REVIEW ARTICLE

Sensing Deepfake Detection: A Survey of Detection Architectures, Adversarial Challenges, and Critical Applications in Political, Educational, and Military Domains

Alexandros Gazis[1,*], **Stylianos Pappas**[2], **Theodoros Vavouras**[3], **Asim Ali**[4] and **Nikos E. Mastorakis**[2,5]

[1] Department of Electrical and Computer Engineering, Democritus University of Thrace, Xanthi 67100, Greece
[2] Electrical Engineering and Computer Science, Hellenic Naval Academy, Piraeus 18539, Greece
[3] School of Italian Language and Literature, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece
[4] Department of Artificial Intelligence, CECOS University, Peshawar 25000, Pakistan
[5] Faculty of Engineering, Technical University of Sofia, Sofia 1756, Bulgaria

Abstract

Deepfake technology has advanced swiftly, enabling the rapid production of hyper-realistic synthetic media that pose considerable threats to digital security, privacy, military operations, and information integrity. This paper extensively examines visual intelligence and computer vision methodologies for deepfake detection, covering recent developments in deep learning, adversarial strategies, and feature extraction. It reviews prevalent generation architectures—including GANs, autoencoders, neural rendering, and diffusion models—alongside novel adversarial tactics that enhance realism while evading detection, particularly in military and intelligence contexts. We also investigate visual artifacts and manipulation traces, scrutinizing physical discrepancies, digital fingerprints, and physiological signals as critical detection indicators. The article offers a comprehensive analysis of CNN-based, transformer-based, and frequency-domain detection methods, highlighting their advantages, drawbacks, and practical relevance. Furthermore, we examine assessment measures and generalization challenges, while emphasizing prospective research avenues such as explainable AI, self-supervised learning, and federated learning. This study serves as a significant resource for academics and practitioners combating deepfake disinformation in civilian, military, and hybrid threat environments, providing insights into detection improvements and impending issues in hostile AI.

Keywords: deepfake detection, visual sensing and signal analysis, multimodal feature extraction, generative adversarial networks, adversarial robustness, physiological

Submitted: 09 February 2026
Revised: 20 May 2026
Accepted: 11 August 2026
Published: 23 September 2026

Vol. 3, **No.** 3, 2026.
10.62762/TSCC.2026.303040

***Corresponding author:**
✉ Alexandros Gazis
agazis@ee.duth.gr

Citation
Gazis, A., Pappas, S., Vavouras, T., Ali, A., & Mastorakis, N. E. (2026). Sensing Deepfake Detection: A Survey of Detection Architectures, Adversarial Challenges, and Critical Applications in Political, Educational, and Military Domains. *ICCK Transactions on Sensing, Communication, and Control*, 3(3), 176–196.

signal-based detection, frequency-domain analysis, explainable AI, privacy-preserving detection, federated learning.

1 Introduction

Over the last five years, there has been tremendous discussion around Artificial Intelligence (AI), the Fourth Industrial Revolution, and the transition into Industry 5.0. In collaboration with the Internet of Things (IoT) and Big Data, these rapid advancements are expected to transform our everyday lives as well as the industrial and military landscapes. However, rather than delving into AI in general, this paper focuses directly on one of its most recent and controversial developments: deepfakes [1].

Specifically, a deepfake refers to an audiovisual file generated using AI techniques that try to mimic how someone looks, moves, speaks, and behaves, producing an artificial media file that imitates a real person, including military officials or authority figure (the individual being mimicked). As such, a deepfake is a form of complex synthetic media in which a person's face, voice, or both is digitally manipulated using AI to create highly realistic but entirely fabricated audiovisual content. The term "fake" in this context does not refer to issues of consent or unauthorized usage, but rather to the fact that the content does not reflect a truthful or real event; it is not authentic in terms of the information it conveys. Unfortunately, while deepfakes represent a marvel of engineering in terms of machine learning algorithms (especially deep learning), they are most often used to swap faces in videos (e.g., making a person appear to say or do something they never did), mimic someone's voice, or even generate entirely fake online personas [2].

There are also applications of deepfake technology in educational contexts, including the simulation of historical figures for immersive training environments, as broadly catalogued in surveys of deepfake creation and deployment [3], though the same underlying generative mechanisms that enable such applications also raise significant societal and security challenges—including threats to privacy, democratic integrity, and national security [4]—that make reliable detection increasingly urgent. However, the term is more commonly associated with negative connotations, including the spread of misinformation enabled by increasingly sophisticated generation methods documented in the deep learning literature [5], identity theft [6], reputational harm and erosion of public trust in media—as documented from a communication science perspective [10]—compounded by the growing difficulty of reliably distinguishing synthetic from authentic content [7], and, in its simplest form, satirical or humorous misuse. Modern AI techniques now produce hyper-realistic media, making it increasingly difficult to determine, in military intelligence, industrial applications, and broader defense analysis contexts, whether an image, audio, or video is genuine. This growing realism poses significant threats to digital security, personal privacy, media credibility, and public trust [8–10]. One of the most widely used technologies for generating deepfake content is Generative Adversarial Networks (GANs). Specifically, the term GANs is used to explain two competing neural networks, a generator and a discriminator, that work against each other to develop a highly realistic synthetic media. The generator produces fake content based on training data, while the discriminator attempts to detect whether the content is real or generated, forcing the generator to improve with each iteration [11]. Additionally, latent diffusion models (LDMs) are one of the new trends that have arisen as a powerful alternative, generating high-quality content by iteratively denoising random noise [12]. Early visual approaches to deepfake detection demonstrated that optical flow patterns carry discriminative temporal cues; convolutional architectures applied directly to flow maps between consecutive frames can reliably expose motion inconsistencies introduced by face-synthesis pipelines [13]. An equally important part of the deepfake creation pipeline involves autoencoders, models used to compress and reconstruct input data (typically images or video frames). These are often trained specifically on facial features to enable facial swapping and mimicry of expressions by reconstructing one person's face with the structure and movement of another [14].

