

Analysing the Authentication Mechanism for GAN-Generated Adversarial Network Content through a Unified Industrial Framework: Issues, Challenges, and Future Recommendations

Harshita Maurya, Sakshi Prajapati, Ruchi Kumari

Shri Ramswaroop Memorial University, Lucknow, India

Abstract

Generative adversarial networks (GANs) have evolved from an esoteric research area into mainstream production tools for generating images, video, audio, and text that are often indistinguishable from authentic content. As this realism grows, organizations in banking, healthcare, media, insurance, law, e-commerce, and governments have an urgent industrial problem: they need reliable methods to authenticate a given piece of digital content as being genuine or not. In this article, we examine the technical foundations of GANs, present the problem in the context of the literature on forensic-detection, provenance, and policy, review the industry-specific threats raised by synthetic media, and survey the main families of detection and authentication techniques, namely, pixel-level forensics, frequency-domain analysis, metadata inspection, deep-learning-based classifiers, watermarking, digital signatures, and provenance systems. We argue that detection is structurally insufficient due to limitations in generalizability: while detectors trained on one generative model generalize poorly to new ones, thus engendering a continuous arms race between the generator and detector, and propose a five-layer Unified Industrial Authentication Framework that combines cryptographic source verification, provenance, multi-model AI detection, integrity verification, and a formalized authentication confidence score as a viable industry solution to address this gap rather than a theoretical one. We validate our approach against known threats to validity, discuss its advantages and limitations, and conclude with industry applications, open challenges, and directions for future work.

Keywords: Generative Adversarial Networks, deepfake detection, digital provenance, cryptographic signatures, C2PA.

1. Introduction

1.1 Background and Evolution of GANs

Generative adversarial networks (GANs) were developed by Goodfellow et al. in 2014 as a system where two neural networks engage in a minimax game: the generator creates synthetic samples from random noise, while the discriminator attempts to distinguish between real and generated samples [1]. As training progresses, the generator learns to output samples with a statistical distribution similar to the training data. While initial GANs produced low-resolution images, within five years the field has seen breakthroughs enabling photo-realism: progressive growing of GANs made high-resolution face synthesis possible [2], BigGAN showed the ability to perform class-conditional generation on ImageNet scale [3], and StyleGAN introduced a style-based generator architecture with control over facial attributes, and subsequently improved

in StyleGAN2 [4, 5]. Variants such as CycleGAN have also been developed for image-to-image translation with unpaired images [6].

The space of generative models has since expanded to incorporate other approaches. Denoising diffusion probabilistic models [7] were shown to outperform GANs on image-synthesis benchmarks [8], and latent diffusion has allowed high-resolution text-to-image generation to become widely available [9]. However, GANs have not been eclipsed: the GigaGAN large-scale text-to-image GAN shows that GANs can compete in that space while benefiting from much faster sampling than diffusion [10]. For the problem of authentication, the practical implication is that content can originate from multiple generator families, which leave different statistical traces. Initial work showed that a classifier trained on a single CNN generator generalizes to other CNN-based generators, while the authors caveat that this may only be the case “for now” [11]; subsequent work has looked at how detectors trained on GAN images generalize to diffusion images, especially after compression and resizing [12].

1.2 Investment, Market Size and Growth

Reliable public figures for money spent on GANs specifically are hard to find: no government or academic body tracks the expenditure on GANs separately from other generative AI. The closest alternative is commercial market-research revenue estimates. Grand View Research estimates the global GAN market to be worth USD 5.52 billion in 2024, USD 7.27 billion in 2025, and USD 36.01 billion in 2030, at a compound annual growth rate (CAGR) of 37.7% from 2025 to 2030 [13]. An independent estimate from Acumen Research and Consulting finds a similar value for 2025 (USD 7.06 billion) and predicts a 38.0% CAGR to 2035, with cloud deployment comprising 63% of 2025 revenue and North America representing about 37% [14]. Figure 1 plots the Grand View series. Due to being vendor forecasts rather than audited figures, these should be read as indicative of trend and order of magnitude.

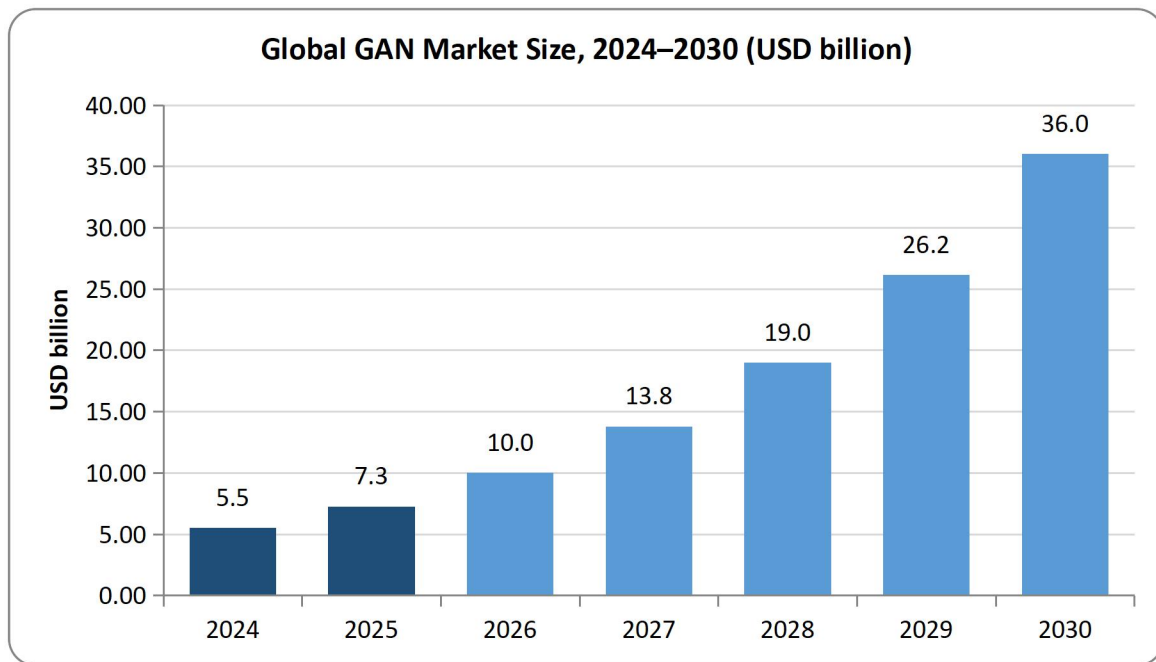


Figure 1. Global GAN market size, 2024–2030 (USD billion). 2024–2025 values and the 2030 value are from Grand View Research [13]; 2026–2029 values are derived by applying the reported 37.7% CAGR to the 2025 base and are therefore derived estimates, not separately published figures. To edit: right-click the chart in Word and choose Edit Data.

GANs sit inside a much larger wave of generative AI investment. The Stanford AI Index 2026 reports that global corporate AI investment reached USD 581.7 billion in 2025, up roughly 130% from USD 253 billion in 2024; private investment alone was USD 344.7 billion (up 127.5%), and generative AI captured nearly half of all private AI funding while growing by more than 200% [15, 16]. The same report finds that generative AI reached 53% population-level adoption within three years of mass-market introduction, faster than the personal computer or the internet, and that 70% of surveyed organizations now use it in at least one business function [15]. Table 1 summarizes the key figures.

Table 1. Key market and threat indicators.

Indicator	Value	Source
Global GAN market, 2024	USD 5.52 billion	Grand View Research [13]
Global GAN market, 2025	USD 7.27 billion	Grand View Research [13]
Projected GAN market, 2030	USD 36.01 billion (CAGR 37.7%)	Grand View Research [13]
Projected GAN market, 2035	USD 177.31 billion (CAGR 38.0%)	Acumen Research [14]
Global corporate AI investment, 2025	USD 581.7 billion (about +130% YoY)	Stanford AI Index 2026 [15, 16]

Indicator	Value	Source
Organizations using generative AI	70% (at least one function)	Stanford AI Index 2026 [15]
US GenAI-enabled fraud losses	USD 12.3 billion (2023) to USD 40 billion (2027 projection)	Deloitte [17]

1.3 Why Growth Creates a Security Problem

Falling generation costs and rising quality shift the economics toward attackers. Deloitte's Center for Financial Services projects that generative-AI-enabled fraud losses in the United States could rise from USD 12.3 billion in 2023 to USD 40 billion by 2027, a 32% compound annual growth rate, in its forecast built on scenarios of generative AI adoption [17]. A widely reported illustration is the Arup case in Hong Kong, in which an employee transferred about USD 25 million after a video call with deepfake versions of company executives [17]. Financial institutions rely on facial verification at onboarding, newsrooms publish photographs as evidence, insurers accept photographic damage claims, courts admit digital media, and hospitals depend on diagnostic imaging. In each setting a synthetic or manipulated artifact that passes as genuine can cause direct financial, legal, medical or reputational harm.

1.4 Future Scope

Three directions are likely to shape the field over the coming years. First, generation is moving toward multimodal, real-time synthesis (video, voice and live video calls), where verification windows are short. Second, GANs are increasingly used in beneficial roles, notably synthetic data generation for training models where real data is scarce or privacy-restricted, which is a major driver of the market growth cited above [13, 14]. Third, GANs are becoming one component in hybrid pipelines alongside diffusion and transformer models, for example as fast decoders or adversarial fine-tuning stages [10]. Market forecasts expect Asia Pacific to be the fastest-growing region at a 39.6% CAGR to 2035 [14]. Each of these trends increases the volume and variety of synthetic content that organizations must be able to authenticate.

1.5 Problem Statement and Contribution

While GANs pioneered high-fidelity synthetic media and remain a useful lens for explaining the authentication problem, the enterprise threat landscape now includes diffusion models such as Stable Diffusion and Midjourney, as well as generative video and large language models. The threats, detection challenges and layered framework proposed in this article apply equally to content from these newer model families, although the discussion continues to use GANs as its primary illustrative example.

