes capture the dominant forms of visual manipulation seen in current deepfake systems and provide a functional basis for discussing both the threat landscape and the effectiveness of proactive defense mechanisms such as watermarking.

Visual deepfake production begins by training generative neural networks to model realistic facial appearance, motion, and context, and then applying these models to manipulate or synthesize imagery. Early foundational methods used latent-variable frameworks such as the Variational Autoencoder (VAE) introduced by Diederik P. Kingma and Max Welling [36] to encode images into a compact latent space and decode back to images. In 2014, the landmark Generative Adversarial Network (GAN) framework proposed by Ian J. Goodfellow et al. [37] enabled adversarial training of a generator and discriminator to create highly realistic imagery. More recently, the Denoising Diffusion Probabilistic Model (DDPM) proposed by Jonathan Ho, Ajay Jain and Pieter Abbeel [38] leveraged a noise-addition forward process and a learned reverse denoising process to produce high-fidelity images.

These architectures (and their many variants) are then adapted for deepfake workflows: identity transfer, expression manipulation, motion synthesis and so on. In practise a generator network takes as input a source (image, video, or audio) and a target (face, pose, identity) and produces manipulated output; optionally a refinement network or discriminator ensures realism. Post-processing (e.g., blending, colour correction, temporal smoothing) often completes the pipeline. Because these models continually improve (higher resolution, fewer artifacts, stronger temporal coherence) the resulting deepfakes become more difficult to detect.

Passive deepfake detection. Passive deepfake detection seeks to identify manipulated content after its generation by analyzing statistical, semantic, temporal, or forensic artifacts left in images or videos. Early methods focused on handcrafted features – for example, detecting unnatural eye blinking, facial warping or inconsistent lighting – but around 2019-2021 deep-learning models (especially convolutional neural networks) became dominant. A notable example is Learning Self-Consistency for Deepfake Detection by Zhao et al. [39], which introduced a "source-feature inconsistency" framework and achieved strong performance across in- and cross-dataset settings. Other influential methods utilize GAN-fingerprint detection and convolutional-trace extraction [40] to detect residual traces of generation pipelines. While passive methods have achieved impressive results in controlled

settings, their generalization to unseen generative models, compression settings, and "in-the-wild" data remains a major challenge [41]. In practice, passive detectors may falter when new deepfake generation techniques eliminate specific artifacts, or when the detector is confronted with a generator it never saw in training – thus motivating proactive approaches embedded at creation time.

Steganography and watermarking. Steganography provides the algorithmic and conceptual foundations from which modern digital watermarking emerged. While steganography and watermarking are often distinguished by their goals-concealing a secret message versus embedding an identifiable and typically detectable mark-the underlying techniques, signal-processing primitives, and security assumptions developed in classical steganography directly shaped the watermarking methods that later became central to proactive deepfake defense. For this reason, it is useful to briefly survey classical steganographic work not as a separate discipline, but as the historical substrate from which contemporary robust watermarking evolved.

Early spatial-domain steganography introduced the basic paradigm of embedding information through controlled, imperceptible modifications to a host signal. Least significant bit (LSB) substitution, described in early taxonomies such as that of Bender et al. [42], demonstrated how small pixel-level perturbations can carry information without perceptual degradation. The optimal pixel adjustment process (OPAP) of Chan and Cheng [43] refined this idea by explicitly minimizing distortion after embedding, providing one of the first systematic methods for optimizing imperceptibility under a payload constraint. These works introduced two principles that watermarking inherits directly: embedding strength must remain below a perceptual threshold, and embedding patterns should minimize statistical artifacts that enable detection.

Transform-domain steganography further established techniques that later became central to watermarking. JPEG-domain algorithms such as JSteg showed how embedding into discrete cosine transform (DCT) coefficients provides resilience against re-compression, while the F5 algorithm by Westfeld [45] and the OutGuess framework by Provos [46] demonstrated the importance of statistical preservation-embedding must maintain global coefficient distributions to resist steganalysis. These methods pioneered the idea that embedding should be aligned with the statistical structure of the host representation, a principle mirrored in robust watermarking methods that target perceptually important but redundancy-rich spectral components.

Watermarking as a distinct field took shape when these steganographic insights were repurposed for robustness rather than secrecy. The foundational spread-spectrum framework of Cox et al. [47] extended the idea of embedding noise-like patterns into transform coefficients but explicitly analyzed the watermark-attacker interaction using detection theory. This reframed data hiding as a communication problem, where the watermark is a signal transmitted through a channel subject to adversarial distortion. Subsequent work by Podilchuk and Zeng [48] introduced image-adaptive watermarking driven by human visual system (HVS) models, adjusting watermark strength according to local texture and masking to jointly optimize imperceptibility and robustness. Wolfgang, Podilchuk, and Delp [49] further developed perceptual watermarking for images and video, systematizing HVS-based design rules and demonstrating practical schemes for copyright marking and tamper detection. Together with numerous DCT – and wavelet-domain schemes from the late 1990s, these works established robustness under compression, filtering, and geometric distortions as primary design goals for robust watermarking [49].

A second major line of classical watermarking – quantization index modulation (QIM), introduced by Chen and Wornell [51] – formalized embedding as modulation of host coefficients according to quantizer "cosets." This method inherits the steganographic principle of controlled coefficient manipulation but adds an information-theoretic robustness analysis, showing that watermarking can be modeled as communication with side information and optimized under explicit distortion constraints. QIM and its variants remain structurally important for modern proactive watermarking, including watermarking for synthetic media, because they directly model an adaptive adversary attempting removal.

Deep learning based steganography and watermarking. Deep-learning approaches have begun to transform watermarking by leveraging convolutional neural networks (CNNs), generative adversarial networks (GANs), and end-to-end embedding/extraction architectures to learn embedding strategies that optimize invisibility, robustness, and capacity simultaneously. For instance, the work by Xin Zhong and Shih [50] proposes a fully automated deep CNN-based image watermarking pipeline that embeds a watermark and extracts it under distortions without explicit attack modelling. In model-level watermarking, Nagai et al. [44] introduced a method for embedding watermarks into deep neural network parameters themselves, thereby extending watermarking beyond signal carriers into learned models. These advances illustrate a shift: instead of hand-crafting embedding rules in spatial or transform domains, deep-learning methods can jointly learn embedding/extraction, simulate attacks during training, and optimize for multiple objectives. However, this also introduces new concerns such as over-fitting to known attack types, interpretability of embedding, and generalization to unseen transformations – challenges that remain active research questions in proactive watermarking.

