information

MDPI

Review

A Review of Tools and Technologies to Combat Deepfakes

Dmitry Erokhin * and Nadejda Komendantova

International Institute for Applied Systems Analysis, A-2361 Laxenburg, Austria
* Correspondence: erokhin@iiasa.ac.at

Abstract

Deepfakes and adjacent synthetic-media capabilities have become a systemic challenge for information integrity, security, and digital trust. Countermeasures now span passive detection methods that infer manipulation from content traces, active provenance systems that cryptographically bind metadata to media, and watermarking approaches that embed detectable signals into content or generative processes. This review presents a rigorous synthesis of tools and technologies to combat deepfakes across modalities (image, video, audio, and selected multimodal settings), drawing primarily from the peer-reviewed literature, standardized benchmarks, and official technical specifications and reports. The review analyzes detection methods, provenance and authentication technologies, with emphasis on cryptographic manifests and threat models, watermarking and content provenance, including diffusion-era watermarking and industrial deployments, adversarial robustness and attacker adaptation, datasets and benchmarks, evaluation metrics across tasks, and deployment and scalability constraints. A dedicated section addresses legal, ethical, and policy issues, focusing on emerging transparency obligations and platform governance. The review finds that no single countermeasure is sufficient in realistic adversarial settings. The strongest practical approach is a layered defense that combines provenance, watermarking, content-based detection, and human oversight. The study concludes with limitations of the current evidence base and prioritized research directions to improve generalization, interoperability, and trustworthy user experiences.

Keywords: deepfakes; synthetic media; multimedia forensics; deepfake detection; provenance; authentication; content credentials; watermarking; adversarial machine learning; evaluation metrics

Academic Editors: Marcin Paprzycki, Robin Haunschild, Giorgio Maria Di Nunzio, Vasco N. G. J. Soares, Paulo Quaresma and Luigi Laura

Received: 24 February 2026
Revised: 25 March 2026
Accepted: 1 April 2026
Published: 3 April 2026

1. Introduction

The term deepfake has evolved from describing face-swapped videos to serving as an umbrella for AI-generated or AI-manipulated media, including images, video, audio voice cloning, and multimodal composites [1]. This definitional breadth creates concrete engineering consequences. A detector tuned to face-swapping artifacts may fail against diffusion-generated imagery. An audio anti-spoofing system optimized for telephony channel variability may not handle high-fidelity studio-grade voice clones. Media provenance mechanisms must support multiple file types, edits, and distribution transformations.

From a security and information integrity standpoint, deepfakes are best understood as an asymmetric capability. Attackers can cheaply generate high volumes of plausible false media, while defenders must either prove authenticity or detect manipulation under uncertainty, including adversarial post-processing [2]. This asymmetry is compounded by the open-world setting of online distribution, where content is frequently resized,

recompressed, transcoded, screenshotted, or otherwise transformed in ways that can erase fragile forensic signals and complicate model assumptions [3].

Contemporary research and practitioner consensus increasingly treats deepfake defense as an end-to-end socio-technical system problem, rather than a narrow classification problem. In this framing, content-based detection remains important but is complemented by authenticity infrastructure that enables cryptographically verifiable provenance, and by watermarking or fingerprinting approaches that support scalable labeling and later auditing [3].

This review is motivated by the widening gap between the speed and scale at which synthetic media can be produced and the fragmented, modality-specific, and often brittle countermeasures available in practice. The objective of this study is to synthesize the current state of the art in deepfake mitigation across passive detection, provenance and authentication, watermarking, benchmarks, robustness techniques, and governance frameworks, and to evaluate how these approaches complement one another under realistic adversarial conditions. The practical applications of this research include platform content moderation, newsroom and fact-checking workflows, forensic and evidentiary screening, election and public-information integrity systems, synthetic-media labeling, and policy design for transparency and accountability.

2. Background

Deepfake combat technologies emerge from the broader field of multimedia forensics, which historically sought to infer camera origin, detect splicing, identify manipulation traces, and assess integrity under common transformations (Figure 1). Survey work highlights that deepfakes accelerate traditional forensics vs. anti-forensics dynamic, shifting the problem from detecting relatively constrained edits to detecting high-capacity generative processes that can synthesize or rewrite content with semantically coherent detail [2].

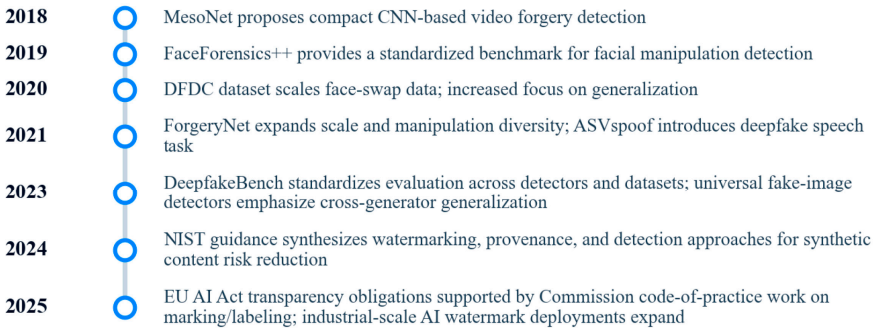

Figure 1. Milestones in deepfake countermeasures (selected).

Institutional programs including Defense Advanced Research Projects Agency [4] initiatives such as MediFor and SemaFor formalized parts of the problem as building automated integrity assessment for images and video and developing semantic forensics capabilities to analyze authenticity under advanced manipulation [1]. In parallel, National Institute of Standards and Technology has supported public evaluation efforts (e.g., OpenMFC) designed to reduce bias from self-evaluation and to structure recurring, task-based benchmarking for manipulation and deepfake detection [5].

A key conceptual shift is the recognition that authenticity is not purely a property inferable from pixels or waveforms. It is often an attribute of process and provenance. The Coalition for Content Provenance and Authenticity [6] technical specification, for example, defines a signed manifest structure composed of assertions, claims, and claim signatures bound to media through cryptographic hashes (hard bindings) or other identifiers (soft bindings, including fingerprints or invisible watermarks). This architecture reframes trust

as a validation problem anchored in signer identity, credential validity, and tamper-evident bindings rather than (only) statistical detection.

3. Methodology

This review followed a narrative literature review approach, adopting an evidence-oriented methodology, while acknowledging constraints inherent to rapidly evolving industrial deployments and mixed publication cultures in security and machine learning.

Literature and technical documentation were identified through structured searches of Google Scholar, arXiv, IEEE Xplore, ACM Digital Library, CVF Open Access, OpenReview, and official repositories and portals from standards bodies and public institutions (e.g., NIST data and publications, European Commission digital strategy pages, EUR-Lex). Searches were executed without a strict start-date restriction, with emphasis on 2018–2026 due to the emergence and scaling of modern deepfakes and countermeasures.

Representative terms included combinations of "deepfake detection", "synthetic media detection", "face forgery detection", "audio deepfake detection", "ASVspoof", "content provenance", "content credentials", "watermarking diffusion models", "robust watermark", and "adversarial attack deepfake detector".

Sources were included if they met at least one of the following: peer-reviewed publication or widely adopted benchmark paper with transparent experimental methodology, official technical standard/specification or security/threat-model documentation, official institutional report providing synthesis or guidance grounded in technical evidence, dataset paper/documentation enabling reproducible evaluation, or open-source tool documentation that implements or operationalizes standards or well-cited research methods. Sources were excluded when they were primarily speculative, marketing-only without technical detail, duplicated higher-quality primary references, or provided unverifiable performance claims. News coverage was used sparingly and only when it captured otherwise undocumented operational facts.

4. Deepfake Detection Methods

Deepfake detection methods are commonly grouped into passive (content-based) detection and active (metadata/provenance/watermark-assisted) detection, with hybrid systems increasingly common in deployment [3]. Official guidance emphasizes that provenance and watermarking can provide affirmative signals of origin, but their absence is not proof of manipulation. Therefore, content-based detection remains necessary in many settings (see Table 1 for an overview of major detection algorithms).