The core industry problem can now be stated explicitly. Existing authentication mechanisms are increasingly unable to establish the origin and integrity of digital content, because synthetic media can be produced without visible artifacts that a human reviewer or legacy forensic check would catch. The AI detectors built to close this gap are themselves fragile: their reliability on unseen generator families is a documented concern [11, 12], and they degrade under routine compression and re-encoding, deliberate manipulation and adversarial adaptation, as detailed

in Section 7. Industries therefore need an authentication mechanism that does not rely on detection in isolation, and instead combines content forensics with cryptographic provenance and integrity verification. This motivates the five-layer Unified Industrial Authentication Framework presented in Section 8, which is the central original contribution of this article.

1.6 Research Objectives

This article pursues four objectives:

- Characterize how GAN- and diffusion-generated content undermines the authenticity of digital evidence in banking, healthcare, media, insurance, law, e-commerce and government.
- Evaluate existing detection and authentication techniques (pixel and frequency forensics, learned classifiers, watermarking, digital signatures and provenance) and identify where each one fails.
- Propose a five-layer Unified Industrial Authentication Framework that combines cryptographic source verification, provenance tracking, multi-model AI detection, integrity verification and an authentication confidence score.
- Assess the framework against threats to validity, and set out deployment challenges (real-time performance, explainability, privacy, cross-platform durability) and future work.

2. LITERATURE REVIEW

2.1 Forensic and Frequency-Domain Detection

First-generation GAN detectors were built on traces that generators leave in the image itself. Rössler et al. built the FaceForensics++ benchmark to explore the detection of manipulated facial images [18], and Yu et al. showed that GANs leave individual fingerprints that allow generated images to be attributed to a specific model [19]. In the frequency domain, Durall et al. reported that CNN-based generators fail to reproduce the spectral distribution of real images [20], and Frank et al. traced high-frequency artifacts to the upsampling operations used in GAN architectures, showing that these artifacts were consistent across architectures, datasets and resolutions and could be used for automated detection [21]. These methods are interpretable and cheap to run, but they depend on artifacts tied to a particular generation mechanism.

2.2 Learned Detectors and the Generalization Problem

Deep classifiers trained to separate real from generated images are the dominant approach. Wang et al. found that a classifier trained on one CNN generator (ProGAN) generalized well to other CNN-based generators, while cautioning that this might hold only “for now” [11]. That caution proved justified. Ojha et al. showed that the standard real-versus-fake training paradigm, when trained on GAN images, fails to detect images from newer generative model families, and explained the failure by the classifier being tuned to fake-specific patterns while treating anything else as real [22]. Corvi et al. studied the forensic traces of diffusion models and how detectors developed for GAN images perform on them, including under the compression and resizing typical of social networks [12]. Ojha et al. proposed classification in the feature space of a large pretrained vision-language model, which improved on the prior state of the art by 15.07

mAP and 25.90% accuracy on unseen diffusion and autoregressive models [22]. Detection is therefore improving, but every advance is measured against generators that already exist.

2.3 Watermarking and Provenance

A second line of work avoids guessing from pixels and instead attaches evidence at creation. The Coalition for Content Provenance and Authenticity (C2PA) specifies Content Credentials, cryptographically signed records of a file's origin and edit history; its deployment guidance is at specification version 2.4 and treats the standard as distinct from watermarking and fingerprinting, with the combination described as “Durable Content Credentials” [23]. Invisible watermarking is the complementary mechanism, but its robustness is contested: Zhao et al. showed that regeneration attacks, which add noise and then reconstruct the image with a generative model, remove pixel-level invisible watermarks across the four schemes they tested, and argued for semantic-preserving watermarks instead [24]. Because C2PA pairs signed manifests with soft bindings (watermarks and fingerprints) so that provenance can be recovered when a manifest becomes separated from its asset, provenance should not be assumed to remain embedded and available across every distribution path, for example when metadata is stripped. C2PA also verifies the origin and integrity of a signed asset; it does not establish that the depicted event is true.

2.4 Gap Identification

Three gaps follow. (1) Generalization gap: detectors trained on one generator family transfer unreliably to the next [12, 22]. (2) Robustness gap: both learned detectors and invisible watermarks degrade under compression, resizing or regeneration [12, 24]. (3) Integration gap: within the sources reviewed here, forensics, watermarking and provenance are studied largely as separate mechanisms, and none is presented as combined into a single, industry-facing decision output such as a confidence score with an audit trail. Academic benchmarks such as FaceForensics++ [18] measure detection accuracy, whereas an insurer, bank or court needs a defensible accept, review or reject decision.

2.5 Current Research Work

The present work responds directly to these gaps. Rather than proposing another single detector, it treats authentication as a layered decision problem and contributes the five-layer Unified Industrial Authentication Framework presented in Section 8. Cryptographic source verification and provenance tracking establish where content came from when credentials exist (addressing the provenance dependence noted above). Multi-model AI detection provides a second opinion when they do not, and multi-model detection is proposed to reduce dependence on any single detector that may not generalize (the generalization gap). Integrity verification checks whether signed content has changed since signing or capture, so that alteration can be caught without relying on detectors or watermarks that degrade under compression (addressing part of the robustness gap, though it does not make the detectors themselves more robust). Provenance and integrity checks establish origin and unaltered state, not that the depicted event is factually true. A formalized authentication confidence score combines the layers into one output that a reviewer can act on (the integration gap). The contribution of this article is at

the framework and analysis level; it is assessed against known threats to validity after the framework is presented, rather than through a new benchmark of detector accuracy.

3. UNDERSTANDING GANS AND SYNTHETIC CONTENT

A GAN is composed of two neural networks trained in opposition. The generator receives a random noise vector as input and transforms it into a synthetic sample, such as an image. The discriminator receives both real samples drawn from a training dataset and synthetic samples produced by the generator, and are trained to classify each as real or fake. During training, the generator is updated to produce samples that are more likely to be classified as real, while the discriminator is updated to become better at telling real and generated samples apart. Over many iterations, this adversarial process drives the generator toward producing outputs whose statistical distribution increasingly resembles that of the real training data.

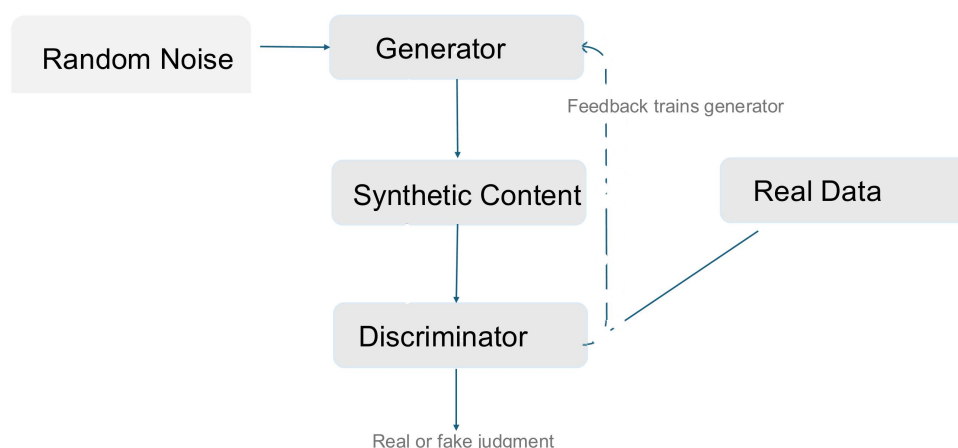


Figure 2. Basic GAN architecture: the generator and discriminator train against each other in a feedback loop.

Since their initial formulation, there has been a proliferation of GAN variants, including conditional GANs, enabling training with images and labels, CycleGAN, which enables image-to-image translation with unpaired training examples, and StyleGAN, which introduced mode mixing capabilities and enabled manipulation of realistic human faces. Diffusion based generative models have recently emerged as an alternative to GANs at the forefront of synthetic media generation, and many of the problems related to authentication described in this article apply to both families of models, although we will continue to use the term GAN in the interests of consistency with the overall article.

The reason authentication with such models is challenging is that the generator is explicitly trained to mislead a detector, and that the detector often uses similar architectures that would be used in a forensic detector. Progress in one necessarily hurts the other.

4. INDUSTRIAL PROBLEMS

General Issues

1. Erosion of Trust in Digital Content

- GAN generated images and videos make it increasingly difficult to discern authentic content from synthetic content.
- It may become increasingly difficult to trust what you see as being real, and the digital world will become polluted with fake news made easy by synthetic media.

2. Spread of Misinformation

- Images and videos generated by GANs may be disseminated rapidly on social media.
- Misinformation may be spread rapidly and cause confusion during critical events.

3. Identity Misrepresentation

- GANs can generate realistic looking faces that do not belong to any person
- Such synthetic faces can enable creation of synthetic social media personas and misrepresentation

4. Privacy and Reputation Risks

- Manipulated video and images can superimpose a person's face onto an image or video, effectively without consent.
- Such misrepresentation using synthetic media can damage personal reputations, and violate privacy

5. Authentication Challenges

- The notion that if an image does not look obviously fake it is authentic is no longer a valid assumption
- With current GAN technology, content that looks authentic to humans can be generated

6. Authentication Gap

- There is a widening gap between the rate of innovation in generative models and the ability to develop and deploy authentication algorithms at the same rate.
- Many current watermarking and other authentication technologies are not capable of detecting GAN generated content and need to be updated.

7. Difficulty in Manual Detection

- Content generated by GANs may show few if any visible artifacts, making manual detection challenging.
- Humans are not always capable of detecting manipulation in media generated by a GAN.