WATERMARKING PROPERTIES AND METRICS

Imperceptability. We define imperceptability as the ability of watermarked images to be indistinguishable from original in visual terms while preserving the informational capacity of watermark. In most real-life cases the visibility of a watermark is not desirable. If the watermark is perceptible enough to be noticed then it could lead to disruption of its semantics and possibly invalidate the initial purpose of content.

The most widely used evaluation metrics for imperceptibility are peak signal-to-noise-ratio (PSNR) and structural similarity index (SSIM).

PSNR represents how the watermarked image is different from original and defined as:

$$PSNR = 20 \times \log_{10}\left(\frac{MAX_I}{\sqrt{MSE}}\right) \quad (1)$$

Where MAX_I is the maximum possible pixel value of the image and MSE is mean square error which is defined as:

$$MSE = \frac{1}{m \times n}\sum_{i=1}^{m}\sum_{j=1}^{n}\left[I(i,j) - K(i,j)\right]^2 \quad (2)$$

For two $m \times n$ monochrome images I and K where $1 \le i \le m$, $1 \le j \le n$[15].

SSIM represents similarity between original and watermarked image and defined as:

$$SSIM(x,y) = \frac{\left(2\mu_x\mu_y + C_1\right)\left(2\delta_{xy} + C_2\right)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\delta_x^2 + \delta_y^2 + C_2\right)} \quad (3)$$

Where μ_x and μ_y denote the mean intensities of images x and y, respectively, while δ_x and δ_y represent their variances. δ_{xy} refers to the covariance between x and y, and C_1 and C_2 are constants introduced to prevent instability when the denominator is small.

Imperceptibility is an important property for all techniques of invincible watermarking.

Table 1

Summary of imperceptibility metrics of watermarking methods reported by authors

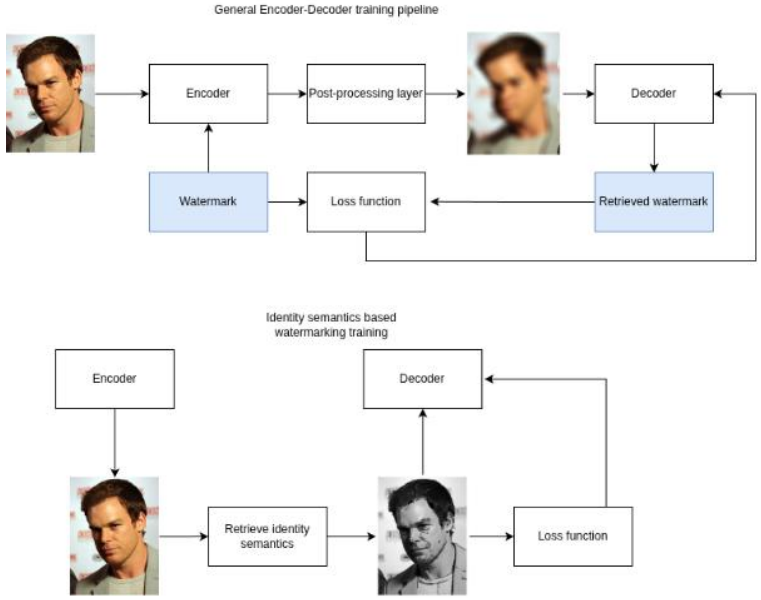
Method	Evaluation datasets	Metrics	
		PSNR (dB)	SSIM
FaceGuard [7]	FaceForensics++	-	0.94
	DFDC	-	0.94
	Trump-Cage	-	0.86
[8]	FFHQ	34.84	0.95
LampMark [9]	CelebA-HQ	45.45	0.995
	LFW	43.14	0.983
FaceProtect [10]	CelebA	42.73	0.989
DeepTag [11]		29.89	0.927
FakeTagger [12]	Encoded Images (StyleGAN)	35.21	0.948
FakeTracer [14]		44.12	0.98
[16]	CelebA-HQ (128×128)	47.39	0.993
	CelebA-HQ (256×256)	45.38	0.994
SepMark [17]	CelebA-HQ (128x128)	38.5112	0.9588
FaceSigns [21]	-	36.38	0.973
[24]	-	45.284	0.98
[26]	-	35.57	0.97
DiffMark [27]	LFW (128 × 128)	41.2869	0.9776
	LFW (256 × 256)	41.9572	0.9769

Robustness. We define robustness as the property that the watermark can be retrieved after the content alterations, both accidental and intentional. The former include compression, blurring, resizing, cropping and other changes that could be applied to the image, the latter in context of deepfakes are GAN transformations. The robustness is critical for real life usage of watermarking because images and videos can endure different changes during capturing, storing, converting, opening in apps or uploading to websites.

Robustness is not uniform across all types of alterations, some methods may remain robust to certain manipulations while being fragile to others. In our taxonomy the difference between deepfake provenance and semi-fragile watermarking methods is in how they differentiate for what manipulations they are robust to. Deepfake provenance watermarks are robust to GAN alterations while semi-fragile watermarks, in contrast, can be destroyed from it.

To achieve robustness during the training different alterations, that watermark should be robust to, are applied on training data. Methods that rely on encoder-decoder architecture have post-processing layers which apply alterations on watermarked images which are encoder output [7], [9], [11], [12], [16], [17]. Methods that are relying on identity features of content achieve robustness by implementing loss functions that focus specifically on identity

semantics that are not relying on other aspects of images [8], [10]. The general overview of how encoder-decoder architectures training achieves robustness presented in pic. 1.



Pic. 1. Training processes for encoder-decoder training pipelines

Robustness is linked to how well a watermark will be retrieved from the image, the metrics for it are checking the accuracy of detection and comparison with the original embedded message. Those metrics include bit error ratio (BET) [17], accuracy (ACC) [7, 8], F1-Score [8], false positive rate (FPR) [7], and false negative rate (FNR) [7]. Since the retrieving of watermarks is the main task for which the model is training, those metrics could also evaluate the general performance of the model.