Early and widely deployed visual detectors rely on discriminative learning over face crops or full frames, using convolutional neural networks (CNNs) and training data from face-swap and reenactment pipelines. Compact architectures such as MesoNet were designed to capture mesoscopic properties and report strong detection rates on then-common deepfake styles, illustrating the viability of lightweight inference for video tampering detection under compression constraints [7]. As detectors matured, research increasingly separated approaches into naïve end-to-end classifiers versus methods that explicitly model forgery artifacts such as blending boundaries, warping, or inconsistent face boundaries [8].

Multiple lines of work exploit the observation that generative and manipulation pipelines can introduce atypical traces in the frequency domain, motivating detectors that incorporate discrete cosine transform (DCT) or other spectral representations. F3-Net (Frequency in Face Forgery Network) exemplifies this family by mining frequency-aware clues and is incorporated into standardized benchmarking as a representative frequency detector [9]. Frequency-based detectors can improve resilience to certain compression regimes but are also subject to adaptive attacks that explicitly target frequency cues [10]. In

adaptive settings, an attacker can deliberately manipulate high-frequency statistics through targeted perturbations in DCT or related spectral bands or apply hybrid spatial-frequency post-processing that suppresses the very artifacts detectors such as F3-Net are trained to exploit. These perturbations can remain visually subtle while still degrading detector performance, including in transfer settings against multiple detector families. This implies that frequency cues are useful but not stable invariants of manipulation. In practice, they should be treated as one signal among several and reinforced through augmentation, detector diversity, calibration, and continuous re-evaluation against adaptive attacks.

Video deepfakes may be detectable via temporal inconsistencies (e.g., motion, optical flow) or physiological cues not reliably synthesized. A prominent physiological approach is remote photoplethysmography (rPPG), which estimates subtle blood-flow signals from facial pixels. FakeCatcher operationalized this idea, reporting high accuracies across multiple face-forgery datasets and demonstrating a pathway toward real-time or near-real-time detection in constrained portrait video settings [11]. The strength of such approaches is interpretability aligned with human physiology. The limitation is that many real-world conditions (lighting, makeup, post-processing, low frame rate) can degrade rPPG estimation, and future generators may partially learn or simulate such signals [3].

Audio deepfakes, including text-to-speech (TTS) and voice conversion (VC), are closely related to the long-standing problem of spoofing in automatic speaker verification (ASV). The ASVspoof challenge series formalizes evaluation tasks for logical access (LA), physical access (PA), and a dedicated deepfake speech task (DF). The ASVspoof 2021 evaluation plan states that LA and PA use a revised tandem detection cost function (t-DCF), while the DF condition uses equal error rate (EER), reflecting differing operational contexts (ASV-constrained vs. general fake audio detection) [12]. Model architectures such as AASIST (Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks) aim to jointly exploit spectral and temporal artifacts via graph attention mechanisms, explicitly motivating single-system alternatives to ensembles [13].

Realistic impersonation often blends synchronized audio and video, motivating detectors that jointly assess cross-modal consistency (e.g., lip-sync vs. speech) and shared identity cues. FakeAVCeleb provides an audio-video dataset explicitly aimed at multimodal deepfake detection development and evaluation, arguing that unimodal datasets can be insufficient for realistic threat scenarios [14]. Multimodal benchmarking efforts are still less mature than visual-only benchmarks, but the trajectory is toward unified protocols and larger-scale multimodal corpora.

Table 1. Table comparing major detection algorithms with DeepfakeBench.

Name	Year	Modality	Approach	Performance Metrics [15]	Strengths	Limitations
Meso4 (MesoNet)	2018	Video (face frames)	Compact mesoscopic CNN classifier	Within Avg AUC 0.6097; Cross Avg AUC 0.6551	Lightweight; fast inference; early compact baseline	Weak vs. modern methods; limited generalization
MesoInception (MesoNet)	2018	Video (face frames)	Compact CNN with inception-style blocks	Within Avg AUC 0.7571; Cross Avg AUC 0.7364	Better than Meso4; still efficient	Cross-domain still modest; older artifact assumptions
Xception (baseline)	2019	Image/Video (face frames)	CNN classifier (Xception backbone)	Within Avg AUC 0.9450; Cross Avg AUC 0.7718	Strong, widely used baseline	Cross-domain drop vs. within-domain
EfficientNet-B4	2019	Image/Video (face frames)	CNN classifier (EfficientNet backbone)	Within Avg AUC 0.9389; Cross Avg AUC 0.7718	Parameter-efficient; strong within-domain	Cross-domain similar to Xception
Capsule (Capsule-Forensics)	2019	Image/Video (face frames)	Capsule-network based forensics representation	Within Avg AUC 0.8217; Cross Avg AUC 0.7488	Alternative inductive bias vs. plain CNN	Cross-domain degradation; heavier than simple CNNs
FWA (also known as. DSP-FWA/Face Warping Artifacts)	2019	Video (face swap)	Detects warping/blending artifacts from face warping pipeline	Within Avg AUC 0.8549; Cross Avg AUC 0.7239	Interpretable artifact cue; strong on some swap pipelines	Brittle to new generation pipelines; limited cross-domain

Table 1. *Cont.*

Name	Year	Modality	Approach	Performance Metrics [15]	Strengths	Limitations
CNN-Aug (CNNDetection-style augmentation)	2020	Image/Video (face frames)	CNN classifier + augmentation strategy	Within Avg AUC 0.8419; Cross Avg AUC 0.7020	Simple; shows importance of training tricks	Sensitive to preprocessing; cross-domain drop
Face X-ray	2020	Image/Video (face frames)	Boundary/blending trace modeling with X-ray style maps	Within Avg AUC 0.9391; Cross Avg AUC 0.6985	Interpretable maps; targets compositing signatures	Cross-domain weak in DBench; implementation sensitivity
FFD (On the Detection of Digital Face Manipulation)	2020	Video (face frames)	CNN-based detection for manipulated faces	Within Avg AUC 0.9434; Cross Avg AUC 0.7733	Competitive cross-domain among visual methods	Still notable generalization gap
F3-Net	2020	Video (face frames)	Frequency-aware detection (DCT/spectral cues)	Within Avg AUC 0.9449; Cross Avg AUC 0.7645	Strong within-domain; leverages frequency cues	Frequency cues can be attenuated by adaptive post-processing
SPSL	2021	Video (face frames)	Frequency/phase spectrum learning	Within Avg AUC 0.9408; Cross Avg AUC 0.7875	Best cross Avg AUC among listed frequency baselines in DBench	Potentially vulnerable to frequency-domain countermeasures
SRM	2021	Video (face frames)	Noise/residual modeling (SRM-style high-frequency cues)	Within Avg AUC 0.9359; Cross Avg AUC 0.7760	Competitive cross-domain; emphasizes residual cues	Depends on augmentation/codec; still generalization gap
CORE	2022	Video (face frames)	Erasing/reconstruction cues + spatial modeling	Within Avg AUC 0.9431; Cross Avg AUC 0.7694	More robust to some manipulations	Cross-domain still below within-domain
RECCE	2022	Video (face frames)	Reconstruction-based cues (spatial)	Within Avg AUC 0.9422; Cross Avg AUC 0.7649	Moves beyond pure classification to reconstruction signals	Needs careful training; cross-domain limited

Beyond temporal synchronization, recent multimodal deepfake detection research has begun to exploit richer behavioral and biometric correspondences between face, voice, and motion [16–20]. Watch Those Words models word-conditioned facial Action Units and head motion, showing that falsification can be exposed by inconsistencies between spoken content and person-specific facial dynamics. POI-Forensics reframes detection as audio-visual identity verification, asking whether the observed face and voice belong to the same subject. AVFF and related attention-based methods learn cross-modal correspondences more generally, while ART-AVDF introduces articulatory representation learning to test whether lip motion and speech remain physiologically compatible. Fine-grained multimodal graph methods further show that the task can extend beyond binary real/fake decisions to identify whether the forgery resides in the audio stream, the video stream, or both.