The practical implications of unreliable content authentication differ across industries, but the general idea remains the same: the decision-making process which assumes something is real can be compromised at the point of assertion. For banking and financial services, GAN-generated faces could be used to bypass Know Your Customer (KYC) checks and remote identity verification, resulting in account fraud and money laundering. The same principle applies to healthcare industry where manipulated or synthetically generated medical images could lead to incorrect diagnostic review (if not authenticated at the point of capture). Similarly, fake photos or videos could be disseminated as proof of some event that has not occurred in media/news industry, accelerating the spread of disinformation. In insurance industry, digitally-manipulated photos of damage or injury would be used to support fraudulent claims. In law industry, the evidential value or admissibility of digital image and video would hinge more and more on the ability to ensure authenticity. In e-commerce, synthetic product images and fake reviews could mislead online shoppers. In government and defense industry, deepfake videos of public officials may present risks of social engineering and disinformation security threats.

5. THREAT TO INDUSTRIAL PROBLEMS

Several concrete attack patterns can be associated with the idea of authentication gap and the ways it can be leveraged.

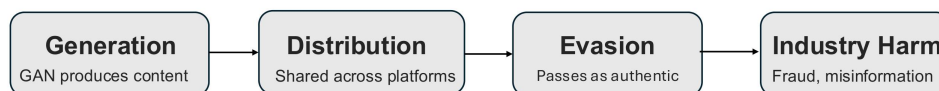


Figure 3. The GAN content threat lifecycle: each stage from generation to industry harm is a potential intervention point for the authentication framework.

5.1 Synthetic Identity and KYC Fraud

An attacker can use a generative model to produce a realistic but entirely synthetic face that does not correspond to any real individual, or a face that blends attributes of multiple real individuals. This synthetic identity can then be used to open financial accounts, apply for credit,

or bypass facial verification systems that rely on liveness or similarity checks not specifically designed to detect generative artifacts.

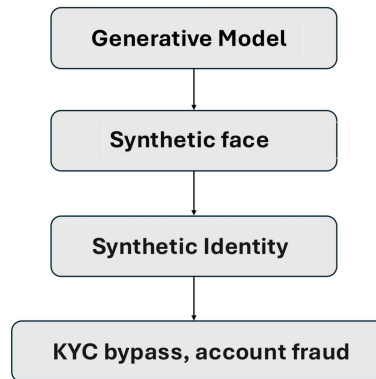


Figure 4. Synthetic identity and KYC fraud pathway.

5.2 Executive Deepfake Video

An attacker generates a video or audio clip that appears to show a company executive authorizing a financial transaction or issuing an instruction. Because such attacks exploit trust in a recognizable individual rather than a technical vulnerability in a payment system, they are particularly difficult to defend against using conventional cybersecurity controls.

5.3 Fabricated Evidence in Insurance Claims

A claimant submits a digitally altered photograph showing damage or injury inconsistent with the actual incident, seeking to inflate or fabricate a claim. Because insurers often rely heavily on photographic evidence, the absence of robust authentication can materially increase fraud losses.

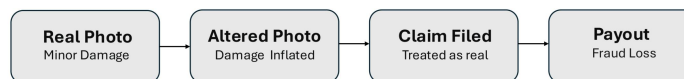


Figure 5. Fabricated evidence pathway in insurance claims.

5.4 Manipulation of Medical Imagery

Although less common than other scenarios, manipulated or synthetically altered medical images could in principle be introduced into a clinical workflow, whether through malicious intent, data corruption, or error, with potential consequences for diagnosis and treatment planning.



Figure 6. Manipulation of medical imagery pathway.

5.5 Fabricated Photojournalism

A synthetic photograph is presented as documentary evidence of an event that did not occur, and is circulated through social media and news channels before it can be authenticated, contributing to the spread of misinformation and eroding public trust in visual media generally.



Figure 7. Fabricated photojournalism pathway.

5.6 Real-World Precedent

These scenarios are not merely hypothetical. In January 2024, a finance employee at the Hong Kong office of the multinational engineering firm Arup was deceived into making fifteen wire transfers totaling approximately 25 million US dollars after joining a video conference in which the company's chief financial officer and several colleagues appeared to be present. Hong Kong police later confirmed that every participant on the call other than the victim was a real-time deepfake, generated from publicly available video and audio footage of the actual executives [25]. The incident, one of the largest documented cases of live deepfake-enabled fraud, illustrates precisely the executive deepfake scenario described in Section 5.2 and underscores that conventional executive verification procedures, namely recognizing a colleague's face and voice on a video call, can no longer be treated as reliable authentication on their own.

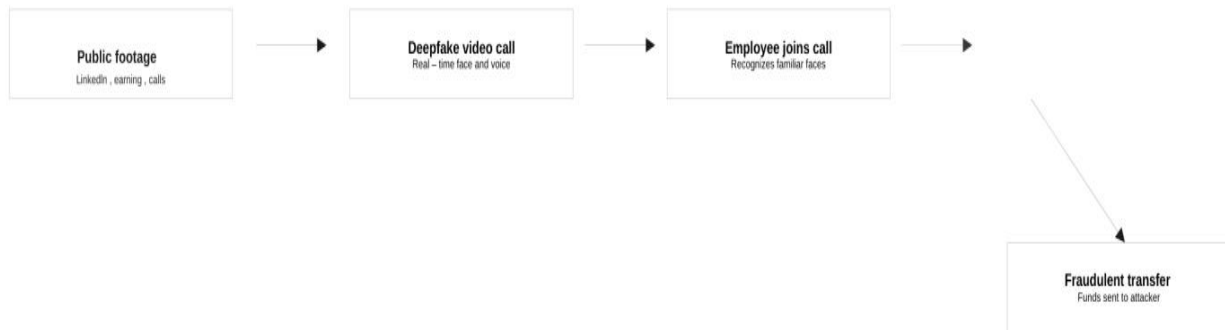


Figure 8. The Arup executive deepfake attack pathway (January 2024).

Beyond financial fraud, synthetic and manipulated imagery has also been documented circulating during periods of geopolitical conflict, where fabricated or AI-generated photos purporting to show battlefield damage, casualties, or troop movements have been shared on social media platforms faster than they could be verified or debunked, illustrating the fabricated photojournalism scenario described in Section 5.5 at a societal scale. Separately, security researchers and financial institutions have documented KYC bypass attempts in which fraudsters use virtual camera software to feed a synthetically generated or face-swapped video stream into a remote identity verification system in place of a live camera feed, directly targeting the liveness assumptions on which many biometric onboarding systems still rely [26, 27]. Together, these examples demonstrate that the threats described in this section are already being actively exploited across the financial, informational, and identity-verification domains discussed in Section 4, rather than only representing a risk of the future.

6. EXISTING GAN AUTHENTICATION AND DETECTION TECHNIQUES

Industry and academic research have produced a range of techniques for detecting GAN-generated or manipulated content. These techniques can be grouped into several broad families, which are frequently combined in practice rather than used in isolation.

6.1 Pixel-Level and Spatial Analysis

Early forensic approaches examine an image at the pixel level for irregularities in texture, edges, lighting consistency, and noise patterns that differ subtly from those found in images captured by real camera sensors. Because camera sensors introduce characteristic noise patterns during image capture, and because generative models do not reproduce these patterns exactly, spatial inconsistencies can sometimes reveal synthetic origin, particularly around fine details such as hair, teeth, and the boundaries between generated and background regions.

6.2 Frequency-Domain Analysis

Generative models, and GANs in particular, have been shown to leave characteristic artifacts in the frequency domain that are not present in natural images. Techniques based on the Fourier

transform can reveal periodic patterns introduced by the upsampling operations commonly used in GAN generator architectures. Frank et al. (2020) demonstrated that such frequency-domain artifacts can serve as a robust and computationally efficient signal to distinguish GAN-generated images from real photographs and often outperform purely spatial-domain classifiers [21].

6.3 Metadata Analysis

Digital images and video files typically carry embedded metadata, including EXIF information describing the capture device, timestamp, and, in some cases, editing history. Discrepancies between claimed metadata and the actual characteristics of a file can indicate manipulation. However, metadata is trivially removable or falsifiable, and its absence or inconsistency is at best circumstantial evidence rather than a reliable authentication mechanism on its own.

6.4 Deep-Learning-Based Detectors

The dominant approach in recent forensic research trains convolutional neural networks or vision transformers directly on datasets of real and GAN-generated images to learn discriminative features. Wang et al. (2020) showed that classifiers trained on images from a single GAN architecture, combined with careful data augmentation, could generalize reasonably well to images produced by other, previously unseen generative models, although this generalization remains imperfect and degrades over time as new architectures emerge [11]. FaceForensics++, introduced by Rossler et al. (2019), established one of the most widely used benchmark datasets for training and evaluating deepfake and manipulation detectors specifically for facial video content [18].

6.5 GAN Fingerprinting

Marra et al. (2019) and Yu et al. (2019) independently demonstrated that individual GAN architectures, and in some cases individual trained model instances, leave subtle, consistent fingerprints in their generated output, analogous to the way physical camera sensors leave characteristic noise fingerprints [19, 28]. These fingerprints can, in principle, not only be used to detect that an image is synthetic but to attribute it to a specific generative model, which has useful forensic and legal implications.

6.6 Watermarking

Some generative model providers have begun embedding imperceptible watermarks directly into generated content at the point of creation, allowing downstream systems to identify synthetic origin even after moderate post-processing. Watermarking is most effective when adopted broadly and consistently by model providers, since it depends on cooperation at the point of generation rather than being inferable purely from the output itself.

6.7 Digital Signatures and Cryptographic Verification

An alternative and complementary approach authenticates content at the point of capture or creation using cryptographic signatures tied to a trusted device or identity, rather than attempting to infer authenticity purely from the content itself. A signed hash of the original file,

generated by a trusted capture device or software, allows any subsequent modification to be detected, since even a single altered pixel changes the cryptographic hash.