Table 2

Summary on robustness metrics of watermarking methods reported by authors

Method	Evaluation datasets	Metrics*				
		FPR	FNR	ACC	F1	AUC
FaceGuard [7]	FaceForensics++	1.7	0.0	99.2		
	DFDC	1.1	0.0	99.5		
	Trump-Cage	1.5	0.0	97.7		
[8]	CelebA			98	0.98	0.99
	CelebA-HQ			99	0.98	0.99
LampMark [9]	CelebA-HQ (128x128)			87.23		0.9839
	CelebA-HQ (256x256)			87.21		0.9855
FaceProtect [10]	All Test Datasets (Mixed Identity/Attribute Manipulation)			0.96	0.95	
DeepTag [11]				Nearly 90%		
[16]	CelebA-HQ (128x128)					0.9866
	CelebA-HQ (256x256)					0.9897

Transferability. We define transferability as the ability of watermarks to be retrieved from output of systems that utilized watermarked content as training data. The concept of

transferability is different from robustness. While the task of robust watermark is to be retrieved after perturbations on original image, the transferable watermark could be fed as training data for generative model and then reappeared as output of it. In our taxonomy the methods of watermarking for deepfake generation models have this property.

Bitwise accuracy (B-Acc) is utilized as evaluation of transferability [13, 14]. B-Acc is defined as $BA(w,\widehat{w})$, the proportion of identical bits between embedded and recovered watermarks w and $\widehat{w}$:

$$BA(w,\widehat{w}) = \frac{1}{n}\sum_{i=1}^{n} 1\left[w_i = \widehat{w}_i\right] \tag{4}$$

Where n is the number of bits in the watermark.

Beside the B-Acc the authors of [13] utilize Fréchet Inception Distance (FID). FID measures the distance between a distribution of extracted features of ground truth and generated images, which correlates with general generation quality and realism. FID is utilized when traditional evaluation metrics are not suitable, such as in case of training data watermarking because the output of the generative model is new, unique data. FID is defined as:

$$FID = \left|\left|\mu_r - \mu_q\right|\right|^2 + Tr\left(\sum_r + \sum_g - 2\sqrt{\sum_r \sum_g}\right) \tag{5}$$

Where μ_r is a mean of extracted feature vectors of training set images and μ_g is for generated, and Tr denotes the trace of matrix [20].

Table 3

Summary on robustness metrics of watermarking methods

Method	Evaluation datasets	Metrics	
		B-Acc	FID
[13]	CelebA	1.000	1.15
FakeTracer [14]		0.94	7.03
[16]	CelebA-HQ (128x128)	0.9802	
	CelebA-HQ (256x256)	0.9698	

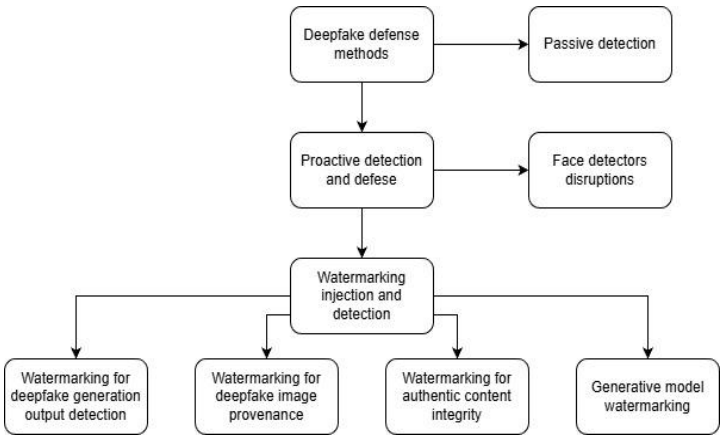
WATERMARKING TECHNIQUES FOR PROACTIVE DEEPFAKE DETECTION

Watermarking is a method of marking content with adding additional information into content itself without changing information in it. Watermarking can be served for copyrighting or tracking purposes. For example, artists that share their work as digital images want to add their name or logotype over content so it cannot be stolen or claimed by somebody else.

In this survey we focus on invincible watermarking for images which adds perturbations to images that are insignificant in terms of visual difference and can't be identified by human eye but still can be tracked by computer algorithms. This kind of watermarking can serve for the provenance purpose by providing an identification of content origin trace. In contrast with metadata-based provenance techniques, watermarks can be used to store provenance identification traces directly in the content, making it robust against metadata stripping or benign content modifications like compression or blurring. Also some of the methods can

serve for detecting if any breaking changes were applied for the image (like GAN-based manipulations for creating deepfakes).

We provide a taxonomy for watermarking-based provenance techniques that is visualized in pic. 2. Every individual category is discussed below.



Pic. 2. Taxonomy for proactive deepfake defense watermarking methods

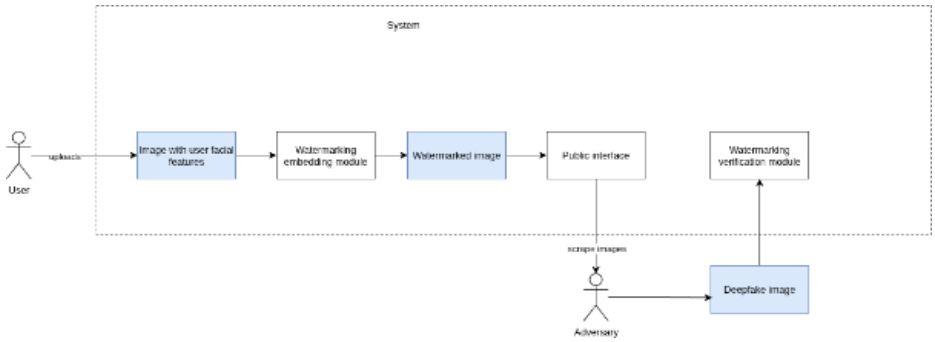
Deepfake generation output detection. We define the watermarking for deepfake generation output detection as a set of techniques to detect if an image is a result of face forgery. In other words, the purpose is to check if the image is the output of the deepfake generation pipeline.

The threat model for which this type of watermarking can be utilized is applied to systems where users share images with facial features to the public [11, 12]. For example, social media platforms. By making such content publicly accessible, platforms inadvertently provide adversaries with the opportunity to scrape these images and use them for deepfake generation.

Some methods allow to localize affected areas of images using watermarks that could be partially removed on transformations affected areas. By detecting if watermark is present at the face area the face forgery could be identified. Authors of WaterLo [25] propose a semifragile watermark that can be erased on parts of video where non benign manipulations were applied. FractalForensics [29] utilizes fractal watermarks which can be preserved on unaffected parts of images.