A directly gait-centric multimodal deepfake-detection literature remains limited [20–22]. The stronger recent line of work focuses on facial motion, head movement, articulatory dynamics, gestural mannerisms, and voice as identity-bearing behavioral signals rather than relying only on pixel artifacts. Recent work on facial biometric anomalies likewise suggests that modern talking-head generators still struggle with complex expressions and motion patterns.

A further promising direction is detection grounded in real-world physical laws [23–25]. Unlike artifact-specific detectors, physics-based detectors would test whether trajectories, accelerations, collisions, gravity-driven motion, object persistence, or human kinematics remain compatible with first-principles constraints. In intuitive terms, a moving car should obey Newtonian motion, a bouncing object should display plausible collision dynamics, and speech-driven facial motion should remain biomechanically coherent. Recent evaluation work on video generation models shows persistent failures on Newtonian and conservation laws, suggesting that physical-law violations may offer a more generator-agnostic cue than many pixel-space artifacts. Although this literature is presently framed more as evaluation of generative models than as turnkey forensic detection, it strongly motivates future deepfake detectors that combine multimedia forensics with physics- and kinematics-based consistency checks.

A recurring result across surveys, benchmarks, and institutional guidance is that detector performance often drops substantially in cross-dataset or in-the-wild conditions, particularly under dataset shift, novel generators, or post-processing [8]. DeepfakeBench was explicitly proposed to address inconsistent preprocessing and evaluation settings in the literature and provides standardized within-domain and cross-domain ROC-AUC comparisons across detectors trained on FaceForensics++ c23.

Even within a unified benchmark, these results do not fully resolve real-world deployment uncertainty, since operational conditions can involve unseen generators, multi-step editing pipelines, streaming artifacts, and deliberate adversarial adaptation. Nonetheless, standardized reporting exposes the magnitude of the generalization gap that dominates practical risk [8].

In addition, reported benchmark accuracy alone is insufficient for deployment decisions because computational complexity, memory footprint, and inference latency strongly affect operational feasibility. Lightweight CNN-based detectors such as MesoNet remain attractive for edge devices and first-pass upload screening because they are comparatively efficient, but they trade off accuracy and generalization. Mid-range CNN backbones such as Xception or EfficientNet often provide a more practical balance between performance and cost for batch moderation and offline triage. By contrast, capsule-based, transformer-based, multimodal, and ensemble systems can improve accuracy or robustness but generally impose higher latency and hardware demands, making them less suitable as always-on front-line filters for live-stream monitoring without cascading or selective triggering. Frequency-aware methods may also incur extra preprocessing overhead because spectral transforms must be computed in addition to feature extraction. For real-world applications, studies should therefore report not only AUC, EER, or accuracy, but also throughput, memory use, model size, and hardware assumptions.

Table 2 summarizes more recent multimodal detectors evaluated on heterogeneous datasets and protocols, their deployment relevance, and limitations.

Table 2. Selected multimodal deepfake detection methods from 2023 to 2025.

Name	Year	Modality	Approach	Strengths	Limitations
Self-Supervised Video Forensics by Audio-Visual Anomaly Detection [26]	2023	Audio-video	Real-only anomaly detection over audio-visual synchronization patterns	Avoids dependence on fake training data; good conceptual robustness	Strongest on speaking-video scenarios; still centered on temporal AV consistency rather than broader behavioral cues
Watch Those Words: Video Falsification Detection Using Word-Conditioned Facial Motion [16]	2023	Audio + video	Word-conditioned facial motion using facial Action Units and head movement	Interpretable, goes beyond lip-sync by modeling facial-expression and speech-content compatibility	Requires visible face and speech, person-specific dynamics may limit broad deployment
Audio-Visual Person-of-Interest DeepFake Detection (POI-Forensics) [17]	2023	Audio + video	Contrastive audio-visual identity verification	Useful for high-value-target protection and identity-centric verification	Assumes reference data for the protected identity, less universal for unknown subjects.
AVFF: Audio-Visual Feature Fusion for Video Deepfake Detection [18]	2024	Audio + video	Two-stage cross-modal learning of audio-visual correspondence	Strong recent multimodal baseline, explicitly designed for improved cross-modal correspondence	More computationally demanding than lightweight visual-only detectors
Fine-Grained Multimodal DeepFake Classification via Heterogeneous Graphs [19]	2024	Audio + video	Heterogeneous graph attention for intra-/inter-modal relationships	Distinguishes whether forgery is in audio, video, or both, useful for triage	More complex pipeline, evaluated on multimodal benchmarks rather than unified visual benchmarks

Table 2. *Cont.*

Name	Year	Modality	Approach	Strengths	Limitations
Audio–visual deepfake detection using articulatory representation learning (ART-AVDF) [20]	2024	Audio + video	Articulatory/physiological correspondence between speech and lip motion	Physically grounded cue family, useful addition to physiological discussion	Mainly suited to speaking-face content rather than arbitrary scenes
Detecting Deepfake Talking Heads from Facial Biometric Anomalies [21]	2025	Audio + video	Facial biometric anomalies under complex motion/expression	Very relevant to modern avatar/talking-head threats	Scope is talking-head impersonation rather than all deepfake

5. Provenance and Authentication

Content provenance systems aim to provide verifiable information about how an asset was captured, generated, edited, and distributed. The C2PA specification formalizes this via a signed manifest model. Assertions and claims are bound into a C2PA Manifest, with a claim signature, and collected into a manifest store that may be embedded or external. Trust decisions are anchored in signature validation, signer identity, and, when present, trusted timestamps and revocation checking [6].

In practical workflows, provenance is intended to travel through a content lifecycle: capture → edit → publish → consume/verify. The Content Authenticity Initiative [27] positions Content Credentials as a cross-industry approach to making this lifecycle visible to users and interoperable across tools and platforms, promoting a digital nutrition label style interface for provenance. Implementations emphasize creator recognition and transparency about capture, AI generation, and editing steps [3,28].

However, provenance systems are not a silver bullet. An analogous adaptive threat exists for provenance systems in the form of manifest stripping and provenance transference. An attacker may remove an embedded manifest from an otherwise credentialed asset and repost the file without credentials or attempt to copy valid provenance data between unrelated assets. The C2PA Security Considerations document explicitly models attacker goals. It also emphasizes that security analysis is an evolving process, and that context matters for risk and likelihood [6]. Complementing security threats, C2PA harms modeling highlights privacy risks, the possibility of overreliance, and the need to avoid treating the presence of manifests as a definitive indicator of truth, since content can be signed yet misleading.

Operationally, provenance adoption depends on tooling. The CAI open-source ecosystem includes libraries and tools such as c2pa-rs and c2patool, enabling creation, signing, embedding, and validation of C2PA manifests for supported media formats [29]. Such tooling is essential for integrating provenance into cameras, editing suites, content management systems, and verification services at scale in a way that is auditable and interoperable.

6. Watermarking and Content Provenance

Watermarking refers to embedding a signal into content such that it can later be detected (zero-bit watermarking) or decoded (multi-bit watermarking). For deepfake mitigation, watermarking is used for origin transparency (flagging AI-generated content), traceability and auditing, and, in some designs, linking to provenance records via soft bindings [6].

Google DeepMind [30] describes SynthID as a watermarking technology that embeds watermarks into AI-generated content (including images, audio, text, and video) and supports detection via dedicated tools. A technical report on SynthID-Image emphasizes deployment constraints at internet scale, including robustness, fidelity, and security considerations, and reports watermarking of over ten billion images and video frames across

services, illustrating the scale at which provenance-adjacent signals may be operationalized [31].

Tree-Ring Watermarking proposes embedding a watermark in the initial noise vector of diffusion sampling, structured in Fourier space for robustness to a range of transformations, with detection performed by inverting the diffusion process to recover the noise signal [32]. Follow-on work such as RingID revisits the mechanism and analyzes robustness sources, while security research also investigates watermark removal and attack surfaces, underscoring that watermarking is an arms race rather than a one-time fix [33].