6.8 Provenance Systems

Provenance systems record a verifiable history of a piece of content, including who or what created it, when, using which device or software, and what edits were subsequently applied. The Coalition for Content Provenance and Authenticity (C2PA), a joint effort involving major technology and media companies, has developed an open technical standard for attaching cryptographically verifiable provenance information, referred to as Content Credentials, to digital media [29]. Adobe's Content Authenticity Initiative has been an active contributor to this standardization effort and has implemented provenance features across several of its creative products [30].

6.9 Comparing the Technique Families

Taken together, these eight technique families sit at different points on a trade-off between cost, robustness, and industry readiness, and it is worth summarizing that trade-off explicitly before turning to the limitations that motivate a layered framework. Pixel-level and spatial analysis is effective on raw, high-resolution, uncompressed assets, but its underlying signal breaks down quickly under ordinary JPEG compression or the re-encoding that social platforms apply automatically, which limits its industry suitability to raw-asset forensics rather than general deployment. Frequency-domain analysis is computationally cheap and fast enough to serve as an initial triage heuristic, and is reasonably resistant to light spatial blurring, but can be evaded by post-generation spatial filtering or by generators specifically refined to suppress the periodic upsampling artifacts it looks for. Metadata analysis costs essentially nothing to run and gives an instant deterministic check, but because metadata is trivially stripped or rewritten using widely available tools, it is only ever a supplementary signal and should not be treated as authoritative on its own. Deep-learning-based detectors, whether convolutional or transformer-based, achieve the highest accuracy against known generator architectures and can pick up subtle deepfake cues that simpler techniques miss, but they generalize poorly to unseen model architectures and remain vulnerable to adversarial manipulation, which means they require constant retraining against newly released generative models to remain useful in industry deployment.

Model fingerprinting, which maps output noise signatures back to a specific generator architecture or even a specific trained instance, is particularly valuable for legal attribution and forensic post-mortem investigation, but fails once a generator's parameters change or once its output has been heavily filtered or re-processed downstream. Watermarking offers a high degree of determinism and direct attribution to the generating vendor, but only where that vendor has chosen to embed a watermark in the first place and where the watermark survives subsequent editing, so its industry value scales directly with how broadly generative-model vendors adopt it. Cryptographic signatures, which sign an image's hash at capture using a hardware root of trust, provide what is close to a gold standard for origin trust, an essentially unforgeable mathematical proof of pixel integrity at the moment of capture, but that guarantee breaks the moment an image is edited or cropped unless the edit is captured in an auditable

provenance wrapper. Provenance systems built on standards such as C2PA extend that guarantee across an entire content lifecycle, offering the most complete visibility into a piece of content's history, but remain subject to the adoption chicken-and-egg hurdle across hardware manufacturers and platforms described in Section 11.6. In short, no individual family is simultaneously cheap, robust, and universally deployable today, which is precisely the gap the layered framework in Section 8 is designed to close by combining several of these families rather than depending on any single one.

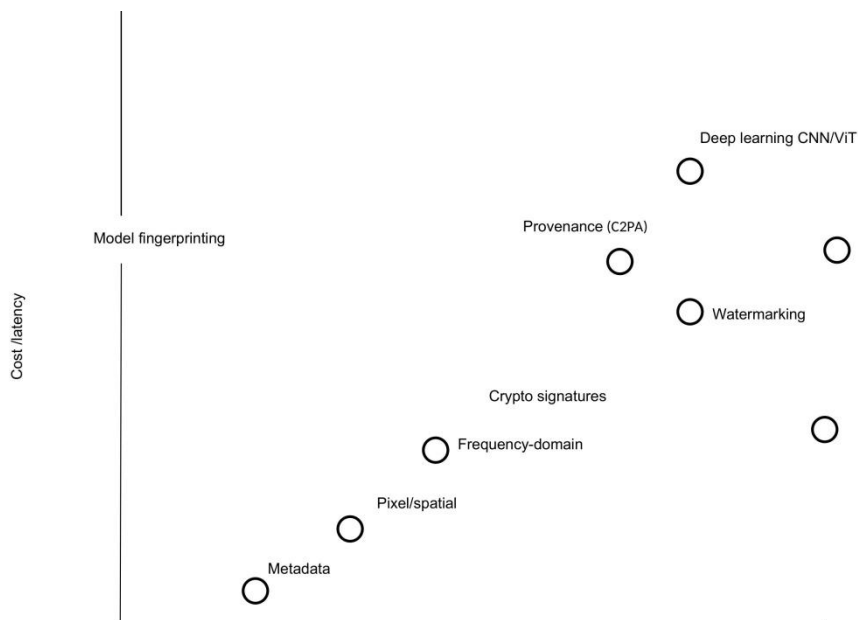


Figure 9. Detection technique families compared on robustness to evasion, compression, and new models (x-axis) versus cost and latency (y-axis).

7. LIMITATIONS OF CURRENT APPROACHES

Despite the encouraging progress, none of the approaches described above are likely to be sufficient on their own to address the industrial authentication problem, and it is important to temper expectations regarding the capabilities of current detection technology.

7.1 Poor Generalization Across GAN Architectures

A detector trained on outputs from one GAN architecture often shows dramatically worse performance on a different, structurally dissimilar architecture that it was not trained on. This is a significant limitation given that new architectures are regularly introduced in the literature, and any practical detector must be able to generalize to models not seen during training.

7.2 The Generator-Detector Arms Race

Because the generator architecture in a GAN can be modified to produce less distinctive outputs, the act of developing new GANs inherently involves reducing the distinguishability of

their outputs. This places the development of both GANs and detectors in an arms race, where each new detector prompting modification of the generator to reduce its effectiveness, and vice versa. As a result, detector performance is rarely static – a detector that may be near state-of-the-art at a given moment will inevitably become significantly less effective as new GANs are developed.

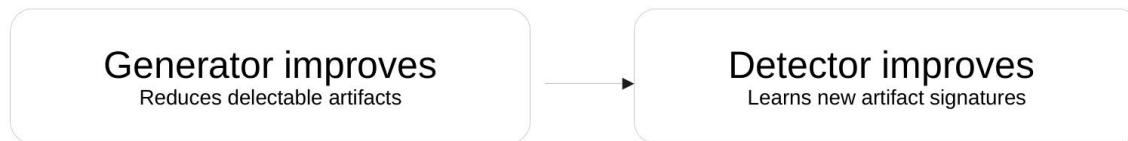


Figure 10. The generator-detector arms race: no single detector remains effective indefinitely.

7.3 Dataset Bias

Detectors trained on narrowly focused sets of data, often heavy on the facial category, generalize poorly to the other sorts of content that the industry use cases in Section 4 would need to deal with.

7.4 Robustness to Compression and Re-encoding

Content that is shared on social media platforms and messaging apps is regularly subjected to compression, shrinking, cropping, and shifting between formats. All of these transformations can obscure or eliminate the fine forensic details that many detection approaches are dependent upon; it is rare to find a detection technology that can match its laboratory performance on the real-world deployment conditions.

7.5 False Positives and False Negatives

No detector can attain perfection. False positives, where real content is falsely identified as synthetic, can damage the trust in legitimate content and its creators; false negatives, where synthetic content goes unmarked, undermine the very existence of the detector. Industry applications need to take both error modes and their respective costs into account, which tend to vary considerably by use case.

7.6 Metadata Cannot Be Fully Trusted

Because metadata can be edited, stripped, or faked using readily available tools, it cannot act as a standalone authentication mechanism and should only be treated as supporting evidence at best.

8. PROPOSED UNIFIED INDUSTRIAL FRAMEWORK

This section describes the article's core original contribution: a five-layer Unified Industrial Authentication Framework, which is proposed as a pragmatic response to the industry-problem posed in Section 1 and motivated by the generalization/robustness limitations of detection-only approaches as documented in Section 7. Given those limitations, industry applications should not employ any single detection technique but be constructed as a layered framework

combining complementary sources of evidence, culminating in a formalized authentication confidence score, rather than a binary real-or-fake determination. This section also incorporates the framework's application in the industry directly, since each sector is best understood not as a separate topic but as a different emphasis pattern applied to the same five layers.

8.0 Framework Overview

The framework is not a single algorithm; it is a layered trust architecture: source authentication, provenance tracking, multi-model AI detection, integrity verification, and confidence scoring feed into one composite decision output. This choice is industry-validated: by 2026 the Coalition for Content Provenance and Authenticity (C2PA) ecosystem has transitioned from a voluntary specification to a conformance-tested standard, with a public Conformance Program, launched in mid-2025, tracking more than fifty conforming products from over thirty companies, and hardware-level signing now shipping by default in consumer devices such as the Google Pixel 10 and professional hardware such as Sony's PXW-Z300 camcorder [31-34]. This confirms that a layered, hardware-anchored-to-software-detection pipeline is the direction the industry itself is converging on, not merely a theoretical construction.

8.1 Layer 1: Source Authentication

Where possible, content should be tied at the point of capture to a verifiable source, such as a trusted capture device or an authenticated creator identity, using a cryptographic signature. This establishes a baseline of trust anchored at content creation rather than relying solely on post-hoc analysis of the finished artifact.

To prevent tampering, the signing key itself should be protected in hardware rather than software, since a private key accessible to the device's general operating system can be extracted or misused by an attacker who has compromised that system. In practice, that means anchoring the signature in a Hardware Security Module (HSM) on server-side infrastructure, or, for consumer capture devices, in a dedicated hardware root of trust such as Apple's Secure Enclave or Android's StrongBox, which both isolate cryptographic key material from the main operating system and application processor [35]. A camera pipeline built on this model can sign a cryptographic hash of the raw sensor output before any post-processing, compression, or editing occurs, so that the signature locks the signed content to the pixels as they existed at capture, with any downstream processing captured separately as an auditable edit in the provenance chain described in Layer 2, rather than silently invalidating or bypassing the original signature. The C2PA specification explicitly supports this hardware-anchored model, defining how a hardware root of trust at capture can serve as the first, strongest link in an otherwise software-based provenance chain [36].