Deepfake image provenance. We define the watermarking for deepfake image provenance as a set of techniques for securing authentic images by applying watermarking on them in order to track Deepfake content that is created from the original tagged image. The main difference from the previous type is that watermarks contain information that could help track the original image or persona. Ideally, a watermark should be recoverable from both the original and the manipulated versions of an image, indicating whether any alterations have been applied [14], [16]; however, some methods could implement this mechanism without alterations detection [11], [12].

The watermarking is applied to marked content. If it’s being used for deepfake, it can be tracked (pic. 3). In those conditions this watermarking method should be robust to GAN alterations [11] and benign transformations like compression or resizing [12], [14], [16].



Pic. 3. Defense model for deepfake image provenance watermarking

Another threat model involves the usage of facial images not as input but as training data for generative models [14].

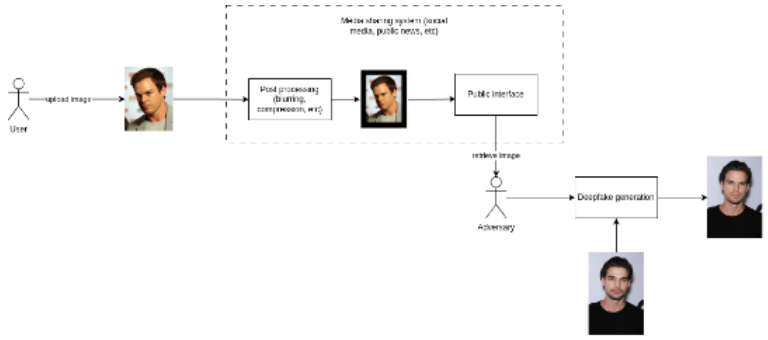
The drawback of most watermarking methods of this type is that recovered watermarks don't provide any information about if it is retrieved from original or deepfake image. As countermeasure to this the authors of [14] propose to use two types of watermarks on the single image, one of which is robust to GAN alterations and another is not. The unresolved problem is that there is no system for watermarks that can be preserved from both training and input images.

Some solutions rely on facial features of images. Authors of [22] utilize the visual hashing algorithm to extract facial features which are embedded as watermark into video backgrounds using pseudo-Zernike moments. Deepmark [23] transforms identity features into digital signatures which can be verified using public keys. Re-Mark [28] is capable of restoring the original facial identity from watermark in manipulated images.

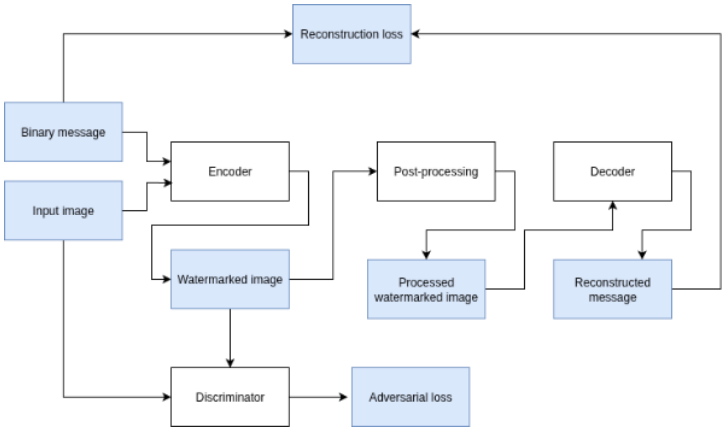
Authentic content integrity. We define the watermarking for content and provenance integrity confirmation as a technique for securing authentic content by applying a watermark on it in order to confirm its authenticity. The image is considered authentic if a watermark is detected. The absence of a watermark does not necessarily mean that the image is fake or manipulated by an adversary; it may simply indicate that it was not created by an authorized party (pic. 4). The main purpose of this method is to confirm authenticity and integrity of content but not to prove its forgery. This type of watermarking is useful for use cases where the maintenance of content authenticity has the highest priority. Examples include public media, social media or community-run services like Wikipedia.

Unlike metadata-based provenance techniques that verify the authenticity of media by strictly comparing its content with its log of manipulations, watermark aims to preserve provenance and trace it in conditions where content can be modified with benign alterations. The former are defined as alterations that can change the content but don't drastically change information in it. Examples can include compression, blurring or resizing.

The authors of [7], [21] proposed methods that utilized architecture that involves the usage of the post-processing layer during training. This layer applies image post-processing operations on the training data like JPEG compression, blurring, cropping or resizing. Decoders that are trained with this type of data are more robust to those manipulations while still fragile to GAN-powered face alterations (pic. 7). This provides a significant advantage compared to metadata-based provenance that trace in watermarking cannot be stripped without damaging the content.



Pic. 4. Threat model for authentic content integrity



Pic. 5. The training pipeline of encoder-decoder architecture which utilizes post-processing layer

In [8] and [10] the authors propose methods that embed watermarks into facial features. These methods rely on the rationale that realistic deepfake generation requires altering the facial features of the source image. Consequently, if a watermark is embedded within these features, it will be disrupted during deepfake generation, thereby preventing the authenticity of the resulting image from being verified.

The proposed approach in [8] uses a neural network with an encoder-decoder structure. An input face image is disentangled into an identity representation and attributes representation using two dedicated networks: identity encoder and attributes encoder. In [9] the watermark embedding and verification are separated into two networks where the former is GAN and the latter is U-Net.

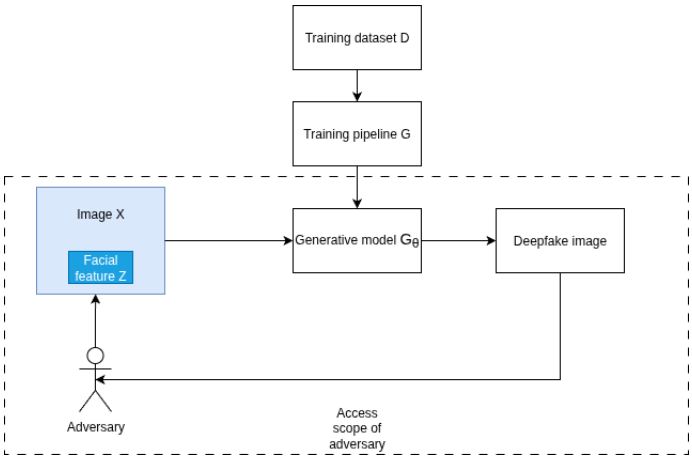
In [9] authors introduce the training-free landmark perceptual watermark for proactive Deepfake detection, a significant departure from previous deep watermarking methods that rely on extensive end-to-end training. The method uniquely exploits the structure-sensitive characteristics of Deepfake manipulations on facial landmarks. It devises a secure and confidential transformation pipeline that projects facial landmarks into binary landmark perceptual watermarks without needing additional training.