NIST guidance stresses that watermarking schemes can be attacked via removal, paraphrasing (for text), or targeted distortions. It further discusses spoofing attacks that can cause detectors to misclassify non-watermarked content as watermarked, potentially undermining trust [3]. Importantly, NIST highlights scale considerations. If each model developer deploys a private, method-specific watermark requiring a unique detector, users face a fragmented ecosystem and high burden. Public watermark schemes may scale more easily but can enable spoofing and evasion via repeated probing. These observations motivate research into interoperable watermark standards, multi-detector services, and governance regimes that reduce incentives for fragmentation.

7. Adversarial and Robustness Techniques

Deepfake defenses operate under active adversaries. Two adversarial classes are particularly important.

Attackers can exploit detector brittleness using compression, resizing, filtering, re-encoding, or deliberate perturbations, often without materially reducing perceptual realism. Work on frequency adversarial attacks explicitly targets frequency-domain cues used by face forgery detectors, showing that attacker adaptation can be domain-specific and that detectors relying on a narrow cue set may be systematically evaded [10]. Standardized evaluation reinforces that cross-manipulation generalization can be poor even without explicit adversarial optimization. When only the manipulation algorithm changes within the same dataset, AUC degradation can be substantial for many detectors [8].

Provenance systems can be attacked by stripping metadata, transferring metadata to unrelated assets, or manipulating user perception of trust signals. C2PA's threat model explicitly includes stripping C2PA manifests and copying metadata across assets, and it discusses mitigations such as manifest repositories while noting that such mitigations introduce additional security and privacy tradeoffs [6]. Watermarking can be attacked by removal or spoofing. NIST reports empirical evidence that spoofing watermark presence has been demonstrated for various watermark types and warns that spoofing undermines trust even if the goal is only synthetic vs. not [3].

Across the literature, robustness is pursued via strong and diverse data augmentation aligned with expected distortions, domain generalization and cross-dataset training, self-supervised or foundation-model feature reuse to reduce overfitting to generator-specific artifacts, detector ensembles with calibration and uncertainty estimation for triage, and continuous monitoring and re-training pipelines that treat detector decay as inevitable under target drift. DeepfakeBench directly analyzes how factors such as augmentation, backbone selection, and pretraining influence performance, illustrating that some simple baselines become competitive when trained under consistent, well-designed protocols, yet still face cross-domain limits [8].

8. Dataset and Benchmark Landscape

Progress in deepfake defense is tightly coupled with benchmark quality. Datasets vary across manipulation type (swap, reenactment, synthesis), data source (studio vs. internet), consent and ethics, compression regimes, and annotation richness.

FaceForensics++ introduced a widely used benchmark for facial manipulation detection, providing a standardized evaluation setting across multiple manipulation methods and compression levels [34]. The DFDC dataset scaled data collection to over 100,000 clips from paid actors with explicit consent, offering a large training and evaluation substrate and reporting generalization considerations from a major public challenge [35]. Celeb-DF aimed to increase realism and difficulty by producing higher-quality deepfakes aligned with internet-circulated content characteristics [36]. DeeperForensics-1.0 emphasized scale and real-world perturbations, supporting robustness evaluation via large video volume and diverse degradations [37]. ForgeryNet pushed scale further, reporting millions of images and hundreds of thousands of videos across multiple tasks and manipulation types [38].

WildDeepfake collected deepfake videos from the internet, highlighting that detectors trained on controlled datasets may fail under real-world distributions and that performance can drop drastically [39].

Diffusion-era image generation motivated benchmarks such as GenImage, which provides a million-scale dataset and proposes cross-generator and degraded image classification tasks to test real-world applicability under generator diversity and distortion [40]. Complementary research argues for universal fake-image detectors that generalize across generative model families using feature spaces not explicitly trained for real-vs-fake discrimination [41].

ASVspoof datasets provide structured evaluations for spoofing and deepfake speech across channel variability and compression, with explicit metrics and rules that enable comparable progress tracking [12]. FakeAVCeleb targets a multimodal threat model by pairing deepfake video with synthesized and lip-synced fake audio, enabling joint audio-video detector development [14].

Recent benchmark development has shifted toward larger, more heterogeneous, and more realistic multimodal settings. AV-Deepfake1M and its follow-on challenge infrastructure target large-scale audio-visual detection and localization [42], DeepSpeak emphasizes consented and identity-matched webcam deepfakes generated with modern voice and video engines [43], and Deepfake-Eval-2024 highlights how sharply detector performance can degrade on current in-the-wild content [44]. These datasets indicate that multimodal evaluation is moving beyond curated celebrity face-swaps toward more operationally realistic speaking-video and cross-platform scenarios (see Table 3).

Table 3. Widely used datasets/benchmarks supporting deepfake defense research and evaluation.

Name	Year	Modality	Approach/Scope	Key Scale or Official Metric	Strengths	Limitations/Notes
FaceForensics++	2019	Video (faces; released largely as extracted face frames)	Benchmark for facial manipulation detection using 4 methods; evaluated under multiple compression settings (commonly raw/c23/c40) with a hidden test set	Over 1.8 M manipulated images (frames) across methods; hidden test set	Standardized, widely used baseline; controlled compression enables apples-to-apples comparisons	Methods reflect the era's generators; still a curated pipeline and limited to specific facial manipulation families
DFDC Dataset	2020	Video (face swap)	Large-scale dataset + challenge setup; multiple method families; explicit subject consent	Over 100,000 clips; 3426 paid actors	Scale + strong consent/ethics framing; includes analysis of competition submissions	Still subject to distribution shift; not exhaustive across all manipulation types
Celeb-DF (v2)	2020	Video (faces)	Higher-fidelity celebrity deepfakes intended to resemble real online content	590 real + 5639 fake videos	Higher visual realism than earlier academic sets; commonly used to test generalization	Celebrity-domain bias; still narrower manipulation diversity than internet scale

Table 3. *Cont.*

Name	Year	Modality	Approach/Scope	Key Scale or Official Metric	Strengths	Limitations/Notes
DeeperForensics-1.0	2020	Video (faces)	Large-scale face forgery benchmark with explicit real-world perturbations and hidden test set	60,000 videos/17.6 M frames (50,000 real + 10,000 manipulated)	Perturbation design stresses robustness; strong scale; hidden test set	Generation pipeline can imprint systematic artifacts; dataset is heavy to store/process
ForgeryNet	2021	Image + Video (faces)	Mega-scale deep face forgery benchmark spanning multiple manipulation types and perturbations	2.9 M images; 221,247 videos	Scale + breadth across tasks/manipulations; rich labels for multiple analysis tasks	Large compute/storage footprint; still curated around benchmark definitions
WildDeepfake	2021	Video (in-the-wild faces)	Internet-collected deepfakes aimed at realistic testing under distribution shift	707 deepfake videos; 7314 face sequences	Captures real-world shift; highlights brittleness of lab-trained detectors	Smaller; collection/label representativeness is inherently challenging
GenImage	2023	Image (general)	Million-scale AI-generated image detection benchmark; tests cross-generator and degraded robustness	Over 1 M pairs of generated/real images; cross-generator + degraded classification	Targets modern generator diversity (diffusion + GAN) and degradations	Not video/audio; abstraction may miss temporal/physiological deepfake cues
ASVspoof 2021 (LA/PA/DF)	2021	Audio	Challenge datasets for spoofed/deepfake speech across LA, PA, DF tracks	Official metrics: min t-DCF (LA/PA) and EER (DF)	Standardized competitive evaluation; operationally motivated metrics	Audio-only; generators evolve quickly → periodic refresh needed
FakeAVCeleb	2021	Audio + Video	Multimodal deepfake dataset with lip-synced fake audio; multiple A/V authenticity combinations	20,000 total samples: ARVR 500, AFVR 500, ARVF 9000, AFVF 10,000	Enables cross-modal consistency checks; addresses unimodal-only gap	Celebrity/source-domain bias; multimodal evaluation protocols still less standardized than unimodal
OpenMFC	2020	Image + Video	Public, ongoing NIST evaluation of systems that detect (and often localize) manipulations; initial focus includes manipulation + deepfake tasks	Common reporting includes ROC/AUC and CD@FAR = 0.05; evolving tasks (e.g., MD/DD/StegD)	Recurring evaluations + infrastructure; neutral scoring/leaderboard tooling; supports longitudinal comparison (with caveats)	Coverage depends on what datasets/tasks are released each cycle; can lag newest consumer generators
AV-Deepfake1M	2024	Audio + video	Large-scale content-driven deepfake detection/localization dataset	>2000 subjects, >1 million videos	Large scale; supports localization as well as detection	Challenge-style synthetic pipeline; still not fully in-the-wild
DeepSpeak	2204	Audio + video	Consented webcam-style audiovisual deepfakes with identity matching	>100 h total; >50 h real from 500 consenting participants; >50 h fake; 14 video synthesis engines and 3 voice cloning engines	Modern, consented, identity-matched, multimodal	Webcam domain may not capture all platform conditions
Deepfake-Eval-2024	2025	Image + audio + video	In-the-wild benchmark of deepfakes circulated in 2024	45 h of video, 56.5 h of audio, 1975 images; 88 websites; 52 languages	Highly realistic and current; exposes real-world performance collapse	Emerging benchmark, less historically standardized than FaceForensics++/DFDC