8.1.1 Handling Unverified or Absent Source Signatures

A necessary refinement to the layer as formulated is that the framework should not treat the absence of a verifiable source signature as equivalent to a positive finding of synthetic content. Most content that is actually processed in a real pipeline, such as a user-submitted insurance photo, a social media upload, or a legacy archival file, will simply never have been signed at the point of capture, since hardware-rooted signing is still an emerging capability rather than a

universal one [33, 34]. Treating an unsigned file as 'fake' would generate an unworkable rate of false positives against ordinary, innocent content.

The framework should treat an absent or unverifiable source signature as a downgrade to 'suspicious, unverified,' not a classification of 'fake.' This is what the Authentication Confidence Score (ACS) formula in Section 8.5 does mechanically: setting $S = 0.0$ lowers the composite score and pulls the classification into 'suspicious,' but does not by itself force the composite score to zero, since Layers 2 through 4 can still independently support authenticity. This distinction is operationally important: a compliance officer or reviewer should see 'insufficient provenance, escalate for review' rather than 'flagged as fake,' since the correct remedy in most such cases is human review or request for additional evidence, not automatic rejection.

A second refinement concerns adversarial tampering with the signature itself. An attacker who succeeds in stripping or breaking a cryptographic hash binding does not thereby pass authentication; the framework is designed so that this failure mode is itself detected and surfaced rather than silently ignored. A broken or mismatched hash is captured directly by Layer 4 ($I = 0.0$) and, where the break occurs mid-chain, it also registers as a discontinuity in Layer 2's provenance graph, since the gap-penalty term in the provenance-integrity calculation is specifically designed to catch this: an otherwise complete-looking chain with a detected break is penalized rather than scored as though the break were absent. The practical effect is that an attacker who defeats the hash does not obtain a clean 'authentic' verdict; the attempt instead produces a chain with a visible discontinuity, which the framework routes to 'suspicious, manual review' rather than either 'trusted' or a silent pass-through. This treats broken cryptographic evidence as informative, a signal of probable tampering, rather than simply missing.

8.2 Layer 2: Provenance Tracking

Building upon the C2PA model, the framework records a verifiable chain of custody over creation, editing, processing, and publishing. Each step is cryptographically tied to the previous step in the chain, so that any gap is visible to a verifier.

8.3 Layer 3: Multi-Model AI Detection

Rather than using a single classifier, the framework runs content through multiple, architecturally-diverse detectors in parallel, spanning analysis in the spatial, frequency, and metadata domains, as well as convolutional and transformer-based deep learning detectors. Diversity within the detection approaches reduces the likelihood of a single generative model capable of evasion across all of them simultaneously.

Running several heavy models in parallel is not free, and an industry deployment cannot ignore the cost. Transformers have meaningful inference latency and compute cost, and invoking 4 or more on every piece of content is economically difficult to justify, and often impossible to achieve within the approximately 3-second window of a live bank KYC video check. A realistic deployment thus tiers the detection stage, rather than invoking every model on every item. A first pass applies only light and low-latency heuristics, such as frequency-domain artifact checks and basic metadata consistency checks, both computationally light relative to a deep neural network. Only content flagged by this first pass as ambiguous or suspicious, or content

associated with inherently higher-risk transaction such as a large wire transfer or a new-account KYC check, is escalated to the full battery of heavier convolutional and transformer-based detectors. This tiered, risk-adaptive design keeps routine, low-risk content flowing through the system at a low cost and low latency, while reserving the most expensive detection resources for the smaller volume of content where the marginal accuracy gain is most likely to matter.

8.4 Layer 4: Integrity Verification

Independently of whether content is judged synthetic or genuine, the framework verifies whether a file has been modified since it was originally signed or registered, using cryptographic hash comparison. This layer addresses manipulation of an otherwise-authentic content, a distinct problem from synthetic generation but relevant to industry authentication needs.

8.5 Layer 5: Authentication Confidence Scoring

Rather than emitting a simple real-or-fake label, the framework aggregates evidence from the preceding layers into a composite authentication confidence score representing source authenticity, provenance integrity, AI detection confidence, and metadata consistency. Presenting authentication as a graded confidence score rather than a binary verdict better reflects the genuine uncertainty involved, and provides downstream decision-makers (compliance officers, editors, claims adjusters) with the information they need to apply their own risk thresholds rather than deferring to an automated decision.

The composite score is formalized as a normalized scalar bounded between zero and one: $ACS = (w_1 \times S) + (w_2 \times P) + (w_3 \times (1-D)) + (w_4 \times I)$, subject to the constraint that the four weights w_1 , w_2 , w_3 , and w_4 be non-negative and sum to one. Each of the four terms is itself a normalized value between zero and one, defined as follows.

Source authenticity, S , takes the value 1.0 when content is signed by a verified hardware root of trust such as an HSM or a Secure Enclave; 0.5 when signed only by a software-based key or an otherwise unverified public-key-infrastructure anchor; and 0.0 when unsigned or signature fails verification. Provenance integrity, P , is calculated as the number of valid, cryptographically linked C2PA assertions in the content's history divided by the total number of history claims asserted for that content, multiplied by a factor of one minus a gap-penalty indicator which is set to 1 whenever an otherwise unbroken chain of custody includes a detected gap, and 0 otherwise. This formulation rewards both completeness and continuity of the provenance record rather than completeness alone. The AI generator risk index, D , is evaluated conservatively as the maximum synthetic-probability output across all parallel detectors run in Layer 3 (whether vision transformer, frequency-domain analyzer, or space CNN), so that a single strongly positive detector is sufficient to raise the overall risk index even if other detectors in the ensemble disagree. This ensemble max-risk aggregation is a deliberately conservative choice, because understating synthetic risk is typically more costly to an organization than occasionally overstating it. Hash integrity, I , takes 1.0 when a cryptographic hash comparison confirms the content matches its originally registered payload, and 0.0 when a hash mismatch indicates unauthorized pixel manipulation has occurred.

The weight vector w_1 through w_4 should be tuned to an organization's own threat profile, rather than universal tuning. The relative ordering of the weights, rather than any particular numeric value, should differ across sectors. A bank performing KYC onboarding should place more weight on source authentication and AI detection than on provenance, since live-camera verification and deepfake prevention matter more than a long-term provenance history which is typically unavailable for a first-time customer. A legal team assessing evidence admissibility should place the greatest weight on source authentication and provenance, since admissibility questions turn on an unforgeable chain of custody rather than on a statistical judgment of whether an image looks artificial. A social media platform screening user-generated content (which is rarely signed or provenance-tracked at the point of upload) should weight AI detection most heavily, since it's effectively the only signal available on that scale. An insurance claims processor should adopt a more evenly balanced profile across all four evidence types, reflecting comparable concern about synthetic image generation and after-the-fact editing of otherwise genuine claim photographs.

Because the same underlying evidence can be weighted differently depending on which of these profiles an organization adopts, the composite score of a single piece of content could reasonably fall on either side of an organization's own trust threshold, depending on which evidence sources that organization prioritizes. This is an argument for why the weight vector must be tuned locally (using an organization's own historical and operational data) rather than fixed as a single industry-wide default. No specific weight values or resulting scores are asserted here as recommendations; an organization adopting this framework would need to derive its own weights and thresholds empirically, for example by testing candidate weight vectors against its own labeled outcomes, and should expect to revisit the calibration periodically as its evidentiary mix (such as the proportion of content arriving with hardware signed provenance) shifts over time.

For enterprise systems with richer forensic telemetry, this linear weighted model can be extended into a Bayesian odds-update formulation which accommodates conditional dependencies among layers rather than treating them as independent linear contributions. In this formulation, the posterior odds of authenticity given the accumulated evidence equal the prior odds of authenticity for that content channel, multiplied by the product, across all evidence sources, of the likelihood of observing that evidence under the authentic hypothesis divided by the likelihood of observing it under the synthetic hypothesis. The likelihood-ratio structure allows the system to express, for example, that a Layer 4 hash mismatch should sharply discount an otherwise high Layer 3 detection confidence, rather than average the two signals as the linear model would. Developing well-calibrated likelihood ratios for each evidence source requires substantially more labeled data than the linear model does, so the Bayesian formulation is a refinement path for organizations with mature forensic telemetry, rather than a starting point for initial deployment.

The overall architecture can be summarized as an ingestion stage which routes content through metadata analysis, forensic analysis, and provenance checking; an AI detection stage which applies multiple specialized detectors; a risk assessment stage which aggregates these signals; a cryptographic verification stage which checks integrity against any existing signature; and a

final decision engine which outputs an authentication confidence score, plus recommended classification of trusted, suspicious, or unverified.