Deepfake generation models watermarking. This type of watermarking is applied for the dataset of generative model [13] in contrast to previously discussed methods for which watermarking is applied for input data. The main reason for the usage of this type of methods is to discourage the malicious usage of generative models since the output can be identified as generated. The threat model, for which this type of methods can be applied, involves a malicious usage of generative models like for deepfake generation. The applying of watermarking here should discourage the misusage since the output would be revealed as generated by model.

From the adversary point of view the attack is successful if an image X with facial features Z is fed to a generative model $G_0 = G(D)$ where G is model pipeline and D is a non-watermarked training dataset and the produced image $\widehat{X}$ satisfies (pic. 6):

$$\widehat{X} = G_0(Z),\ Det(\widehat{X}) = 0,\ s(X,\widehat{X}) \geq s_0 \tag{6}$$

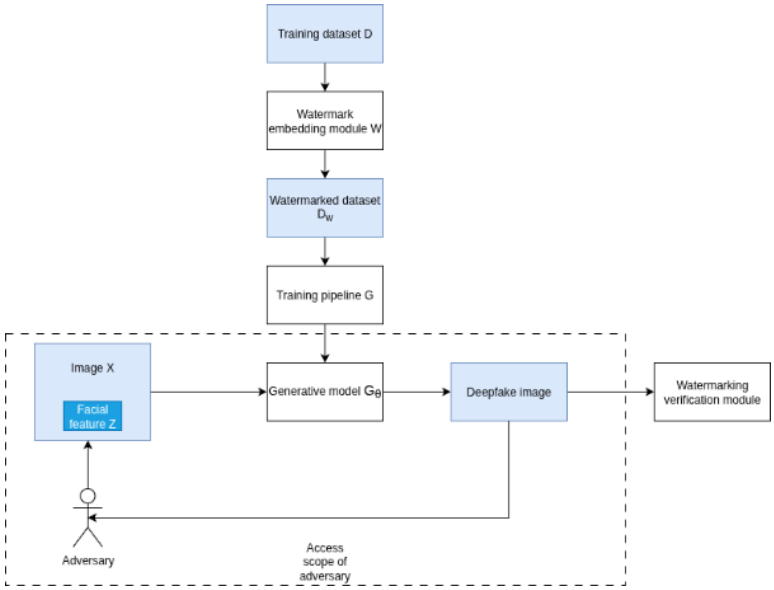
Where Det is a deepfake detector, s is perceptual similarity with threshold s_0.



Pic. 6. The threat model for publicly exposing the generative model

In order to make those attacks non-effective, the watermark embedding function $W(D)$ is being applied on the dataset before feeding it into a training pipeline. Then the watermark extraction and verification function $Det_W(\widehat{X})$ is utilized to identify the output of the generative model making the attack with it non-effective since its origin is revealed (pic. 7):

$$\widehat{X_w} = G_0(Z),\ Det_W(\widehat{X_w}) \approx 1,\ s(X,\widehat{X}) \geq s_0 \tag{7}$$



Pic. 7. The defense model for publicly exposing the generative model with watermarked training data

LIMITATIONS AND CHALLENGES

Despite its promise as a primary defense against the malicious use of deepfakes, the practical implementation of proactive watermarking is fraught with significant challenges that span technical limitations, security vulnerabilities, and unresolved governance issues. This section synthesizes the key problems and obstacles identified across the literature, categorized for clarity.

Technical and foundational limitations. At the most fundamental level, the very building blocks of modern watermarking schemes exhibit inherent weaknesses that undermine their security guarantees.

A core obstacle is the lack of adversarially robust embedding models. The machine learning models required to build a robust and publicly-detectable watermark are not fundamentally secure. As demonstrated in [19], state-of-the-art image embedding models are vulnerable to adversarial attacks that can force collisions, making two different images appear the same to the model. This vulnerability breaks the unforgeability of the entire watermarking scheme, a critical security property.

Furthermore, the literature reveals a fundamental trade-off between evasion and spoofing errors. For watermarking methods that use subtle, imperceptible perturbations, any attempt to make the watermark harder to remove inherently makes it easier for an attacker to make an authentic image appear watermarked, and vice versa. This inherent conflict suggests that a perfectly secure imperceptible watermark may be theoretically impossible [18].

The attack surface is broadened by the discovery that all robust watermarking schemes, by necessity, must embed their signals in the image's spectral amplitudes. This creates a universal, predictable "carrier" that attackers can target directly without needing to know anything about the specific watermarking algorithm, a vulnerability exploited by the UnMarker attack [6].

Finally, for any classifier-based deepfake detector, a fundamental conflict exists between robustness and reliability. A detector cannot simultaneously achieve high performance and high robustness to input perturbations. This trade-off becomes more pronounced as generative models improve and the statistical differences between real and fake content diminish, limiting the long-term viability of classifier-based approaches [18].

Security Vulnerabilities and Adversarial Attacks. The theoretical limits of watermarking are compounded by a range of practical attacks that have proven highly effective at defeating current state-of-the-art systems.

Watermarks with a low perturbation budget can be effectively removed via diffusion purification attacks, which compromise the watermark with minimal changes to the image [18]. Conversely, watermarks with a high perturbation budget (i.e., more visible or structural ones) are vulnerable to model substitution attacks, where an attacker uses a surrogate detector to craft an adversarial perturbation that successfully transfers to and fools the authentic, black-box detector.

Most concerning is the emergence of universal no-box transfer attacks. Even without any access to the detector model or its API, attackers can successfully remove watermarks by ensembling multiple surrogate models to generate a single, highly transferable perturbation [5]. Building on this, universal spectral attacks like UnMarker can directly disrupt the spectral amplitudes of an image to erase the watermark without needing any feedback or knowledge of the scheme, defeating even the most advanced semantic watermarks like TreeRing [6].

In addition to evasion, systems are vulnerable to spoofing attacks, where an attacker can cause authentic content to be misclassified as AI-generated. This poses a significant

reputational risk to developers if their models are falsely associated with malicious or obscene content.