9. Evaluation Metrics

Evaluation choices encode implicit threat models. A common failure mode in the deepfake literature is reporting accuracy on a single dataset split without stress-testing compression, distribution shift, adversarial post-processing, and false-alarm constraints relevant to real deployments. DeepfakeBench explicitly notes that inconsistent pipelines and heterogeneous metric usage (frame-level vs. video-level) hinder comparability, motivating unified reporting across multiple metrics (AUC, average precision (AP), accuracy (ACC), and EER) [8].

Public evaluation programs emphasize operational metrics. OpenMFC describes detection tasks scored via receiver operating characteristic (ROC), area under the curve (AUC), and correct detection (CD) at a specified false alarm rate (FAR), such as CD@FAR = 0.05, aligning evaluation with thresholded decision-making [5]. Audio deepfake evaluation likewise uses domain-specific measures. ASVspoof 2021 specifies min t-DCF for ASV-constrained tasks (LA/PA) and EER for the deepfake speech task (DF), making explicit that different operational settings require different objective functions [12].

Since deepfake defense frequently serves as a triage mechanism feeding human review, calibration (mapping scores to interpretable probabilities), cost-sensitive evaluation, and uncertainty estimation become essential, especially when false positives can cause reputational harm or censorship. NIST guidance explicitly frames detection and labeling as part of broader risk-reduction systems, not standalone technical wins [3].

10. Deployment and Scalability

Deploying deepfake defenses requires balancing latency, cost, privacy, and user experience. Figure 2 summarizes the proposed multi-layered operational workflow for deepfake detection. The process begins at media intake, where an incoming asset is first checked for signed provenance. This branch is prioritized because valid provenance can provide affirmative, cryptographically verifiable information about the origin, capture, generation, and editing history of the media. If signed provenance is present, the system validates signatures, certificate status, and the bindings between the manifest and the asset. When that validation is trusted, the result can be surfaced in the user interface through provenance indicators or Content Credentials and then subjected to applicable platform or organizational policy checks. If provenance is present but the validation is not trusted, the asset is flagged as presenting a possible tampering or trust-chain risk and is not accepted as authentic by default.

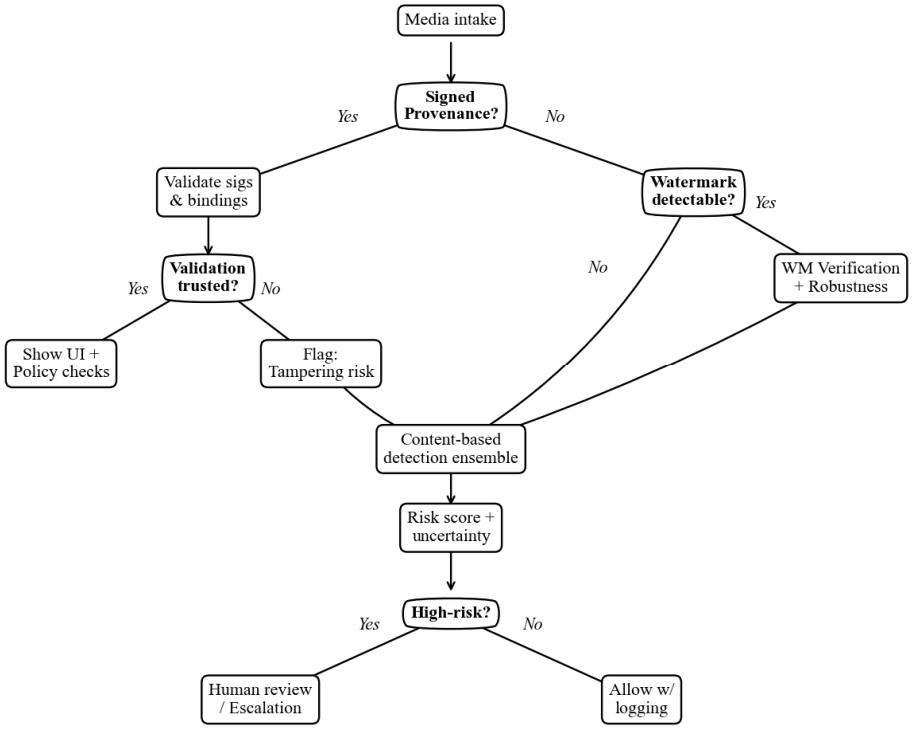

Figure 2. Multi-layered workflow for deepfake detection.

If no signed provenance is available, the workflow next checks whether a watermark is detectable. A detectable watermark does not by itself settle authenticity. Instead, it triggers a watermark verification and robustness stage. This stage tests whether the watermark is genuine, whether it survives common transformations such as resizing, compression, or re-encoding, and whether the signal may have been spoofed. If no watermark is detected, or if provenance is absent or untrusted, the asset proceeds to the content-based detection ensemble. This ensemble represents the combination of complementary detectors discussed throughout the review, including spatial and frequency-domain image or video detectors,

temporal and physiological video methods, audio anti-spoofing systems, and multimodal consistency checks.

The ensemble then produces a risk score together with an uncertainty estimate. This is important because operational decisions should not rely only on raw classification scores but also on confidence and calibration. In the final stage, high-risk or high-uncertainty cases are routed to human review or escalation, while lower-risk cases are allowed subject to logging and later audit. The overall logic of the figure is that provenance provides the strongest positive evidence when available, watermarking adds scalable transparency signals, and content-based detection remains the fallback for suspicious or uncredentialed media.

SynthID-Image's reported scale illustrates that watermarking can be deployed as part of platform infrastructure, but NIST cautions that scaling is harder when detector access is fragmented across private schemes and that public schemes introduce spoofing and probing risks [31].

OpenMFC explicitly argues that neutral benchmark evaluations are needed because system performance is sensitive to data/task variation and self-evaluation may not reflect field performance. It further frames recurring evaluation as necessary due to evolving manipulation technologies [5].

Content provenance requires developer-grade libraries and command-line tools to embed and validate manifests. C2patool and c2pa-rs illustrate how standards become operationalized into software components for desktop, mobile, and embedded integration [45]. Table 4 reviews tools and standards implementing or operationalizing deepfake countermeasures.

Table 4. Tools/standards implementing or operationalizing deepfake countermeasures.