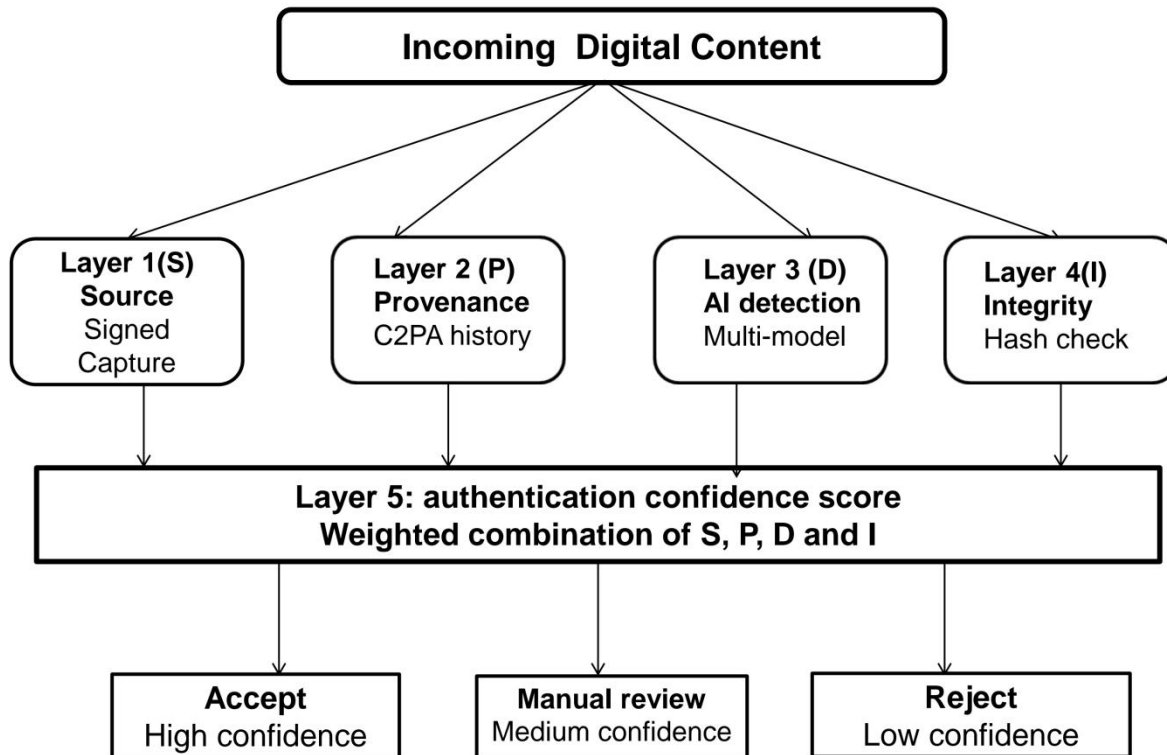


Figure 11. Multi-Layer Digital Content Authentication Framework

8.6 Industry Application of the Framework

Each sector can be seen as a different emphasis pattern across the five layers, driven by what evidence is realistically available and what failure mode is most costly. A qualitative table showing which evidence layer each sector should prioritize is provided as a standalone reference in the accompanying Excel workbook, so it can be adapted and re-sorted independently of this manuscript.

8.6.1 Banking and Financial Services

Layered authentication can strengthen KYC onboarding, remote identity verification, and video-based approval workflows, using liveness detection, GAN and deepfake detection, and, where feasible, source authentication for submitted identity documents and video calls. Live deepfake injection into liveness checks is a measured, fast-growing attack vector: the Association of Certified Fraud Examiners and SAS's March 2026 survey of 713 anti-fraud professionals found 77% reporting a sharp rise in deepfake-enabled social engineering, while only 7% considered their organization more than moderately prepared to counter it [26, 27]. This makes Layer 3 (tiered multi-model detection) and Layer 1 (device-level signing) the dominant layers for this sector; provenance history is largely unavailable for a first-time customer.

8.6.2 Healthcare

Provenance tracking and cryptographic signing at the point of image capture can help ensure that diagnostic scans and patient records reaching clinicians have not been altered since acquisition, supplementing existing controls in radiology and diagnostic imaging systems. Layers 1 and 2 dominate here, since the threat model is mainly post-acquisition tampering of genuine images, rather than wholly synthetic scans.

8.6.3 Media and Journalism

News organizations can adopt provenance standards (such as C2PA) to attach verifiable Content Credentials to photographs and video at the point of capture, allowing readers and platforms to verify authenticity before publication or resharing, and can apply multi-model AI detection to user-submitted content of unknown origin. This is increasingly the industry's own chosen direction: major AI platforms (Adobe Firefly, OpenAI's image/video tools, Google Imagen) now embed Content Credentials by default, and platforms such as TikTok joined the C2PA Steering Committee in 2026 to help scale adoption [33, 34]. Layer 3 is still necessary for user-submitted content of unknown origin which was never signed at capture.

8.6.4 Legal and Evidentiary Use

Cryptographic signing and provenance chains can strengthen the evidentiary weight of digital media submitted in legal proceedings, while AI detection and integrity verification can flag content warranting closer forensic examination before admission. Layers 1, 2, and 4 dominate here, since admissibility questions turn on unbroken custody and pixel integrity rather than a statistical detection score.

8.6.5 Insurance

Authentication confidence scoring of claim photographs and videos can help insurers triage claims for manual review, focusing investigative resources on submissions with low authentication confidence, rather than treating all submissions as equally trustworthy. A relatively even mix across all four evidence layers is appropriate here, since insurers face both wholly fabricated claim photographs and after-the-fact edits of genuine ones.

8.6.6 Government and Public Communications

Provenance-backed publishing of official video and audio statements can help the public and media distinguish authentic government communications from deepfake impersonations, complementing broader efforts such as the labeling requirements for synthetic content introduced under the European Union's Artificial Intelligence Act [37]. Provenance-backed publishing (Layers 1-2) is the anchor for this sector, reinforced by regulatory convergence, since the EU AI Act's Article 50 transparency obligations for synthetic content take effect in August 2026, and C2PA's AI-assertion type is designed to directly satisfy that requirement [33, 37].

9. VALIDATION – THREATS TO VALIDITY

This section contains a review of the limitations of the claims made in Section 8, following standard threats-to-validity practice for framework-proposal, rather than purely empirical, research. It also specifies the statistical methodology by which the Layer 3 detector ensemble, and thus the entire framework's recommendations, should be validated before industrial deployment.

9.1 Statistical Evaluation Methodology for Detector Validation

As context for the claims in Table 2 and the proposed evaluation methodology, it is vital to recognize that no single metric captures a detector's entire real-world utility. Detectors' operating characteristics should be expressed in terms of the full multi-metric family, rather than one scalar like accuracy. Precision, the fraction of flagged content that is synthetic, and recall, the fraction of synthetic content that is correctly flagged, should each receive separate reporting, as operational costs of false positives and false negatives are often significantly imbalanced (Section 7.5). The F1 score, which is the harmonic mean of precision and recall, serves as an adequacy metric in cases where these utilities have comparably high costs. The area under the receiver operating characteristic (AUC-ROC) curve, which summarizes a detector's operating characteristics across threshold values, is preferable to individual points along that curve (such as the default accuracy threshold) when it is desirable to select an operating point according to a specific cost-benefit regime (such as the risk-adaptive escalation framework of Section 8.3).

When comparing two candidate anomaly detectors, or when determining whether a detector retrained on new data significantly improves upon a prior version, paired comparisons using a significance test on the differing counts (McNemar's test) are preferable to raw accuracy comparisons, as the same content is tested on both detectors, violating the independence assumption. Reported accuracy figures, F1 scores, and AUC-ROC values should each have an accompanying confidence interval (e.g., a Wilson or bootstrap interval) when evaluated on smaller, harder-to-generalize subsets of content (e.g., content generated by a newly emerged generation family), as their variance may not decrease predictably with larger sample sizes.

When choosing between two operating points on the precision-recall curve, or other trade-off curves like the false-positive rate vs. false-negative rate, a cost-sensitive utility analysis should be performed instead of choosing between metrics individually. For example, the expected costs for a given threshold would be calculated as $E[\text{cost}] = (\text{FPR} \times C_{\text{fp}} \times N_{\text{authentic}}) + (\text{FNR} \times C_{\text{fn}} \times N_{\text{synthetic}})$, where C_{fn} and C_{fp} are the costs to the business of a false negative

and false positive, respectively, and $N_{\text{synthetic}}$ and $N_{\text{authentic}}$ are estimates of the volumes of synthetic and authentic content. A bank's KYC pipeline, where a false negative could allow account takeover fraud, would have a much higher C_{fn} than C_{fp} , requiring a lower threshold to balance these costs, while a media publisher might reverse this ordering to avoid marking legitimate submissions as synthetic.

9.2 Threats to Construct Validity

The qualitative ratings in Table 2 below, regarding robustness, generality, and cost, are based on synthesizing the literature, rather than performing a controlled experiment with one unified evaluation protocol. Different papers report different accuracy scores for the techniques they analyze, using different evaluation sets, generator families, and compression levels. As such, when one technique family has a 'Moderate' rating and another has a 'High' rating, this reflects relative, rather than absolute, differences in their ability to generalize.

9.3 Threats to Internal Validity

The calculation of weights in Section 8.5 presumes that the evidence values derived from the detector responses are correct, without separately evaluating this assumption. The discussion focuses exclusively on the implications of incorrect weight choices, rather than incorrect evidence values. If the Layer 3 ensemble's synthetic-probability values or the Layer 2 provenance-completeness ratios are incorrect, then any weighting choice fails to address this detector performance issue, which is a separate concern from the weight-choice issue discussed in this section.

9.4 Threats to External Validity

The threat to external validity is primarily that the framework, and the illustrative sector weight profiles it proposes, might not generalize to real-world data as effectively as the experiments suggest. This framework's recommendations should be regarded as preliminary until they have been piloted on relevant data with ground-truth labels for each sector (e.g., with fraud data where the fraudulent edits can be identified, to evaluate the framework's ability to detect them).

9.5 Detector Generalization Threat

The threat described in detail in the paper, that detectors trained on one generator family often perform poorly on another [14, 34], is exacerbated by the fact that the field is shifting away from GANs towards diffusion models as the dominant generation architecture. The literature now demonstrates that detectors fine-tuned on GAN frequency artifacts generalize poorly to diffusion outputs and vice versa, due to each architecture's unique statistical properties [38-40]. Accuracy figures for individual generator families should thus be interpreted as conditional on that family, rather than as general performance claims, and this threat will only grow as new architectures emerge.

Table 2 below summarizes all technique families' reported robustness, provides a qualitative rating of the generalization threat affecting each, and supplies a qualitative confidence rating for the validity of each entry. It extends the technique comparison of Table 1 in Section 6.9 by including this validity analysis.

Table 2. Qualitative performance profile of the eight detection and authentication technique families, synthesized from Section 6.9.