The primary technical challenges in building convincing deepfakes are the availability of high-resolution training data and the substantial computational power required. This is becoming less of a barrier, however, as computing capabilities grow, not merely following Moore's Law as in the past, but now progressing even faster under Huang's Law [15]. As a result, deepfakes are becoming increasingly sophisticated, making it harder, even for digital analysts and observers, to distinguish authentic content from fabricated material [16, 17].

A special note must be made regarding the so called "arms race", the ongoing competition between deepfake generation techniques and the detection algorithms designed to counter them. As new methods emerge to improve the realism of facial expressions, body movements, and clear voice alignment, detection tools are also evolving in parallel, attempting to keep pace with these rapid advancements. Accordingly, several research initiatives are actively developing deepfake detection models that use the same AI foundations to monitor, identify, and flag deepfake content. Modern detection techniques leverage AI tools to analyze inconsistencies in facial expressions, unnatural lighting conditions, and subtle biological signals such as blinking patterns or facial micro-expressions. These physiological and visual cues are carefully studied as potential indicators of synthetic media, spanning a wide range of generation and detection scenarios [18, 19].

As a result, there is a growing need to pinpoint and examine the current state of the art detection mechanisms, especially those that rely on visual intelligence. Visual intelligence for deepfake detection spans both static and dynamic modalities, including standalone images, audio files, and complex video streams composed of multiple synchronized visual and audio frames. Whether deepfakes are spreading on social media platforms or undermining public trust in news dissemination, robust detection tools are essential for safeguarding users and preventing the spread of misinformation [20, 21]. Detection research, therefore, aims to build models that are robust, generalizable, and explainable. Robust models are able to detect deepfakes and manipulation techniques with high certainty and across various contexts. Generalizable models are capable of performing well on deepfakes generated by new datasets or methods not seen during training. Explainable models offer interpretable outputs to help users understand why a content is flagged as suspicious, thus increasing human oversight and trust [22]. Existing surveys covering both deepfake generation and detection across image, video, and audio domains offer essential technical grounding for contextualizing the scope of this review [23]. Table 1 provides a brief contextual overview of selected deepfake detection works spanning visual forensics, spatial-temporal modelling, and frequency-domain analysis, which are included to situate the detection methods reviewed in this paper within the broader deepfake detection landscape.

Although this survey focuses on visual intelligence and computer vision methods, modern deepfakes often exploit spatial, temporal, and frequency-domain vulnerabilities simultaneously. Therefore, Tables 1 and 2 briefly present selected detection studies—covering foundational media forensics, spatial-temporal transformer approaches, and frequency-guided learning—as contextual background for the visual architectures analysed in the following sections. The main analysis, however, remains focused on visual architectures, spatial-temporal artifacts, frequency-domain cues, physiological visual signals, and image- and video-based detection methods.

The importance of explainability is especially critical in cases where false positives occur. A false positive refers to a scenario where content is flagged as suspicious due to minor inconsistencies, despite not being a deepfake. These cases require human intervention for final validation and interpretation, as research on human-machine collaborative deepfake detection demonstrates that neither crowds nor automated systems alone achieve reliable discrimination [34]. However, there are also notable limitations in current detection approaches, which will be discussed later in this paper. One of the primary challenges is poor generalization; many models overfit specific types of deepfake generations and subsequently fail to detect new or novel techniques in early development stages [35]. Another challenge is the lack of robustness: simple post-processing tricks, lossy compression, or adversarial attacks can mask synthetic signals by simulating human-like behavior or injecting noise to confuse detection algorithms [36, 37]. Furthermore, performance often degrades in real-world environments, especially in visual content, where lighting changes, occlusions, motion blur, or low resolution significantly affect detection accuracy [27].

Our goal is to provide both young researchers and experienced practitioners with a clear understanding of current capabilities, emerging methods, and future needs in the field of deepfake detection, with a particular focus on visual intelligence approaches. To clearly establish the scope and novelty of this study, our main contributions are summarized as follows:

- Comprehensive Overview of Generation Architectures: We provide a detailed examination of state-of-the-art deepfake generation techniques, including GANs, autoencoders, neural rendering, and recent advances in diffusion models.

Table 1. Summary of visual and multimodal deepfake detection studies, included to situate the visual detection methods reviewed in this paper within the broader deepfake detection landscape.

No.	Subject / IT domain	Year	Ref.
1.	Deepfake Detection, Deep Learning Survey	2024	[24]
2.	Multimodal Fusion, Advanced Machine Learning Detection	2023	[25]
3.	Audio-visual Fusion, Deep Learning	2024	[26]
4.	Deepfake Definitions, Performance Metrics & Dataset Meta-Review	2024	[27]
5.	Digital Forensics, AI Evaluation	2023	[28]
6.	Visual Media Forensics, Deepfake Overview	2020	[29]
7.	Spatial-Temporal Transformer, Deepfake Video Detection	2023	[30]
8.	Face Forgery Detection, Generalization & Robustness	2023	[31]
9.	Spatial-Frequency Relation Reasoning, Deepfake Detection	2023	[32]
10.	High-Frequency Fine-Grained Transformer, Face Forgery	2023	[33]

- Analysis of Adversarial Threats and Artifacts: We explore adversarial strategies used to enhance the realism of synthetic content and scrutinize critical manipulation traces, such as facial texture distortions, physiological signals, and image inconsistencies.
- Categorization of Detection Methodologies: We systematically classify deep learning-based detection methods into CNN-based, transformer-based, and frequency-domain approaches, evaluating their performance, robustness, and limitations.
- Evaluation and Generalization Challenges: We review benchmark datasets and commonly used evaluation metrics, explicitly addressing the major challenges models face when attempting to generalize across diverse types of deepfake content.
- Future Research Avenues: We outline key prospective research directions, highlighting the importance of explainable AI and self-supervised learning to shape future advances in visual deepfake detection.

2 Background and motivation

This section is divided into two parts. On the one hand, Section 2.1 explains the role of visual intelligence in deepfake detection. On the other hand, Section 2.2 focuses the survey methodology and selection criteria used in this review.