Name	Modality	Approach	Key Properties/Performance	Strengths	Limitations
Coalition for Content Provenance and Authenticity (C2PA) Technical Specification (Content Credentials)	Cross-modal (incl. image/video/audio/PDF; also unstructured text embedding is specified)	Cryptographically signed manifests/claims + hard/soft bindings	Defines validation states incl. Well-Formed/Valid/Trusted; supports embedded + decoupled approaches (via bindings)	Interoperable standard; explicit trust/validation concepts; extensible provenance model	Metadata/manifest can be stripped; meaningful value depends on ecosystem adoption + UX around trust and missing credentials
C2PA Security Considerations	Cross-modal	Threat modeling + mitigations guidance (non-normative)	Explicit threat scenarios include stripping manifests and copying C2PA metadata; mitigation discussion includes manifest repositories and validation behavior	Makes attacker goals explicit; helps implementers reason about secure deployment	Non-normative; likelihood/impact are context-dependent; guidance evolves with versions
C2PA Harms Modeling	Cross-modal	Socio-technical/harms framing for provenance ecosystems	Focuses on privacy, governance, misuse/over-interpretation risks (provenance $\neq$ truth)	Explicit risk framing for stakeholders	Not a detector; mitigations depend on implementers and deployment choices
Content Authenticity Initiative (CAI)	Cross-modal	Ecosystem adoption + open tooling to support Content Credentials	Promotes capture/edit/publish/consume style workflows + consistent provenance UX (implementation guidance + tools)	Cross-industry participation; bridges standards $\leftrightarrow$ product adoption	Adoption uneven; preservation depends on platforms/pipelines not stripping data
ContentCredentials.org (Content Credentials adoption + verification hub)	Cross-modal (site explicitly references photos/videos/audio/documents)	Education + pin UX + inspection/verification entry points	Positions a recognizable Content Credentials pin and links to verification tooling; states spec is hosted by C2PA	Clear UX concept for consumers; helps standard visibility	Still depends on preservation of credentials + compatible viewers
Adobe Content Credentials tooling (apps + web/extension)	Image/video/audio (tooling scope varies by product)	Productized creation/viewing of Content Credentials in workflows	Described as a digital nutrition label metadata type; Adobe applies credentials automatically for Firefly outputs; also provides viewing tools (e.g., extension/inspect)	Strong workflow integration for creators; improves attribution and transparency in supported apps	Platforms/pipelines may strip metadata; durability claims can depend on proprietary/extra mechanisms and interoperability

Table 4. *Cont.*

Name	Modality	Approach	Key Properties/Performance	Strengths	Limitations
c2patool	Audio/Image/Video	CLI to inspect and add C2PA manifests	Reads summary JSON + low-level reports; can add a manifest to supported files	Practical integration/testing; supports developer workflows	Requires key/cert handling for signing; limited by supported formats and pipeline preservation
c2pa-rs/c2pa-python	Audio/Image/Video (supported formats)	SDKs to create/sign/validate manifests	Rust library implements a subset of spec; supports create/sign/parse/ validate + embedding in supported formats; Python lib can read/validate and create/sign/attach	Reusable components for embed/mobile/ server pipelines	Integrators must manage key security, dependency hygiene, and correct validation configuration (trust lists, revocation, etc.)
Google DeepMind SynthID	Image/Audio/Video/Text (family of tools; text has separate constraints)	Invisible watermarking + detection	DeepMind describes embedding imperceptible watermarks in AI content; Google also documents limitations for text watermarking and conditions that reduce detectability	Scalable transparency for supported generators; simple detect or not story for downstream	Generator/tooling specific; watermarking is not universal and can be degraded by transformations (esp. text), so fragmentation remains a risk
Tree-Ring Watermarks	Image (diffusion)	Watermark via diffusion noise initialization; detect via inversion	Reported robustness to common transforms in experiments, with minimal quality impact	Model-integrated watermark design; strong experimental intuition	Detection can be computationally heavier; literature includes ongoing attack/removal work
OpenMFC + MediScore	Image/Video (OpenMFC scope)	Public evaluation + scoring software	NIST documents detection metrics including ROC/AUC and CD @ FAR = 0.05; MediScore provides validator/scorer components	Neutral benchmarking + recurring evaluation; unified scoring tooling helps comparability	Dataset/task coverage bounded; not a turnkey detector; cross-year comparisons can be non-apples-to-apples as tasks evolve
Microsoft Video Authenticator	Image/Video	Frame-level deepfake detection scoring	Outputs confidence score; Microsoft states it was created using FaceForensics++ and tested on DFDC	Fits disinformation response workflows	Limited public technical transparency; external auditing of drift/performance is hard without broader release + benchmarks

11. Legal, Ethical, and Policy Considerations

Technical countermeasures operate within governance frameworks that determine incentives for adoption, disclosure obligations, and institutional responsibilities for labeling and enforcement.

Regulation (EU) 2024/1689 establishes harmonized rules on AI and includes transparency obligations for AI-generated or manipulated content. The European Commission frames Article 50 obligations as aiming to ensure transparency of AI-generated or manipulated content such as deepfakes and is developing a Code of Practice to support compliance with marking and labeling requirements, including machine-readable marking and professional labeling of deepfakes in contexts of public interest [46]. This policy direction is aligned with technical guidance emphasizing interoperability and user comprehension. Labeling that is stripped or inconsistently displayed across platforms fails to meet the intended transparency goals [3].

The Digital Services Act (DSA) and related instruments treat disinformation and manipulation risks through platform accountability mechanisms and codes of practice. The Commission's materials emphasize transparency structures (e.g., DSA Transparency Database processes) and risk-mitigation expectations, while the Code of Practice on Disinformation explicitly references malicious deepfakes as a manipulative behavior that signatories should address [47].

Provenance and watermarking can introduce privacy risks if metadata leaks sensitive information or if systems enable tracking or coercive identification. NIST guidance discusses privacy considerations for watermarking, particularly for non-zero-bit watermarks that can carry data, and notes that covert watermarks are persistent and may complicate

privacy controls [3]. C2PA harms modeling similarly emphasizes risks to vulnerable groups, potential misuse in adverse legal/political contexts, and the need to avoid overreliance, explicitly stating that the presence of valid manifests does not equate to truthfulness of the content [6].

In legal and journalistic contexts, provenance can support chain-of-custody arguments by providing cryptographic assurances that recorded claims and bindings were not tampered with, but standards documents also warn that attackers can remove manifests or disrupt availability. Post hoc interpretation remains difficult when provenance is absent [6]. These tensions imply that policy must address not only labeling requirements but also preservation incentives for platforms and tooling ecosystems that keep provenance intact during distribution transformations.

12. Discussion

This review yields four main findings. First, content-based deepfake detectors can achieve strong in-domain performance, but their reliability drops under cross-dataset shift, unseen generators, compression, and adversarial post-processing. Second, provenance and authentication systems provide the strongest affirmative evidence of origin when signatures, bindings, and trust chains validate correctly, but their effectiveness depends on ecosystem adoption and the preservation of metadata across platforms and edits. Third, watermarking can support scalable transparency and auditing for AI-generated content, yet it remains vulnerable to removal, spoofing, and fragmentation across proprietary schemes. Fourth, the most robust practical strategy is not a single detector but a layered architecture that combines provenance verification, watermark checks, content-based analysis, uncertainty estimation, logging, and human escalation.

The core technical insight emerging across benchmarks, standards, and guidance is that deepfake defense is not a single-model winner-take-all problem. Content-based detectors can be highly accurate in-domain yet fragile under domain shift, and robustness improvements frequently come from engineering discipline (unified preprocessing, augmentation, and evaluation) rather than from a single novel architecture [8]. In-the-wild datasets and public evaluation programs reinforce that realistic deployment conditions can differ substantially from research datasets, necessitating recurring evaluations and continuous updates [39].

Recent review evidence further supports the role of artificial intelligence as a central countermeasure against deepfakes, while also showing that headline accuracy values should be interpreted cautiously. Representative deepfake-detection models with training, validation, and testing accuracies range from 94.2/91.8/90.5% for an RNN-based model to 99.1/97.5/97.2% for an ensemble model [48]. Additional examples include 98.5/96.2/95.8% for a CNN, 96.8/95.3/94.7% for a GAN-based model, 92.6% test accuracy for a real-time EfficientNet configuration, and 97.3/96.1/95.5% for a transfer-learning ResNet50 approach. These results confirm that AI-based detectors can achieve strong performance under controlled conditions, but the gap between training and validation/testing scores also suggests residual overfitting and limited direct comparability across heterogeneous datasets and protocols.