Technique Family	Robustness to Compression	Cross-Model Generalization	Relative Cost / Latency
Pixel-level / spatial analysis (6.1)	Low	Low	Low
Frequency-domain analysis (6.2)	Moderate	Low–Moderate	Low
Metadata analysis (6.3)	Low (easily stripped)	Not applicable	Very low
Deep-learning detectors (6.4)	Moderate	Moderate (degrades on unseen architectures)	High
GAN fingerprinting (6.5)	Low–Moderate	Low (architecture-specific)	Moderate
Watermarking (6.6)	Moderate–High*	Not applicable	Low
Cryptographic signatures (6.7)	High**	Not applicable	Low
Provenance / C2PA (6.8)	High**	Not applicable	Moderate

* Contingent on broad vendor adoption of the watermark standard. ** Contingent on an unbroken, auditable provenance chain; the guarantee degrades once content is edited outside that chain.

Table 3. Sector-wise priority of evidence layers in the framework.

Sector	Highest Priority Evidence	Second Priority	Lowest Priority	Why

Sector	Highest Priority Evidence	Second Priority	Lowest Priority	Why
Banking / KYC onboarding	AI Detection (Layer 3)	Source Authentication (Layer 1)	Provenance (Layer 2)	Live-camera deepfake injection is the dominant, fast-growing threat; first-time customers have no provenance history to draw on
Legal / evidentiary	Source Authentication (Layer 1)	Provenance (Layer 2)	AI Detection (Layer 3)	Admissibility turns on unforgeable chain of custody, not on a statistical 'looks synthetic' score
Social media platforms	AI Detection (Layer 3)	Provenance (Layer 2)	Source Authentication (Layer 1)	Content is rarely signed at upload; detection is often the only signal available at scale, and platforms frequently strip provenance metadata on ingest
Insurance claims	Roughly balanced across all four layers	---	---	Both wholly fabricated images and edited-but-genuine photos are common threats, so no single layer dominates
Healthcare imaging	Provenance (Layer 2)	Integrity Verification (Layer 4)	AI Detection (Layer 3)	Threat model is mainly post-acquisition tampering of genuine scans, not synthetic generation

Sector	Highest Priority Evidence	Second Priority	Lowest Priority	Why
Government / public communications	Provenance (Layer 2)	Source Authentication (Layer 1)	AI Detection (Layer 3)	Anchoring official statements to a signed, published chain of custody matters more than statistical detection after the fact

10. DISCUSSION

10.1 What this Framework Is

The proposed framework is best understood as a decision-support layer, not an automated verdict engine. It does not claim to decide, with certainty, whether a piece of content is real or synthetic; instead, it aggregates four different types of evidence—cryptographic origin, custody history, statistical detection, and byte-level integrity—into a normalized confidence score explicitly leaving the final accept, reject, or escalate decision to a human reviewer or a downstream policy engine calibrated to that organization's risk tolerance.

10.2 Advantages

- Fails gracefully rather than catastrophically: because no single layer's failure zeroes out the score, an attacker must defeat multiple evidence types to force a false 'trusted' classification.
- Matches real-world evidence availability: most content won't have hardware-signed provenance at least for years to come, and the framework degrades that content to 'suspicious,' not 'fake,' which is the operationally correct response given current, uneven C2PA adoption [33].
- Sector-tunable rather than one-size-fits-all: the same underlying evidence can legitimately produce different classifications in banking versus journalism versus insurance, matching how these industries reason about risk.
- Economically realistic: the tiered Layer 3 design keeps routine content moving at low cost while reserving expensive transformer ensembles for content flagged as ambiguous or high-risk.
- Adversarially informative rather than silent: as established in Section 8.1.1, an attacker who breaks a hash or strips a signature produces a visible discontinuity that routes to manual review, rather than a clean pass-through.

10.3 Limitations of This Work

- No empirical validation against real fraud outcomes: the framework and its illustrative weight profiles are reasoned about, not tested against labeled ground-truth datasets of confirmed fraud or confirmed synthetic media in production.

- Provenance layers assume adoption that is still incomplete: Layers 1 and 2 lose most of their value on content from unsigned devices or after a platform strips Content Credentials on upload, a documented, current problem rather than a hypothetical one.
- Detector generalization remains an open problem the architecture can't solve on its own: layering multiple detectors reduces, but doesn't eliminate, the risk that a novel generator defeats the entire Layer 3 ensemble simultaneously.
- Weight-tuning requires data most organizations don't yet have: setting the ACS weights responsibly requires a labeled evaluation set reflecting an organization's actual content mix.
- The Bayesian extension in Section 8.5 is not yet practicable for most deployments, since well-calibrated likelihood ratios require more labeled data than most organizations can supply.

10.4 Future Work Arising from this Study

Future work should prioritize: piloting the framework against real labeled fraud and synthetic-media datasets per sector to replace illustrative weights with empirically fitted ones; extending the Layer 3 detector ensemble evaluation to diffusion and hybrid generators rather than GAN-centric benchmarks; and tracking the ongoing C2PA Conformance Program data as an independent signal of how much of Layers 1 and 2's promise is realized in production rather than only specified on paper. A fuller treatment of future research directions is given in Section 12.

11. CHALLENGES AND ETHICAL CONSIDERATIONS

11.1 Real-Time Performance

Many industry use cases, such as live video verification during account onboarding, require authentication decisions within seconds, placing constraints on the computational complexity of the multi-model detection layer described in Section 8.

11.2 Explainability

Users are unlikely to trust an authentication system that produces a score without justification. Industry deployments should prioritize detection approaches that can indicate which specific evidence, such as a frequency-domain anomaly or provenance chain gap, contributed to a given confidence score.

11.3 Privacy

Authentication systems (particularly those involving biometric or identity verification) must collect and retain no more personal information than strictly necessary and should be designed in accordance with applicable data protection regulations.

11.4 Cross-Platform Durability

Content often passes through multiple platforms and undergoes repeated compression and re-encoding from capture to final consumption. Provenance and watermarking mechanisms need

to survive this journey, which is an open technical challenge, particularly for watermarking schemes.

11.5 Continuous Model Updating

Because of the generator-detector arms race described in Section 7, detection models embedded in any industry authentication system require ongoing retraining/monitoring against newly released generative models, implying a recurring operational cost rather than a one-time deployment.

11.6 Standardization and the Adoption Hurdle

The effectiveness of provenance-based approaches in particular depends heavily on coordinated adoption across camera manufacturers, software vendors, and publishing platforms. Continued development of open standards such as C2PA, and emerging regulatory requirements for synthetic content labeling, will likely play a significant role in determining how quickly industry-wide authentication infrastructure matures.

This dependency creates a chicken-and-egg adoption problem that industry stakeholders shouldn't underestimate: a provenance chain is only as strong as its weakest unadopted link: camera and smartphone manufacturers have little incentive to embed signing capability at capture until publishers and platforms are prepared to display and honor Content Credentials, and vice versa. Early movers bear disproportionate cost for benefits that only materialize once adoption reaches critical mass across the ecosystem. Overcoming this dynamic will likely require a combination of regulatory pressure (such as synthetic content labeling requirements), coordinated commitments from a critical mass of device manufacturers and major platforms acting together, and demonstrable business value (such as fraud-loss reduction) that justifies adoption well before network effects alone would make the case.

12. FUTURE SCOPE

The preceding sections have deliberately focused on GANs as a technically tractable and historically foundational lens for the authentication problem, while noting throughout that the underlying threat landscape has already broadened beyond GANs alone. This section outlines the directions in which the field (and the layered framework proposed in Section 8) are likely to need to evolve over the coming years.

12.1 Beyond GANs: Diffusion, Multimodal, and Agentic Generation

Diffusion models now match or exceed GAN-based generators in image/video realism using a fundamentally different generative mechanism, and text-to-video/audio are rapidly advancing. Future work should extend the empirical evaluation underlying Section 9 to diffusion-based and hybrid architectures specifically, since the frequency-domain and fingerprinting signatures discussed in Sections 6.2 and 6.5 are known to differ between GAN and diffusion outputs, and a detector ensemble tuned primarily against GAN artifacts may systematically underperform against diffusion-generated content unless retrained/revalidated against it [38–40]. Looking further ahead: agentic AI systems capable of autonomously producing/distributing synthetic

content at scale rather than a human operator generating a single artifact, will require authentication infrastructure that can attribute content to an originating system/workflow, rather than a generative model architecture.

12.2 On-Device and Real-Time Detection

Section 11.1 identifies real-time performance as a present constraint on the Layer 3 multi-model detection stage. Continued work on model compression/quantization and knowledge distillation of transformer-based detectors is a promising path to running a meaningful subset of the detection ensemble directly on capture devices, such as smartphones, rather than only in server-side infrastructure. On-device detection would also reduce the privacy exposure discussed in Section 11.3, since content wouldn't need to leave the device for an initial screening pass, only escalating ambiguous/high-risk content to server-side analysis under the tiered design already described in Section 8.3.

12.3 Standardized Interoperability and Regulatory Convergence

Section 11.6 describes the adoption chicken-and-egg problem facing provenance standards such as C2PA. A plausible future path is convergence between voluntary technical standards and binding regulatory requirements: the synthetic content labeling provisions of the European Union's Artificial Intelligence Act's Article 50 transparency obligations, take effect in August 2026, and comparable frameworks likely to emerge in other jurisdictions, would give camera manufacturers, software vendors, and platforms a shared compliance incentive to adopt a common provenance format rather than competing, incompatible schemes [33, 37]. Future research/standards work should focus on interoperability testing across vendors and defining minimum verifiable-provenance requirements that regulators and industry bodies can reference directly, reducing the ambiguity that makes early adoption costly for any single actor.