2.1 Motivating the Role of Visual Intelligence in Deepfake Detection

Before expanding on the technical mechanisms of deepfakes presented in the following sections, it is important to understand the broader technological shift that has enabled their development [38]. Deepfakes are part of a larger trend of AI-driven synthetic media, content generated by machine learning models to replicate, simulate, or fabricate realistic visuals and audio. Although Generative Adversarial Networks (GANs) have become the most well-known and effective tools for creating deepfakes, the issue extends beyond the models themselves. The primary concern with deepfakes is not their technical sophistication per se, but their potential to cause harm and disrupt society—a concern documented from societal and legislative perspectives [39]. These fabricated media files challenge our ability to distinguish real from fake and erode digital trust across multiple domains—from political misinformation to military deception, digital fraud, and identity theft. The breadth of this threat is reflected in the rapidly expanding diversity of manipulation techniques, which now encompasses over 40 distinct generation methods as documented in large-scale detection benchmarks [40], and continues to grow, as documented in early surveys of deepfake technology emergence [41] and confirmed by the most recent comprehensive generation and detection surveys [44]. For instance, in the United States, the threat of deepfake videos targeting political figures and electoral processes has been a growing concern, with scholars warning of their potential to suppress voter participation and undermine democratic integrity [42]. In Europe, synthetic voice imitations have been used to impersonate well-known CEOs and authorize fraudulent wire transfers, as documented in threat scenario analyses of synthetic media in financial systems [43]. Similarly, in various regions, manipulated videos of public figures have emerged on social media platforms, inciting political tensions and undermining public trust in authentic visual content—a phenomenon whose technical underpinnings are comprehensively surveyed across generation and detection dimensions in the recent deepfake literature [44], and whose real-world

political consequences have been documented in studies showing that synthetic video erodes trust and amplifies deception [45]. As time progresses and these tools become more advanced, synthetic content will grow in realism, making it increasingly difficult to detect through conventional means. In this context, visual intelligence, the use of computer vision and AI to interpret, analyze, and understand the properties of visual content, becomes a critical factor for military situational awareness and threat assessment in mitigating the risks posed by deepfake technologies.

Visual intelligence is not just another buzzword or a one-size-fits-all solution. Unlike audio-based or metadata-level analysis, it focuses directly on the image and video content itself, examining the fine-grained patterns relevant to military-grade image and video analysis that are often difficult for the human eye to detect. It is, therefore, not only a tool for validating human judgment but also a means of extending human capability in monitoring and understanding deepfakes. This includes pixel-level monitoring of inconsistencies, unnatural facial movements, lighting mismatches, and temporal anomalies across video frames [25].

Modern visual intelligence techniques, derived from the deep learning domain, particularly convolutional neural networks (CNNs), transformers, and hybrid architectures, are capable of identifying both spatial and temporal cues from large volumes of labeled and, in some cases, even unlabeled, media datasets. A major area of focus is the study of visual artifacts and frequency domain analysis, which involves detecting signal distortions introduced during the synthesis process that are not always identifiable through spatial domain analysis alone [46]. Additionally, physiological signal analysis is an emerging field that involves studying involuntary human signals such as heartbeat-induced skin tone changes, facial micro-expressions, and gaze dynamics [47, 48]. These signals are extremely difficult for generative models to replicate accurately, which makes them a promising direction for future detection systems due to the uniqueness of human emotional and physiological responses. Figure 1 illustrates the processing pipeline from multimodal inputs (images, video, audio) through feature extraction and domain-specific analysis (spatial, temporal, frequency, and physiological cues) to interpretable and robust decision-support outputs. The interpretability component represents explainability mechanisms at a conceptual level and does not correspond to a specific saliency or attribution method.

Specifically, the diagram illustrates the end-to-end processing pipeline: multimodal inputs (images, video, and audio) undergo feature extraction across spatial, temporal, frequency, and physiological domains. These cues are processed by detection models (CNNs, Transformers, and Frequency-based) to produce interpretable and robust decision-support outputs, ultimately classifying media as real or fake while providing confidence scores and explainable visual grounding.

Furthermore, instead of relying solely on black-box models that produce binary outputs (real/fake), it is important to consider visually grounded models that highlight exact regions or temporal segments where anomalies are detected. This approach supports human oversight by providing visual explanations and can help establish and rebuild trust in applications that are heavily scrutinized civilian and military applications, such as journalism, law enforcement, and digital forensics [49].

The core challenge at hand is achieving real-time, explainable, and scalable detection. As such, this paper aims to explore this challenge in depth and provide insight into adaptive approaches in visual intelligence for deepfake detection. It also explains the most promising current methods while allowing the reader to assess whether these tools and techniques can help address the ongoing arms race between content generation and the detection of so called deepfake media.

2.2 Survey Methodology and Selection Criteria

To ensure a comprehensive and objective review of the deepfake detection landscape, a systematic search was conducted across major academic databases, including IEEE Xplore, ScienceDirect, ACM Digital Library, and arXiv.

- Search Keywords and Time Window: The search focused on publications between 2019 and 2026, using primary keywords such as "Deepfake Detection," "Visual Intelligence," "Generative Adversarial Networks (GANs)," and "Adversarial Machine Learning".
- Inclusion and Exclusion Criteria: Studies were included if they proposed novel detection architectures (CNN, Transformer, or Frequency-based), provided empirical results

on benchmark datasets (FF++, Celeb-DF, DFDC), or addressed critical adversarial challenges. Purely social-science-oriented studies without technical implementation were excluded to maintain the focus on visual intelligence.

- Data Aggregation and Consistency: Regarding the performance metrics reported in the comparative performance analysis presented earlier in this paper, results are compiled from studies utilizing standardized evaluation protocols. To maintain consistency, we prioritized results obtained from high-quality (c23) or raw data splits and standardized metrics (AUC/EER) to ensure that the comparative analysis reflects a reliable "apples-to-apples" assessment across different architectural approaches.