Accuracy must also be considered together with deployment speed. Under estimated standard-GPU conditions at 224 $\times$ 224 resolution, CNN-based detectors achieve about 88–90% accuracy at roughly 28–30 fps, whereas transformer-based, multimodal, and hybrid detectors reach approximately 92–95% accuracy but often operate at only 18–22 fps [49]. Since real-time video analysis typically requires around 24 fps, the most accurate models may not always be the most suitable for live or platform-scale screening without cascading, compression, cloud offloading, or region-of-interest processing. Models performing above

90% on constrained benchmarks can fall to around 60% on more challenging in-the-wild datasets, reinforcing that discussion of deepfake detection should extend beyond accuracy alone to include cross-dataset robustness and inference efficiency.

Provenance and content credentials offer a qualitatively different security value proposition. Rather than guessing authenticity from content traces, they enable cryptographic verification of recorded claims and edits when adopted end-to-end [6]. Yet provenance systems face adoption, usability, and attack-surface constraints, especially manifest stripping and inconsistent platform preservation, that can force defenders back to content-based detection in precisely the settings where detection is hardest.

Watermarking occupies a middle ground, capable of operating at large scale (as exemplified by industrial deployments) but subject to removal, spoofing, and fragmentation across providers. NIST's scale analysis and spoofing discussion underscore that watermarking requires ecosystem-level coordination and careful threat modeling to avoid undermining trust through inconsistent or forgeable signals [3].

The evidence supports a layered conclusion. The most realistic near-term path to robust deepfake mitigation is compositional combining provenance standards, watermarking, and adaptive detection, tied together by governance frameworks that require disclosure and incentivize metadata preservation.

This review faces several limitations. First, parts of the most operationally important ecosystem like platform enforcement, large-scale detector performance, and real-time deployment constraints remain under-documented in peer-reviewed venues, creating unavoidable reliance on official documentation and selected technical reports rather than fully reproducible performance studies. Second, performance values across papers are often not directly comparable due to differences in preprocessing, face detection, sampling strategies, and frame-level versus video-level aggregation. The review therefore prioritizes unified benchmarks (e.g., DeepfakeBench) where possible but cannot eliminate comparability issues entirely. Third, policy interpretation is inherently time sensitive. While this review cites primary EU legal texts and Commission materials, ongoing guidance development and national implementations can evolve beyond the cited snapshots.

13. Future Research Directions

Future research should prioritize closing persistent gaps in current evidence by developing detection methods that generalize beyond generator-specific artifacts, leverage domain-general representations, and quantify uncertainty robustly, including through abstention under distributional shift. Since real-world attacks are inherently multimodal, progress also depends on larger and more diverse benchmarks, building on early efforts such as FakeAVCeleb, together with standardized evaluation protocols that reflect operational conditions. These protocols should go beyond headline accuracy to report robustness against adaptive attacks, including frequency-targeted evasion, watermark spoofing, provenance stripping, and adversarial removal, while also capturing latency, calibration, and false-alarm-constrained performance. In this context, evaluation metrics must align with application risk through scenario-specific reporting, such as FAR-constrained measures in initiatives like OpenMFC or the use of t-DCF alongside EER in ASVspoof. Research should further advance interoperable provenance and watermark recovery mechanisms that survive platform transformations, reduce user burden, and limit ecosystem fragmentation, while enabling efficient detection at scale. As threat models increasingly assume the stripping of forensic traces, practical mitigation will also depend on platform behavior and user interfaces that communicate trust signals clearly and without overclaiming. Addressing these challenges will require interdisciplinary collaboration across machine learning, cryptography, usable security, and platform governance. Another promising research direction

is physics-aware forensic detection, in which models test whether synthesized scenes obey first-principles constraints such as Newtonian motion, collision dynamics, gravity, and biomechanical consistency, thereby moving beyond generator-specific visual artifacts.

14. Conclusions

Deepfake defense has matured into a multi-layer ecosystem spanning passive detection, provenance/authentication standards, watermarking, and governance mechanisms. The strongest empirical signal across benchmarks is that detector generalization remains fragile under realistic distribution shift, motivating layered architectures that use provenance and watermarking where available and reserve content-based detection as an adaptive backstop. Standards efforts (notably C2PA/Content Credentials) provide a credible foundation for cryptographic provenance but must be deployed with explicit threat models and harms awareness to avoid privacy harms and misplaced trust. Emerging legal frameworks, especially in the EU, are shaping the incentives for marking and labeling AI-generated content, increasing the importance of interoperable technical solutions that preserve provenance through real-world media transformations. In practical terms, the layered framework synthesized in this review can inform the design of moderation systems, fact-checking and newsroom verification pipelines, audiovisual forensic screening, public-sector communication safeguards, and compliance-oriented transparency mechanisms for AI-generated content.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No data used in the study.

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

References

1. Altuncu, E.; Franqueira, V.N.; Li, S. Deepfake: Definitions, performance metrics and standards, datasets, and a meta-review. *Front. Big Data* **2024**, *7*, 1400024. [CrossRef] [PubMed]
2. Verdoliva, L. Media forensics and deepfakes: An overview. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 910–932. [CrossRef]
3. National Institute of Standards and Technology. *Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency (NIST AI 100-4)*; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2024. [CrossRef]
4. Defense Advanced Research Projects Agency. MediFor: Media Forensics. 2026. Available online: https://www.darpa.mil/research/programs/media-forensics (accessed on 31 March 2026).
5. Guan, H.; Guan, H.; Delgado, A.; Lee, Y.; Yates, A.N.; Zhou, D.; Kheyrkhah, T.; Fiscus, J. *User Guide for NIST Media Forensic Challenge (MFC) Datasets*; US Department of Commerce: Washington, DC, USA; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2021. [CrossRef]
6. Coalition for Content Provenance and Authenticity. Content Credentials: C2PA Technical Specification (Version 2.3). 2025. Available online: https://spec.c2pa.org/specifications/specifications/2.3/specs/C2PA_Specification.html (accessed on 31 March 2026).
7. Afchar, D.; Nozick, V.; Yamagishi, J.; Echizen, I. Mesonet: A compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*; IEEE: New York, NY, USA, 2018; pp. 1–7. [CrossRef]
8. Yan, Z.; Zhang, Y.; Yuan, X.; Lyu, S.; Wu, B. Deepfakebench: A comprehensive benchmark of deepfake detection. *arXiv* **2023**, arXiv:2307.01426. [CrossRef]
9. Qian, Y.; Yin, G.; Sheng, L.; Chen, Z.; Shao, J. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European Conference on Computer Vision*; Springer International Publishing: Cham, Switzerland, 2020; pp. 86–103. [CrossRef]
10. Jia, S.; Ma, C.; Yao, T.; Yin, B.; Ding, S.; Yang, X. Exploring frequency adversarial attacks for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2022; pp. 4103–4112. [CrossRef]