12.4 Federated and Privacy-Preserving Detection

Centralizing large volumes of user content for detector training/validation, as described in Section 9.1, raises the privacy concerns discussed in Section 11.3—particularly for biometric/identity-verification use cases. Federated learning, in which detection models are trained across distributed data holders without the underlying content ever leaving its original custody, and privacy-preserving techniques such as differential privacy and secure multi-party computation, are promising directions for allowing industry-wide detector improvement without requiring any one organization to aggregate content centrally. Maturing these techniques to the accuracy levels achievable with centralized training is an open research problem.

12.5 Block chain and Distributed-Ledger Provenance Anchoring

The provenance chain described in Layer 2 (Section 8.2) is, in the framework as presented, verified against records held by the signing organization or the C2PA ecosystem. An open direction for future work is anchoring provenance assertions (or cryptographic commitments to them) in a public, tamper-evident distributed ledger: this would allow any verifier to confirm that a provenance record hasn't been retroactively altered without needing to trust the record-holder's own infrastructure. This would strengthen the legal/evidentiary use case described in

Section 8.6.4 in particular, though it introduces its own scalability, cost, and key-management considerations that would need to be resolved before industry-scale deployment.

12.6 Continual and Adversarial Robust Detection

Section 11.5 notes that the generator-detector arms race described in Section 7.2 implies a recurring retraining cost rather than a one-time deployment. Future work on continual learning techniques that update detectors incrementally as new generator families emerge, without suffering catastrophic forgetting of previously learned artifact signatures, would reduce this cost. Similarly, adversarial training against deliberately evasive generators (rather than only naturally occurring generator output) is likely to become increasingly important as generator developers (intentionally or as a side-effect of pursuing greater realism) continue to reduce the very artifacts on which current detectors depend, a dynamic already visible in the scale of deepfake-enabled fraud growth documented through 2025 and 2026 [26, 27, 41, 42].

13. CONCLUSION

GAN-generated content is not just a technological curiosity: it's a trust/authentication problem with direct consequences for banking, healthcare, media, insurance, legal systems, e-commerce, and government. This article has argued that no single detection technique is likely to remain effective against the full range of present and future generative models, given the demonstrated generalization limitations of existing detectors and the ongoing arms race between generative and forensic technologies. An effective industry solution should combine cryptographic source authentication, provenance tracking, multi-model AI detection, integrity verification, and continuous monitoring, aggregated into a graded authentication confidence score rather than a binary verdict. Critically, the framework must treat an absent signature or broken hash as a signal to escalate for review (since most content in circulation today was never signed at capture at all). As generative models continue to improve, industry investment in this kind of layered, continuously updated authentication infrastructure will be essential to preserving trust in digital content.

References

- [1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 27, 2014, pp. 2672–2680.
- [2] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive Growing of GANs for Improved Quality, Stability, and Variation," in *International Conference on Learning Representations (ICLR)*, 2018.
- [3] A. Brock, J. Donahue, and K. Simonyan, "Large Scale GAN Training for High Fidelity Natural Image Synthesis," in *International Conference on Learning Representations (ICLR)*, 2019.
- [4] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4401–4410.

- [5] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and Improving the Image Quality of StyleGAN,” in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8110–8119.
- [6] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks,” in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2017, pp. 2223–2232.
- [7] J. Ho, A. Jain, and P. Abbeel, “Denoising Diffusion Probabilistic Models,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020, pp. 6840–6851.
- [8] P. Dhariwal and A. Nichol, “Diffusion Models Beat GANs on Image Synthesis,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 34, 2021, pp. 8780–8794.
- [9] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-Resolution Image Synthesis with Latent Diffusion Models,” in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10684–10695.
- [10] M. Kang, J.-Y. Zhu, R. Zhang, J. Park, E. Shechtman, S. Paris, and T. Park, “Scaling Up GANs for Text-to-Image Synthesis,” in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10124–10134.
- [11] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, “CNN-Generated Images Are Surprisingly Easy to Spot... for Now,” in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8695–8704.
- [12] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, “On the Detection of Synthetic Images Generated by Diffusion Models,” in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5.
- [13] Grand View Research, “Generative Adversarial Networks Market Size Report, 2030,” 2025.
- [14] Acumen Research and Consulting, “Generative Adversarial Networks Market Size, Report 2035,” Feb. 2026.
- [15] Stanford Institute for Human-Centered AI, “The 2026 AI Index Report,” Apr. 2026.
- [16] Stanford HAI, “Economy,” The 2026 AI Index Report, 2026.
- [17] Deloitte Center for Financial Services, “Generative AI is expected to magnify the risk of deepfakes and other fraud in banking,” Deloitte Insights, 2024.
- [18] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “FaceForensics++: Learning to Detect Manipulated Facial Images,” in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2019, pp. 1–11.
- [19] N. Yu, L. S. Davis, and M. Fritz, “Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints,” in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2019, pp. 7556–7566.
- [20] R. Durall, M. Keuper, and J. Keuper, “Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks Are Failing to Reproduce Spectral Distributions,” in

- Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7887–7896.
- [21] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, “Leveraging Frequency Analysis for Deep Fake Image Recognition,” in Proc. Int. Conf. Machine Learning (ICML), PMLR vol. 119, 2020.
- [22] U. Ojha, Y. Li, and Y. J. Lee, “Towards Universal Fake Image Detectors that Generalize Across Generative Models,” in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24480–24489.
- [23] Coalition for Content Provenance and Authenticity (C2PA), “Content Credentials Deployment Guidance,” 2026.
- [24] X. Zhao, K. Zhang, Z. Su, S. Vasan, I. Grishchenko, C. Kruegel, G. Vigna, Y.-X. Wang, and L. Li, “Invisible Image Watermarks Are Provably Removable Using Generative AI,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 37, 2024.
- [25] CNN Business, “Finance Worker Pays Out \$25 Million After Video Call With Deepfake ‘Chief Financial Officer’,” Cable News Network, May 16, 2024.
- [26] Association of Certified Fraud Examiners and SAS, “AI-Driven Fraud Survey (713 anti-fraud professionals, 8 regions),” Mar. 2026, as reported in “Deepfake-Era KYC: Why Documents and Selfies are No Longer Enough.”
- [27] ASIS International / Shufti, “Identity Fraud Index Report: Deepfake Identity Fraud Poised to Increase Nearly 500 Percent in 2026,” Jun. 2026.
- [28] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi, “Do GANs Leave Artificial Fingerprints?,” in Proc. IEEE Conf. Multimedia Information Processing and Retrieval (MIPR), 2019.
- [29] Coalition for Content Provenance and Authenticity (C2PA), “C2PA Technical Specification,” Joint Development Foundation, 2024.
- [30] Adobe, “Content Authenticity Initiative: Overview and Implementation Guide,” Adobe Inc., 2023.
- [31] C2PA, “The C2PA Launches Content Credentials 2.3 and Celebrates 5 Years of Impact Across the Digital Ecosystem,” Linux Foundation Projects, Feb. 9, 2026.
- [32] C2PA, “Content Credentials Deployment Guidance (Conformance Program status report),” c2pa.org, May 14, 2026.
- [33] Content Authenticity Initiative, “The State of Content Authenticity in 2026,” contentauthenticity.org, 2026.
- [34] J. Wondrasek, “C2PA Adoption in 2026: Hardware Platforms and Verification Reality,” softwareseni.com, Mar. 3, 2026.
- [35] Apple Inc., “Secure Enclave Overview,” Apple Platform Security Guide, 2023.
- [36] Coalition for Content Provenance and Authenticity (C2PA), C2PA Technical Specification, Section on Trust Models, Joint Development Foundation, 2024.

- [37] European Commission, “Regulation (EU) 2024/1689 (Artificial Intelligence Act) - Provisions on Synthetic Content Transparency and Labeling,” Official Journal of the European Union, 2024.
- [38] Springer Nature, “Deepfake Detection Across Image, Video, and Audio: A Comprehensive Survey with Empirical Evaluation of Generalization and Robustness,” Artificial Intelligence Review, 2026.
- [39] MDPI Applied Sciences, “Advancing GAN Deepfake Detection: Mixed Datasets and Comprehensive Artifact Analysis,” vol. 15, no. 2, p. 923, 2025.
- [40] Liu et al., “A Review of Deepfake and Its Detection: From Generative Adversarial Networks to Diffusion Models,” International Journal of Intelligent Systems, Wiley, 2025.
- [41] Zyphe, “30+ Sourced Identity Verification Statistics for 2026 (citing Sumsb, Entrust, and FBI IC3 data),” 2026.
- [42] LexisNexis Risk Solutions, “Deepfake Identity Fraud Report,” Jul. 14, 2026.
- [43] X. Zhang, S. Karaman, and S.-F. Chang, “Detecting and Simulating Artifacts in GAN Fake Images,” in Proc. IEEE Int. Workshop on Information Forensics and Security (WIFS), 2019.
- [44] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, “DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection,” Information Fusion, vol. 64, pp. 131–148, 2020.
- [45] L. Verdoliva, “Media Forensics and DeepFakes: An Overview,” IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 5, pp. 910–932, 2020.
- [46] Y. Mirsky and W. Lee, “The Creation and Detection of Deepfakes: A Survey,” ACM Computing Surveys, vol. 54, no. 1, pp. 1–41, 2021.
- [47] H. Farid, “Creating, Using, Misusing, and Detecting Deep Fakes,” Journal of Online Trust and Safety, vol. 1, no. 4, 2022.
- [48] National Institute of Standards and Technology (NIST), “Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency,” NIST AI 100-4, 2023.
- [49] R. Chesney and D. Citron, “Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security,” California Law Review, vol. 107, p. 1753, 2019.
- [50] D. Cozzolino, J. Thies, A. Rossler, C. Riess, M. Niessner, and L. Verdoliva, “ID-Reveal: Identity-aware DeepFake Video Detection,” in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2021.
- [51] c2paviewer.com, “What is C2PA [technical overview, regulatory and CISA advisory summary],” 2026.