3 Definition and characteristics of deepfakes

To define the term, deepfakes refer to synthetic media generated using deep learning algorithms, primarily Generative Adversarial Networks (GANs) and autoencoders. These media types are mainly produced in the form of images, videos, and audio files. The term itself combines “deep learning” and “fake,” denoting AI-generated content that attempts to convincingly replicate real people or real civilian and military events. In most cases, deepfakes are created to replace, simulate, or mimic human appearances and voices in a way that produces hyper-realistic outputs, making it extremely difficult to distinguish them from genuine media. The key difference between deepfakes and traditional computer-generated imagery (CGI) or other visual effects is that deepfakes are data-driven: they can be automatically generated from diverse input datasets without requiring manual animation or scripting. Their increasing realism has driven the development of more generalizable detection methods, such as those exploiting blending boundary artifacts common to most face-swap pipelines [50], while their applications simultaneously extend to more serious societal concerns. As a result, the realism achieved by these models, based on the richness of the training data, often leads to outputs that are nearly indistinguishable from authentic audiovisual recordings. In recent years, facial manipulation represents a particularly consequential class of deepfake targeting public figures and military officials, encompassing sequential manipulation steps that detection systems must be capable of identifying and reversing [51]. This includes face-swapping, expression transfer, and full-face mimicry across multiple video frames. Additionally, voice synthesis has garnered increasing attention, particularly in military impersonation scenarios. These voice-based deepfakes not only replicate a person’s vocal tone but also mimic their accent, cadence, and speech patterns; when coupled with synthesized lip motion, the resulting audio-visual forgeries expose cross-modal inconsistencies that joint audio-visual detectors are specifically designed to exploit [52]. As noted earlier, the low barriers to entry due to the availability of pre-trained models and open-source software have made the creation of deepfakes easier than ever, while making them correspondingly difficult to detect and trace. Moreover, the multi-modality of deepfake generation, i.e., the combination of facial, vocal, and movement synthesis, enables the creation of real-time deepfakes. The increasing complexity of such multi-modal forgeries has correspondingly driven the development of detection frameworks capable of reasoning jointly across visual, audio, and motion cues [53]. Although this currently requires high-end hardware and computational resources, these are no longer out of reach for the average user. As such, it is now possible to conduct live manipulation of a target’s identity or appearance during events or broadcasts, or even military briefings.

Whether the intent is malicious or benign, the applications and use cases of deepfakes are vast. On the one hand, there are beneficial applications such as educational simulations, where historical figures are recreated for training or immersive learning environments, or entertainment uses like parody videos and special effects. On the other hand, the potential harms may outweigh the benefits. Nonconsensual adult content, political propaganda, defamation, and identity fraud are increasingly common and often more impactful than the legitimate uses of this technology. We will further explore this in the following chapters.

Deepfakes exhibit a range of characteristics that can be analyzed through visual intelligence systems, such as:

1. **Differences in visual artifacts relevant to military media forensics**: Inconsistencies such as abnormal skin texture, irregular lighting, and mismatched shadows around facial features often occur briefly, usually in less than a second. These are difficult to detect without advanced computer vision techniques [54].

2. **Temporal inconsistencies**: Unnatural blinking rates, inconsistent head movement, and lip-sync errors

often become apparent over time. These are typically identified through frame sequence analysis used in military surveillance and video stream models such as Recurrent Neural Networks (RNNs) and transformer architectures [55].

3. **Frequency domain anomalies**: By operating in the frequency domain, analysts can detect unnatural signal patterns introduced during the synthesis process that are not visible in the spatial domain; recent approaches combine frequency-domain masking with spatial interaction to further improve generalization [56].

4. **Inconsistent biological signals**: Deepfakes often fail to replicate subtle human signals such as involuntary eye movements, facial microexpressions, and muscle dynamics. These cues are highly individual and difficult for generative models to reproduce [57].

5. **Anatomical irregularities**: Distorted or misaligned features, such as fingers, ears, or limbs, are often visible in deepfakes. These physical inaccuracies, common when body parts are rendered in motion or at unusual angles, have been documented in challenging real-world deepfake collections and are among the detectable indicators that automated systems are designed to identify [58].

6. **Compression and encoding artifacts**: During online distribution, the interaction between synthetic data and compression algorithms can introduce visible distortions. These artifacts serve as useful indicators when analyzing media integrity [59].

4 Impact and Potential Threats: Scope and Organization of the Survey

The consistent evolution of deepfake technologies, combined with our limited understanding of their broader impact, is closely associated with significant future threats [60]. This section outlines the dangers posed by deepfakes across nearly every aspect of human activity, from social to political and technological domains, and places the technical contributions of deepfakes within a larger framework of actions that will be further discussed in the following sections. First and foremost, the entropy of digital ecosystems, that is, the overwhelming proliferation of synthetic content, introduces a high degree of uncertainty in the verification of digital content. This severely undermines the value of digital evidence, especially in critical contexts such as journalism, legal courts, and public information sources. Another serious threat posed

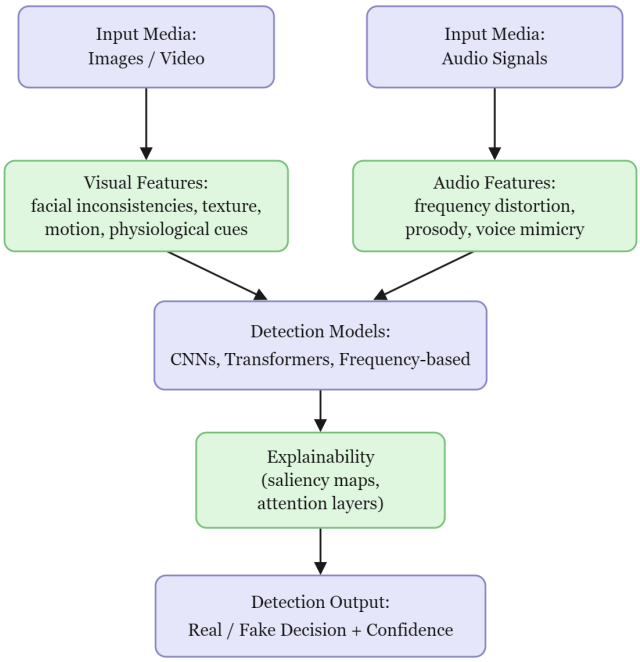


Figure 1. Conceptual sensing and control framework for visual intelligence-based deepfake detection: from multimodal signal acquisition and feature extraction to interpretable decision-support outputs.

by hyper-realistic synthetic data is its weaponization, particularly in the realm of geopolitical influence and military influence. Deepfakes can be used to manipulate elections, fabricate diplomatic incidents or false military escalations, and incite public unrest—threats documented from both strategic and legal perspectives [61, 62].