Governance, policy, and implementation challenges. The technical fragility of watermarking is exacerbated by a chaotic and immature governance landscape.

A primary issue is the lack of standards and independent verification. Currently, there are no common technical standards, evaluation benchmarks, or independent testing protocols. This makes regulatory compliance unverifiable and allows companies to deploy weak schemes while claiming to meet governance goals [60], [2].

This leads to a significant control and trust deficit, often described as the "judge and jury" problem, where watermark detection is almost exclusively controlled by the same entity that provides the generative model. The ecosystem is further characterized by fragmentation and non-interoperability, with different companies developing proprietary, incompatible methods that render detection useless at scale.

Empirical studies confirm these concerns, showing low and inadequate adoption in practice. A recent analysis of 50 popular AI image generators found that only a minority (38%) implement any form of machine-readable marking, with most relying on easily stripped metadata. Visible deepfake labels are even rarer (18%) [1]. These implementation gaps are widened by ambiguities in legal frameworks like the EU AI Act, which contain unclear definitions and responsibilities, hindering consistent enforcement.

Practical and economic obstacles. Finally, a set of practical and economic factors creates powerful disincentives against the widespread adoption of robust watermarking.

AI providers face conflicting industry incentives, balancing societal safety against user demand for "clean," unwatermarked content. This encourages "symbolic compliance"– announcing watermarking initiatives for public relations value without implementing truly robust solutions [60], [1].

The prevalence of open-source models poses another major obstacle, as watermarking can often be easily disabled by end-users. Lastly, the high computational and economic costs associated with robust cryptography, secure audit infrastructure, and defending against increasingly sophisticated attacks present a significant barrier to entry for all but the largest technology companies.

CONCLUSION

This survey has provided a comprehensive examination of proactive watermarking techniques as a critical defense against the escalating threat of malicious deepfakes. As passive forensic methods struggle to keep pace with the rapid evolution of generative models, proactive approaches that embed verifiable information at the point of content creation have emerged as an indispensable line of defense for establishing digital provenance and trust.

Our analysis has systematically categorized the landscape of deepfake watermarking, distinguishing between methods designed for manipulation detection, provenance tracing, content integrity verification, and the watermarking of generative models themselves. This taxonomy highlights the diverse threat models and technical trade-offs inherent in each approach. While the field has demonstrated significant innovation – with methods achieving high levels of imperceptibility and robustness in controlled settings – our review underscores that the practical, scaled deployment of these technologies is fraught with formidable challenges.

The most pressing issues are not merely algorithmic, but span fundamental security vulnerabilities, a fractured governance ecosystem, and misaligned economic incentives.

Technically, a fundamental trade-off persists between watermark robustness and imperceptibility, while the rise of powerful, universal adversarial attacks like diffusion purification and spectral disruption threatens to undermine even the most advanced schemes. This technical fragility is compounded by a lack of industry-wide standards, independent verification protocols, and interoperability, creating a "judge and jury" problem where accountability is limited. Furthermore, the prevalence of open-source models and weak industry adoption ("symbolic compliance") create significant gaps in the defense ecosystem.

In conclusion, while proactive watermarking remains one of the most promising and necessary tools in the fight against synthetic media manipulation, it is not a panacea. Future progress depends on a multi-faceted effort. Research must advance beyond novel embedding algorithms to focus on creating fundamentally secure, adversarially robust models. Simultaneously, progress requires a concerted push from policymakers, industry consortia, and the research community to establish clear standards, benchmarks, and a governance framework that can foster a truly trustworthy digital information ecosystem. Without such a holistic approach, even the most sophisticated watermarking solutions will fall short of providing the reliable, large-scale protection that society urgently needs.

REFERENCES

1. Rijsbosch, B., van Dijck, G., & Kollnig, K. (2025). Adoption of watermarking measures for AI-generated content and implications under the EU AI Act. *arXiv*. https://arxiv.org/abs/2503.18156
2. Fernandez, P., Level, A., & Furon, T. (2024). What lies ahead for generative AI watermarking. In *ICML 2024 Workshop on Generative AI and Law*.
3. Nemecek, A., Jiang, Y., & Ayday, E. (2025). Watermarking without standards is not AI governance. *arXiv*. https://arxiv.org/abs/2505.23814
4. Tilo, D. (2025, April 9). Singapore firm nearly lost $500,000 after deepfake video scam: Police. *HRD Asia*. https://www.hcamag.com/asia/specialisation/hr-technology/singapore-firm-nearly-lost-500000-after-deepfake-video-scam-police/531450
5. Hu, Y., Jiang, Z., Guo, M., & Gong, N. Z. (2024). A transfer attack to image watermarks. *arXiv*. https://arxiv.org/abs/2403.15365
6. Kassis, A., & Hengartner, U. (2025). UnMarker: A universal attack on defensive image watermarking. In *Proceedings of the IEEE Symposium on Security and Privacy* (pp. 2602–2620). IEEE.
7. Yang, Y., Liang, C., He, H., Cao, X., & Gong, N. Z. (2021). FaceGuard: Proactive deepfake detection. *arXiv*. https://arxiv.org/abs/2109.05673
8. Zhao, Y., Liu, B., Ding, M., Liu, B., Zhu, T., & Yu, X. (2023). Proactive deepfake defence via identity watermarking. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 4602–4611).
9. Wang, T., Huang, M., Cheng, H., Zhang, X., & Shen, Z. (2024). LampMark: Proactive deepfake detection via training-free landmark perceptual watermarks. In *Proceedings of the ACM International Conference on Multimedia* (pp. 10515–10524).
10. Lan, S., Liu, K., Zhao, Y., Yang, C., Wang, Y., Yao, X., & Zhu, L. (2024). Facial features matter: A dynamic watermark-based proactive deepfake detection approach. *arXiv*. https://arxiv.org/abs/2411.14798
11. Wang, R., Juefei-Xu, F., Guo, Q., Huang, Y., Ma, L., Liu, Y., & Wang, L. (2020). DeepTag: Robust image tagging for deepfake provenance. *arXiv*. https://arxiv.org/abs/2009.09869
12. Wang, R., Juefei-Xu, F., Luo, M., Liu, Y., & Wang, L. (2021). FakeTagger: Robust safeguards against deepfake dissemination via provenance tracking. In *Proceedings of the ACM International Conference on Multimedia* (pp. 3546–3555).
13. Yu, N., Skripniuk, V., Abdelnabi, S., & Fritz, M. (2021). 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).
14. Sun, P., Qi, H., Li, Y., & Lyu, S. (2023). FakeTracer: Catching face-swap deepfakes via implanting traces in training. *arXiv*. https://arxiv.org/abs/2307.14593
15. Sanjalawe, Y., Al-E'mari, S., Fraihat, S., Abualhaj, M., & Alzubi, E. (2025). A deep learning-driven multi-layered steganographic approach for enhanced data security. *Scientific Reports, 15*(1), 4761.