11. Ciftci, U.A.; Demir, I.; Yin, L. *Fakecatcher: Detection of Synthetic Portrait Videos Using Biological Signals*; IEEE Transactions on Pattern Analysis and Machine Intelligence: Washington, DC, USA, 2020. [CrossRef]
12. Delgado, H.; Evans, N.; Kinnunen, T.; Lee, K.A.; Liu, X.; Nautsch, A.; Patino, J.; Sahidullah, M.; Todisco, M.; Wang, X.; et al. ASVspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan. *arXiv* **2021**, arXiv:2109.00535. [CrossRef]
13. Jung, J.W.; Heo, H.S.; Tak, H.; Shim, H.J.; Chung, J.S.; Lee, B.J.; Yu, H.J.; Evans, N. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: New York, NY, USA, 2022; pp. 6367–6371. [CrossRef]
14. Khalid, H.; Tariq, S.; Kim, M.; Woo, S.S. FakeAVCeleb: A novel audio-video multimodal deepfake dataset. *arXiv* **2021**, arXiv:2108.05080. [CrossRef]
15. SCLBD. *DeepfakeBench [Computer Software]*; GitHub: San Francisco, CA, USA, 2026. Available online: https://github.com/SCLBD/DeepfakeBench (accessed on 31 March 2026).
16. Agarwal, S.; Hu, L.; Ng, E.; Darrell, T.; Li, H.; Rohrbach, A. Watch those words: Video falsification detection using word-conditioned facial motion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*; IEEE: New York, NY, USA, 2023; pp. 4710–4719. [CrossRef]
17. Cozzolino, D.; Pianese, A.; Nießner, M.; Verdoliva, L. Audio-visual person-of-interest deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2023; pp. 943–952. [CrossRef]
18. Oorloff, T.; Koppisetti, S.; Bonettini, N.; Solanki, D.; Colman, B.; Yacoob, Y.; Shahriyari, A.; Bharaj, G. Avff: Audio-visual feature fusion for video deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2024; pp. 27102–27112. [CrossRef]
19. Yin, Q.; Lu, W.; Cao, X.; Luo, X.; Zhou, Y.; Huang, J. Fine-grained multimodal deepfake classification via heterogeneous graphs. *Int. J. Comput. Vis.* **2024**, *132*, 5255–5269. [CrossRef]
20. Wang, Y.; Huang, H. Audio–visual deepfake detection using articulatory representation learning. *Comput. Vis. Image Underst.* **2024**, *248*, 104133. [CrossRef]
21. Norman, J.D.; Farid, H. Detecting Deepfake Talking Heads from Facial Biometric Anomalies. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*; IEEE: New York, NY, USA, 2026; pp. 232–240. [CrossRef]
22. Bohacek, M.; Farid, H. Protecting world leaders against deep fakes using facial, gestural, and vocal mannerisms. *Proc. Natl. Acad. Sci. USA* **2022**, *119*, e2216035119. [CrossRef] [PubMed]
23. Guo, X.; Huo, J.; Shi, Z.; Song, Z.; Zhang, J.; Zhao, J. T2vphysbench: A first-principles benchmark for physical consistency in text-to-video generation. *arXiv* **2025**, arXiv:2505.00337. [CrossRef]
24. Kang, B.; Yue, Y.; Lu, R.; Lin, Z.; Zhao, Y.; Wang, K.; Huang, G.; Feng, J. How far is video generation from world model: A physical law perspective. *arXiv* **2024**, arXiv:2411.02385. [CrossRef]
25. Meng, F.; Liao, J.; Tan, X.; Shao, W.; Lu, Q.; Zhang, K.; Cheng, Y.; Li, D.; Qiao, Y.; Luo, P. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. *arXiv* **2024**, arXiv:2410.05363. [CrossRef]
26. Feng, C.; Chen, Z.; Owens, A. Self-supervised video forensics by audio-visual anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2023; pp. 10491–10503. [CrossRef]
27. Content Authenticity Initiative. Restoring Trust and Transparency in the Age of AI. Retrieved. 2026. Available online: https://contentauthenticity.org/ (accessed on 17 February 2026).
28. Adobe. Content Credentials Overview. 2025. Available online: https://helpx.adobe.com/creative-cloud/apps/adobe-content-authenticity/content-credentials/overview.html (accessed on 31 March 2026).
29. Content Authenticity Initiative. c2pa-rs [Source Code]. GitHub. Retrieved. 2026. Available online: https://github.com/contentauth/c2pa-rs (accessed on 17 February 2026).
30. Google DeepMind. SynthID: A Tool to Watermark and Identify Content Generated Through AI. 2026. Available online: https://deepmind.google/models/synthid/ (accessed on 31 March 2026).
31. Gowal, S.; Bunel, R.; Stimberg, F.; Stutz, D.; Ortiz-Jimenez, G.; Kouridi, C.; Vecerik, M.; Hayes, J.; Rebuffi, S.-A.; Bernard, P.; et al. SynthID-Image: Image watermarking at internet scale. *arXiv* **2025**, arXiv:2510.09263. [CrossRef]
32. Wen, Y.; Kirchenbauer, J.; Geiping, J.; Goldstein, T. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. *arXiv* **2023**, arXiv:2305.20030. [CrossRef]
33. Ci, H.; Yang, P.; Song, Y.; Shou, M.Z. Ringid: Rethinking tree-ring watermarking for enhanced multi-key identification. In *European Conference on Computer Vision*; Springer Nature: Cham, Switzerland, 2024; pp. 338–354. [CrossRef]
34. Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: New York, NY, USA, 2019; pp. 1–11. [CrossRef]
35. Dolhansky, B.; Bitton, J.; Pflaum, B.; Lu, J.; Howes, R.; Wang, M.; Ferrer, C.C. The deepfake detection challenge (dfdc) dataset. *arXiv* **2020**, arXiv:2006.07397. [CrossRef]

36. Li, Y.; Yang, X.; Sun, P.; Qi, H.; Lyu, S. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2020; pp. 3207–3216. Available online: https://openaccess.thecvf.com/content_CVPR_2020/papers/Li_Celeb-DF_A_Large-Scale_Challenging_Dataset_for_DeepFake_Forensics_CVPR_2020_paper.pdf (accessed on 31 March 2026).
37. Jiang, L.; Li, R.; Wu, W.; Qian, C.; Loy, C.C. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2020; pp. 2889–2898. [CrossRef]
38. He, Y.; Gan, B.; Chen, S.; Zhou, Y.; Yin, G.; Song, L.; Sheng, L.; Shao, J.; Liu, Z. Forgerynet: A versatile benchmark for comprehensive forgery analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2021; pp. 4360–4369. [CrossRef]
39. Zi, B.; Chang, M.; Chen, J.; Ma, X.; Jiang, Y.G. Wilddeepfake: A challenging real-world dataset for deepfake detection. In *Proceedings of the 28th ACM International Conference on Multimedia*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 2382–2390. [CrossRef]
40. Zhu, M.; Chen, H.; Yan, Q.; Huang, X.; Lin, G.; Li, W.; Tu, Z.; Hu, H.; Hu, J.; Wang, Y. Genimage: A million-scale benchmark for detecting ai-generated image. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 77771–77782. [CrossRef]
41. Ojha, U.; Li, Y.; Lee, Y.J. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2023; pp. 24480–24489. [CrossRef]
42. Cai, Z.; Ghosh, S.; Adatia, A.P.; Hayat, M.; Dhall, A.; Gedeon, T.; Stefanov, K. AV-Deepfake1M: A large-scale LLM-driven audio-visual deepfake dataset. In *Proceedings of the 32nd ACM International Conference on Multimedia*; ACM: New York, NY, USA, 2024; pp. 7414–7423. [CrossRef]
43. Barrington, S.; Bohacek, M.; Farid, H. The DeepSpeak Dataset. *arXiv* **2024**, arXiv:2408.05366. [CrossRef]
44. Chandra, N.A.; Murtfeldt, R.; Qiu, L.; Karmakar, A.; Lee, H.; Tanumihardja, E.; Farhat, K.; Paik, S.; Lee, C.; Choi, J.; et al. Deepfake-eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024. *arXiv* **2025**, arXiv:2503.02857. [CrossRef]
45. Adobe. C2PA Tool. Open-Source Tools for Content Authenticity and Provenance. 2026. Available online: https://opensource.contentauthenticity.org/docs/c2patool/c2patool-index/ (accessed on 31 March 2026).
46. European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence and Amending REGULATIONS (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). Official Journal of the European Union, L, 2024/1689. EUR-Lex. 2024. Available online: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng (accessed on 31 March 2026).
47. European Commission; Directorate-General for Communications Networks; Content and Technology. How the Digital Services Act Enhances Transparency Online. Shaping Europe's Digital Future. 2026. Available online: https://digital-strategy.ec.europa.eu/en/policies/dsa-brings-transparency (accessed on 31 March 2026).
48. Sunkari, V.; Nagesh, A.S. Artificial intelligence for deepfake detection: Systematic review and impact analysis. *IAES Int. J. Artif. Intell. (IJ-AI)* **2024**, *13*, 3786–3792. [CrossRef]
49. Singh, S.; Dhumane, A. Unmasking digital deceptions: An integrative review of deepfake detection, multimedia forensics, and cybersecurity challenges. *MethodsX* **2025**, *15*, 103632. [CrossRef] [PubMed]