Additionally, there is a growing concern about the erosion of accountability, examined here from a digital governance and audit perspective [63]. The emergence of highly convincing synthetic media may give rise to "plausible deniability," where even genuine offensive acts can be dismissed as fabrications. This phenomenon, often referred to as the "liar's dividend", allows wrongdoers to discredit authentic footage in military accountability scenarios by claiming it is fake. Simultaneously, traditional civilian and military content moderation pipelines, including keyword filtering and automated flagging systems, are rapidly becoming obsolete and ineffective against AI-generated visual content [64].

Another critical issue is the erosion of public trust: GAN-synthesized faces are now rated as indistinguishable from—and even more trustworthy than—real faces, while viewers cannot reliably detect deepfake videos yet remain overconfident in their

ability to do so [65, 66]. As AI systems are increasingly integrated into sensitive domains such as medicine (e.g., for diagnostic imaging), education (e.g., intelligent tutoring systems), or military command systems, skepticism and fear around their misuse may overshadow their potential benefits. The legacy of deepfake abuse may erode public confidence [67], as empirical evidence shows that even trained human observers struggle to reliably distinguish real from synthetic political content, further undermining trust in otherwise legitimate and beneficial applications of AI. However, the core technical discussion of this paper remains centered on visual intelligence and computer vision-based detection.

Lastly, on a more technical note, this article draws on methods from computer vision, signal processing, and cognitive science to offer a more holistic understanding of deepfake detection. It aims to provide a cross-disciplinary perspective, focusing on practical applications that demonstrate both the robustness and the explainability of visual intelligence systems in real-world scenarios. This work does not constitute an empirical evaluation but rather a comparative exploration of different models and architectural approaches, highlighting their strengths and limitations in addressing the deepfake challenge.

5 Overview of Deepfake Generation

This section provides an overview of core deepfake generation mechanisms, beginning with widely used techniques such as face swapping, face reenactment, and audiovisual synthesis. We then review the principal architectures underlying these systems, including GAN-based models, autoencoder-based methods, neural rendering approaches, and recent diffusion models-highlighting how each contributes to the visual realism of synthetic media. Finally, we examine adversarial deepfake generation, which enhances the ability of synthetic content to bypass detection by civilian and military-grade detectors and represents one of the most challenging developments in modern deepfake creation.

We begin by introducing face-swapping techniques, arguably the most iconic and widely recognized form of deepfake generation, where a person's face is overlaid onto another individual's body. This process involves several steps, including facial landmark detection, facial alignment, mask creation, and final image blending. Deep learning models such as autoencoders and GANs -which will be discussed in more detail in subsequent sections- are typically trained to reconstruct the facial features of both source and target identities. The encoder component is responsible for extracting facial features from the source, while the decoder reconstructs them using the spatial configuration of the target. Well-known models in this domain include StyleGAN [68], and FaceSwap GAN [69]. More recent face-swapping approaches incorporate attention mechanisms to preserve motion constraints and improve realism [70]. These models often utilize scale-consistent loss functions to minimize identity distortion and facial expression mismatches. Some lightweight architectures now enable real-time implementation of face swaps for live streaming and video conferencing applications, and military communication applications [71].

Another important technique is face reenactment [72], where facial expressions, movements, and even gaze direction from one individual are transferred onto the face of another person without replacing the face itself. Unlike face-swapping, which changes the visual identity of the subject, reenactment preserves the target's appearance while animating it using the source's motion dynamics. This process is also known as motion transfer or expression cloning; a comprehensive review of such reenactment and expression-swap techniques is provided in [73]. A widely known system in this area is Face2Face [74]. Other advanced approaches employ 3D Convolutional Neural Networks (3D-CNNs) [75], or neural rendering frameworks that leverage 3D morphable model representations [76], to model facial dynamics, often using facial action units or dense optical flow across video frames—the same behavioural signals that identity-based detectors later exploit to protect world leaders from reenactment-driven impersonation [77]. Applications of this technology include subject-agnostic face swapping and reenactment for visual dubbing and animation [78], self-blended image synthesis for advancing detection robustness [79], and the development of multi-scale convolutional and vision-transformer architectures for improved deepfake detection [80], alongside proactive defence strategies such as identity watermarking that exploit these reenactment pipelines to embed protective signals [81], as well as more concerning uses such as journalistic manipulation [82, 128], or forensic falsification [82].

A third category is full body synthesis [83], encompassing a broad spectrum of generation and detection methodologies, where the aim is to