16. Wang, T., Huang, M., Cheng, H., Ma, B., & Wang, Y. (2023). Robust identity perceptual watermark against deepfake face swapping. *arXiv*. https://arxiv.org/abs/2311.01357
17. Wu, X., Liao, X., & Ou, B. (2023). SepMark: Deep separable watermarking for unified source tracing and deepfake detection. In *Proceedings of the ACM International Conference on Multimedia* (pp. 1190–1201).
18. Saberi, M., Sadasivan, V. S., Rezaei, K., Kumar, A., Chegini, A., Wang, W., & Feizi, S. (2023). Robustness of AI-image detectors: Fundamental limits and practical attacks. *arXiv*. https://arxiv.org/abs/2310.00076
19. Fairoze, J., Ortiz-Jimenez, G., Vecerik, M., Jha, S., & Gowal, S. (2025). On the difficulty of constructing a robust and publicly-detectable watermark. *arXiv*. https://arxiv.org/abs/2502.04901
20. Jayasumana, S., Ramalingam, S., Veit, A., Glasner, D., Chakrabarti, A., & Kumar, S. (2024). Rethinking FID: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9307–9315).
21. Neekhara, P., Hussain, S., Zhang, X., Huang, K., McAuley, J., & Koushanfar, F. (2024). FaceSigns: Semi-fragile watermarks for media authentication. *ACM Transactions on Multimedia Computing, Communications and Applications*, *20*(11), 1–21.
22. Lai, Z., Yao, Z., Lai, G., Wang, C., & Feng, R. (2024). A novel face swapping detection scheme using the pseudo Zernike transform-based robust watermarking. *Electronics*, *13*(24), 4955.
23. Tang, L., Ye, Q., Hu, H., Xue, Q., Xiao, Y., & Li, J. (2024). DeepMark: A scalable and robust framework for deepfake video detection. *ACM Transactions on Privacy and Security*, *27*(1), 1–26.
24. Noreen, I., Muneer, M. S., & Gillani, S. (2022). Deepfake attack prevention using steganography GANs. *PeerJ Computer Science*, *8*, e1125.
25. Beuve, N., Hamidouche, W., & Déforges, O. (2023). WaterLo: Protect images from deepfakes using localized semi-fragile watermark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 393–402).
26. Nadimpalli, A. V., & Rattani, A. (2024). Social media authentication and combating deepfakes using semi-fragile invisible image watermarking. *Digital Threats: Research and Practice, 5*(4), 1–30.
27. Sun, C., Sun, H., Guo, Z., Diao, Y., Wang, L., Ma, D., et al. (2025). DiffMark: Diffusion-based robust watermark against deepfakes. *arXiv*. https://arxiv.org/abs/2507.01428
28. Walczyna, T., Zurada, J. M., & Piotrowski, Z. (2025). RE-Mark: An identity-recovery watermarking method for undoing deepfake face-swap. *Authorea Preprints*.
29. Wang, T., Cheng, H., Liu, M. H., & Kankanhalli, M. (2025). FractalForensics: Proactive deepfake detection and localization via fractal watermarks. In *Proceedings of the ACM International Conference on Multimedia* (pp. 7210–7219).
30. Shoaib, M. R., Wang, Z., Ahvanooey, M. T., & Zhao, J. (2023). Deepfakes, misinformation, and disinformation in the era of frontier AI. In *Proceedings of the International Conference on Computer and Applications (ICCA)* (pp. 1–7). IEEE.

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 4 (32), 2026
ISSN 2663 – 4023

Марчук Михайло Борисович
аспірант кафедри захисту інформації
Вінницький національний технічний університет, Вінниця, Україна
ORCID: 0009-0003-4773-6541
smoke222catches@gmail.com

Лукічов Віталій Володимирович
доцент кафедри захисту інформаці]
Вінницький національний технічний університет, Вінниця, Україна
ORCID: 0000-0002-3423-5436
lukichov.vitalyi@vntu.edu.ua

ОГЛЯД МЕТОДІВ ВОДЯНИХ ЗНАКІВ ДЛЯ ПРОАКТИВНОГО ЗАХИСТУ ВІД DEEPFAKE

Анотація. З розвитком генеративних моделей контент, створений за допомогою deepfake-технологій, стає все складніше відрізнити від реальності і разом з цим пасивні методи визначення згенерованого контенту стають все більш неефективними. Зловмисне використання засобів для генерації контенту надає зловмисникам можливості для соціальної інженерії, дезінформаційних кампанії та шахрайства. В зв'язку з цим виникає необхідність в новому класі інструментів, заснованих на попередньому маркуванні автентичного контенту, щоб захистити його від використання у deepfake-контенті або дезінформаційних кампаній. У цьому дослідженні ми проводимо комплексний аналіз рішень для водяних знаків (вотермаркінг) з метою проактивного захисту від deepfake. Ми ідентифікували основні рішення для вотермаркінгу, що стійкі до deepfake, які представлені в науковій літературі та створили таксономію для них. Також ми визначили основні метрики та датасети для тренування моделей вотермаркінгу на базі глибинного машинного навчання для проактивного захисту від deepfake. Ми провели кількісний та якісний аналіз існуючих рішень, їх методів, метрик та застосувань. В кінці ми проаналізували існуючі проблеми та виклика в даній сфері. Даний огляд може слугувати основою для майбутніх досліджень та впровадження політики для генеративного ШІ.

Ключові слова: deepfake; водяні знаки; дослідження зображень; глибинне машинне навчання; стеганографія; інформаційна безпека.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Rijsbosch, B., van Dijck, G., & Kollnig, K. (2025). Adoption of watermarking measures for AI-generated content and implications under the EU AI Act. *arXiv*. https://arxiv.org/abs/2503.18156
2. Fernandez, P., Level, A., & Furon, T. (2024). What lies ahead for generative AI watermarking. In *ICML 2024 Workshop on Generative AI and Law*.
3. Nemecek, A., Jiang, Y., & Ayday, E. (2025). Watermarking without standards is not AI governance. *arXiv*. https://arxiv.org/abs/2505.23814
4. Tilo, D. (2025, April 9). Singapore firm nearly lost $500,000 after deepfake video scam: Police. *HRD Asia*. https://www.hcamag.com/asia/specialisation/hr-technology/singapore-firm-nearly-lost-500000-after-deepfake-video-scam-police/531450
5. Hu, Y., Jiang, Z., Guo, M., & Gong, N. Z. (2024). A transfer attack to image watermarks. *arXiv*. https://arxiv.org/abs/2403.15365
6. Kassis, A., & Hengartner, U. (2025). UnMarker: A universal attack on defensive image watermarking. In *Proceedings of the IEEE Symposium on Security and Privacy* (pp. 2602–2620). IEEE.
7. Yang, Y., Liang, C., He, H., Cao, X., & Gong, N. Z. (2021). FaceGuard: Proactive deepfake detection. *arXiv*. https://arxiv.org/abs/2109.05673
8. Zhao, Y., Liu, B., Ding, M., Liu, B., Zhu, T., & Yu, X. (2023). Proactive deepfake defence via identity watermarking. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 4602–4611).
9. Wang, T., Huang, M., Cheng, H., Zhang, X., & Shen, Z. (2024). LampMark: Proactive deepfake detection via training-free landmark perceptual watermarks. In *Proceedings of the ACM International Conference on Multimedia* (pp. 10515–10524).

10. Lan, S., Liu, K., Zhao, Y., Yang, C., Wang, Y., Yao, X., & Zhu, L. (2024). Facial features matter: A dynamic watermark-based proactive deepfake detection approach. *arXiv*. https://arxiv.org/abs/2411.14798
11. Wang, R., Juefei-Xu, F., Guo, Q., Huang, Y., Ma, L., Liu, Y., & Wang, L. (2020). DeepTag: Robust image tagging for deepfake provenance. *arXiv*. https://arxiv.org/abs/2009.09869
12. Wang, R., Juefei-Xu, F., Luo, M., Liu, Y., & Wang, L. (2021). FakeTagger: Robust safeguards against deepfake dissemination via provenance tracking. In *Proceedings of the ACM International Conference on Multimedia* (pp. 3546–3555).
13. Yu, N., Skripniuk, V., Abdelnabi, S., & Fritz, M. (2021). 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).
14. Sun, P., Qi, H., Li, Y., & Lyu, S. (2023). FakeTracer: Catching face-swap deepfakes via implanting traces in training. *arXiv*. https://arxiv.org/abs/2307.14593
15. Sanjalawe, Y., Al-E'mari, S., Fraihat, S., Abualhaj, M., & Alzubi, E. (2025). A deep learning-driven multi-layered steganographic approach for enhanced data security. *Scientific Reports, 15*(1), 4761.
16. Wang, T., Huang, M., Cheng, H., Ma, B., & Wang, Y. (2023). Robust identity perceptual watermark against deepfake face swapping. *arXiv*. https://arxiv.org/abs/2311.01357
17. Wu, X., Liao, X., & Ou, B. (2023). SepMark: Deep separable watermarking for unified source tracing and deepfake detection. In *Proceedings of the ACM International Conference on Multimedia* (pp. 1190–1201).
18. Saberi, M., Sadasivan, V. S., Rezaei, K., Kumar, A., Chegini, A., Wang, W., & Feizi, S. (2023). Robustness of AI-image detectors: Fundamental limits and practical attacks. *arXiv*. https://arxiv.org/abs/2310.00076
19. Fairoze, J., Ortiz-Jimenez, G., Vecerik, M., Jha, S., & Gowal, S. (2025). On the difficulty of constructing a robust and publicly-detectable watermark. *arXiv*. https://arxiv.org/abs/2502.04901
20. Jayasumana, S., Ramalingam, S., Veit, A., Glasner, D., Chakrabarti, A., & Kumar, S. (2024). Rethinking FID: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9307–9315).
21. Neekhara, P., Hussain, S., Zhang, X., Huang, K., McAuley, J., & Koushanfar, F. (2024). FaceSigns: Semi-fragile watermarks for media authentication. *ACM Transactions on Multimedia Computing, Communications and Applications, 20*(11), 1–21.
22. Lai, Z., Yao, Z., Lai, G., Wang, C., & Feng, R. (2024). A novel face swapping detection scheme using the pseudo Zernike transform-based robust watermarking. *Electronics, 13*(24), 4955.
23. Tang, L., Ye, Q., Hu, H., Xue, Q., Xiao, Y., & Li, J. (2024). DeepMark: A scalable and robust framework for deepfake video detection. *ACM Transactions on Privacy and Security, 27*(1), 1–26.
24. Noreen, I., Muneer, M. S., & Gillani, S. (2022). Deepfake attack prevention using steganography GANs. *PeerJ Computer Science, 8*, e1125.
25. Beuve, N., Hamidouche, W., & Déforges, O. (2023). WaterLo: Protect images from deepfakes using localized semi-fragile watermark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 393–402).
26. Nadimpalli, A. V., & Rattani, A. (2024). Social media authentication and combating deepfakes using semi-fragile invisible image watermarking. *Digital Threats: Research and Practice, 5*(4), 1–30.
27. Sun, C., Sun, H., Guo, Z., Diao, Y., Wang, L., Ma, D., et al. (2025). DiffMark: Diffusion-based robust watermark against deepfakes. *arXiv*. https://arxiv.org/abs/2507.01428
28. Walczyna, T., Zurada, J. M., & Piotrowski, Z. (2025). RE-Mark: An identity-recovery watermarking method for undoing deepfake face-swap. *Authorea Preprints*.
29. Wang, T., Cheng, H., Liu, M. H., & Kankanhalli, M. (2025). FractalForensics: Proactive deepfake detection and localization via fractal watermarks. In *Proceedings of the ACM International Conference on Multimedia* (pp. 7210–7219).
30. Shoaib, M. R., Wang, Z., Ahvanooey, M. T., & Zhao, J. (2023). Deepfakes, misinformation, and disinformation in the era of frontier AI. In *Proceedings of the International Conference on Computer and Applications (ICCA)* (pp. 1–7). IEEE.

Отримано редакцією журналу / Received: 27.01.26
Прорецензовано / Revised: 18.02.26
Схвалено до друку / Accepted: 26.03.26