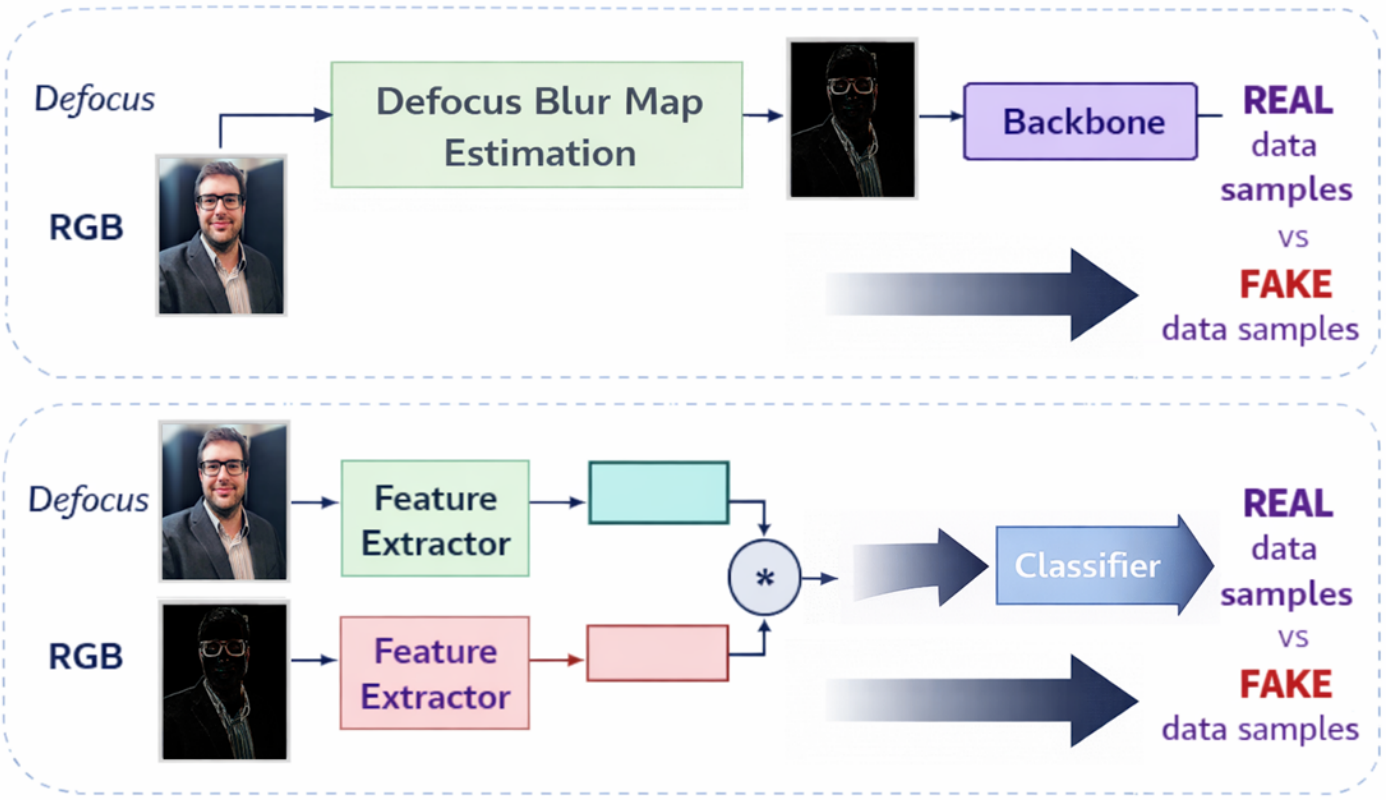


Figure 2. Defocus-based deepfake detection pipeline.
Top: RGB image → defocus blur map estimation → backbone for real/fake classification.
Bottom: defocus extractor + adder/classifier for real/fake artifacts.
As such, the illustration exploits generation models' (GANs, autoencoders, etc.) failure to replicate optical blur.

replicate an individual's entire physical appearance and movements. These systems are significantly more complex due to the need to model body pose, skeletal motion, clothing dynamics, and environmental interactions. Tools like OpenPose [84], and DensePose [85], have become popular in recent years for extracting detailed body pose data. Once the pose is extracted, generative models synthesize each frame to replicate the same set of actions. Generative Adversarial Networks with temporal consistency modules are often employed to reduce jittering and preserve spatial structure [86]. Recent innovations include Neural Radiance Fields (NeRF) [87], and 3D-aware synthesis approaches that exploit three-dimensional semantic representations—a property that, crucially, also introduces inherent spatio-temporal artifacts detectable within 3D morphable model spaces [88].

Finally, audiovisual synthesis is a more traditional deepfake generation method that focuses on synchronizing synthetic speech with matching facial and lip movements. These models combine speech recognition and facial animation to produce hyper-realistic outputs based on either text input or real speech. This process involves encoding speech signals using convolutional or transformer-based architecture and decoding them into corresponding lip motions and facial dynamics. Popular models include Wav2Lip [89], and SyncNet [90], which are capable of precise synchronization and even mimicking emotional tone.

5.1 Common Deepfake Architectures

Although deepfake generation has traditionally focused on generative models capable of synthesizing hyper-realistic visual and auditory data, recent years have seen the rapid evolution of deep learning techniques in this space. Figure 2 illustrates a defocus-based detection pipeline that exploits visual artifacts often produced by principal deepfake generation architectures, such as inconsistent blur in GAN-based, autoencoder-based, neural rendering, and diffusion-based approaches.

This dual-stream architecture highlights how detection systems exploit the technical limitations of generation models (such as GANs and autoencoders) in replicating natural optical blur. Top: A single-stream approach where an RGB image is used for defocus blur

map estimation before passing through a backbone for classification. Bottom: A multi-feature approach utilizing a defocus extractor and a feature adder to identify subtle synthetic artifacts. The pipeline specifically targets inconsistent blur patterns that often appear in AI-generated facial textures and backgrounds.

The primary technologies that underpin deepfake synthesis include Generative Adversarial Networks (GANs) [91], autoencoder-based architectures [92], and neural rendering systems [93], with diffusion models emerging as one of the most promising recent developments. To avoid repetitive explanations and provide a clear overview, we present a structured comparison of these core architectures:

- Generative Adversarial Networks (GANs) [94]: As introduced earlier, and comprehensively surveyed in the literature, GANs utilize an adversarial mechanism that is highly effective for face generation, style transfer, and domain adaptation. Variants like StyleGAN [95], CycleGAN [96], and StarGAN [97], allow for enhanced control over facial attributes, expressions, and lighting. Limitations: They are known to suffer from training instabilities [98], and issues such as mode collapse [99], making them less reliable for high-resolution outputs.
- Autoencoder-based Architectures (see also the detection taxonomy in [22]): These methods compress input data into a latent representation via an encoder and reconstruct it using a decoder. In face-swapping, dual autoencoders share a common encoder but use identity-specific decoders to project different identities while preserving facial geometry. As for their advantages and respective limitations, they are valued for training stability and lower computational requirements compared to GANs [100], but tend to produce blurrier outputs, particularly in textures like skin or under dynamic lighting.
- Neural Rendering [101]: Combining classical graphics pipelines with deep neural networks, this approach offers greater control over geometry, lighting, and camera parameters. Notable examples include deferred neural rendering [102], neural textures, and volumetric scene representations. Advantages: They are particularly well-suited for 3D-aware generation, ensuring viewpoint consistency and supporting multi-angle synthesis.
- Diffusion Models [103]: These models gradually add noise to training data and learn to reverse the diffusion process to reconstruct clean content. Notable examples include DDPM and Stable Diffusion [12], which demonstrate state-of-the-art visual quality. As for their advantages and respective limitations, they offer more stable training and broader diversity than GANs without mode collapse. Although computationally intensive, advancements like DDIMs [104], and rapid improvements in GPU processing power under Huang's Law [15], are enabling more practical deployment.

5.2 Adversarial Deepfake Generation

Adversarial deepfake techniques play a significant role in the deepfake ecosystem, both in terms of enhancing the fidelity of generated media and actively challenging detection mechanisms. These techniques are central to the ongoing "arms race" between generation and detection systems. Generative AI has increasingly been weaponized both as a tool for creating deceptive content and as a means of actively probing and defeating biometric recognition systems, with face recognition representing one of the primary targets of such adversarial exploitation [105], driving a cycle of escalation in which each advance in generation capability demands a corresponding response from detection research.

One key application is adversarial robust deepfake synthesis [106], where the generative model is trained not only to produce realistic content but also to maintain its believability under various transformations. These transformations may include compression, scaling, or partial occlusion of the audiovisual input. This robustness is typically achieved through adversarial loss functions that incorporate feedback not only from the discriminator but also from auxiliary detection networks. These auxiliary networks simulate what a detection system is likely to analyze and highlight, and the generator is optimized to avoid triggering those indicators, effectively embedding a counter-detection mechanism into the synthesis process.

Moreover, adversarial techniques are used to enhance perceptual realism by incorporating multi-objective loss functions [107], a design paradigm established in high-resolution image synthesis frameworks where adversarial, feature-matching, and perceptual losses

Table 2. Contextual overview of selected audio and multimodal deepfake detection surveys, included to situate visual deepfake detection within the broader synthetic media landscape.

Approach	Characteristics	Advantages	Limitations
CNN-Based	Analyzes spatial features and local artifacts (textures).	High spatial Accuracy; Efficient.	Weak temporal Consistency; Noise sensitive.
Transformer	Uses self-attention for global temporal context.	Detects lip-sync and blinking anomalies.	High compute; Slow inference.
Freq. Domain	Detects artifacts in the frequency domain (DCT).	Exposes visually flawless manipulations.	Weak against compression.

are jointly optimized. These may include perceptual loss [108] and identity-preservation constraints, which are particularly critical given the forensic and privacy implications of identity manipulation in synthetic media [109], as well as feature matching loss, which aligns intermediate CNN feature distributions between generated and real images [107], a design consideration of particular relevance to detection systems deployed in national security contexts [110]. Such losses allow the generator to better capture fine-grained identity and texture features, while simultaneously learning to bypass common detection patterns. Another use of adversarial methods involves pixel-level perturbations, which can cause machine learning–based detectors to misclassify fake content as real. The effectiveness of such frequency-level perturbations has been empirically demonstrated even against detectors specifically hardened to resist them [111], revealing the difficulty of building robustness against perturbation-based evasion. These adversarial perturbations can be applied post hoc to already generated deepfakes, making them resistant to classifiers without significantly affecting their visual integrity. Understanding precisely how such perturbations affect detector decisions has motivated parallel work on perturbation-based explainability methods, which use adversarially generated samples to probe and improve the interpretability of detection models [112]. Such perturbations are particularly effective at deceiving CNN-based classifiers or frame-level analysis tools. Recent work has demonstrated that transferable adversarial attacks—where perturbations crafted to fool one detection model remain effective against others with different architectures or training data—pose a practical threat to visual deepfake detectors: black-box and universal perturbations have been shown to bypass the winning DFDC detectors while surviving image and video compression [113]. These emerging methods are especially promising, as they enable adaptive responses to detection threats in real time. For instance, adversarial patches can be dynamically applied to deepfake content, and self-supervised evasion training allows the generative model to become aware of and adapt to existing detection frameworks. The newest generation of adversarial tools moves beyond surface-level feature manipulation and exploits the uncertainty inherent in detection systems to craft more effective evasion strategies. In response, detection frameworks have begun to incorporate ensemble-based uncertainty estimation as a countermeasure [114], aggregating predictions across multiple models to reduce the exploitable confidence gaps that adversarial inputs typically target. As deepfake generation becomes more sophisticated, adversarial techniques will likely remain a critical component in the ongoing development of both offensive and defensive strategies within military, industrial or other AI-generated media landscapes.

5.3 Comparative Performance Analysis of Visual Deepfake Detectors

To provide a clear and structured overview of the primary deep learning-based detection methodologies discussed in this study, Table 2 summarizes the key characteristics, advantages, and limitations of CNN-based models, transformer-based approaches, and frequency-domain techniques.

Beyond these general characteristics, the operational robustness of specialized detection modalities is heavily contingent on environmental conditions and post-processing. For frequency-domain methods, it is empirically observed that while they excel at identifying upsampling and interpolation artifacts, their performance often degrades under heavy lossy compression (e.g., JPEG or H.264), which can mask high-frequency synthetic fingerprints. These methods assume that the underlying grid of the synthesis process remains detectable, an assumption that holds true in prior reports for high-bitrate media but becomes fragile in social media distribution scenarios.

In contrast, physiological-cue methods, such as those

monitoring heart-rate induced skin tone changes or gaze dynamics, are theoretically more resilient to adversarial pixel perturbations as they rely on high-level biological consistency. However, empirical evidence suggests these methods are highly sensitive to resizing and low-resolution inputs, which can destroy the subtle temporal signals required for accurate biological estimation. These robustness profiles are derived from a combination of prior forensic reports and the comparative analysis of cross-dataset performance benchmarks presented in this section.

To facilitate a clearer understanding of the data landscapes used to train and validate modern detection models, we provide a structured summary of the most prominent deepfake datasets. These benchmarks represent a progression from early, manually-intensive manipulations to large-scale, AI-driven forgeries that present significant generalization challenges:

- FaceForensics++ (FF++) [115]: Released in 2019, this dataset contains 1,000 original videos with 4,000 manipulated versions covering four techniques: FaceSwap, DeepFakes, Face2Face, and NeuralTextures. It is particularly useful for evaluating models against different compression levels, though detectors often overfit to its specific manipulation methods.
- Celeb-DF (v2): A 2020 dataset featuring 5,639 high-quality deepfake videos generated using advanced GAN-based face swapping. It was specifically designed to reduce visual artifacts, representing a significant jump in realism that makes traditional boundary-based detection difficult [116].
- Deepfake Detection Challenge (DFDC): One of the largest available benchmarks (2020), consisting of over 100,000 videos with diverse ethnic backgrounds, lighting, and resolutions. It presents extreme challenges for models due to high variability and real-world recording conditions that mimic social media environments.
- UADFV: A 2019 benchmark focusing on early face-swapping (FakeApp), containing 49 real and 49 fake videos. While smaller in scale, it remains a key resource for researchers focusing on physiological inconsistencies, such as unnatural blinking rates.

Figures 3 and 4 provide a literature-based comparative summary of reported detector performance on commonly used benchmarks such as FF++, Celeb-DF, and DFDC. These results are not reproduced experimentally by the present study; rather, they are compiled from prior published works, including [117], and are presented here solely to illustrate comparative trends across the literature under broadly comparable evaluation settings.

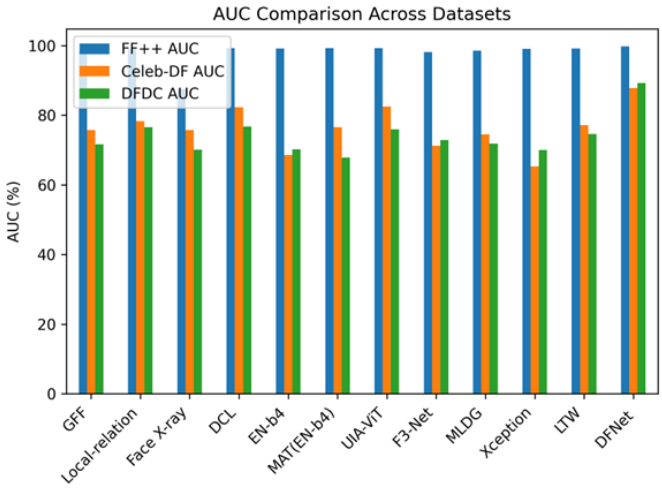


Figure 3. AUC comparison of top deepfake detectors on FF++, Celeb-DF, and DFDC. Higher AUC shows better separation of real vs. fake videos.

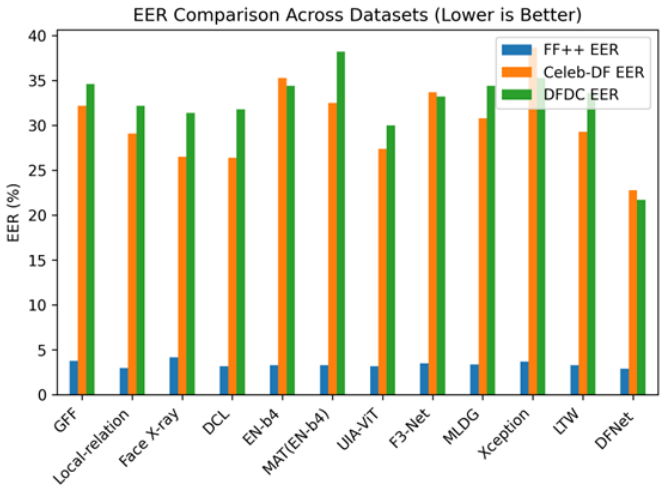


Figure 4. EER comparison of top deepfake detectors on FF++, Celeb-DF, and DFDC. Lower EER means better accuracy where false positives equal false negatives.

To complement the trend-level comparisons in Figures 3 and 4, Table 3 presents a structured summary of representative detector performance under a standardised cross-dataset protocol, illustrating how different architectural families—CNN-based, transformer-based, and frequency-domain methods—trade off intra-dataset accuracy against cross-dataset generalisation. Notably, while

Table 3. Comparative performance of representative visual deepfake detectors on FF++ (c23), Celeb-DF, and DFDC benchmarks. AUC values (%) are reported under the standard protocol of training on FF++ (c23) and cross-dataset testing, unless otherwise noted. – denotes not reported under this protocol in the original study. † denotes training on Blended Images (BI) rather than FF++, per the original paper's experimental design.

Method	Type	Year	FF++ AUC	Celeb-DF AUC	DFDC AUC
Xception [115]	CNN	2019	99.7	65.3	–
Face X-ray† [50]	CNN (boundary)	2020	87.4	80.6	80.9
ISTVT [30]	Transformer	2023	99.3	67.5	69.8
F2Trans [33]	Freq.+Transf.	2023	–	89.9	–

Xception [115] achieves near-perfect AUC on FF++ (c23), its cross-dataset performance on Celeb-DF drops sharply to 65.3%, underscoring the generalisation gap that motivates much of the subsequent research. Face X-ray [50], trained on blended images rather than manipulation-specific data, substantially narrows this gap, confirming the value of manipulation-agnostic training strategies. Transformer-based (ISTVT [30]) and frequency-domain-enhanced (F2Trans [33]) approaches demonstrate improved cross-dataset performance, reflecting the architectural progression discussed throughout this section.

To ensure these performance metrics are accessible to readers across different disciplines, it is helpful to define them in practical terms. The Area Under the Curve (AUC) measures a model's overall ability to distinguish between genuine and manipulated videos. A higher AUC percentage means the model is better at correctly identifying real versus fake content, with 100% representing perfect accuracy and 50% indicating random guessing. Conversely, the Equal Error Rate (EER) represents the specific point where the system's false positive rate (incorrectly flagging a real video as a deepfake) exactly matches its false negative rate (failing to detect an actual deepfake). For the EER metric, a lower percentage is better, as it indicates the model makes fewer overall errors when balancing these two types of mistakes.

The AUC and EER values presented in Figures 3 and 4 are compiled from established benchmarks reported in recent literature, specifically focusing on results achieved under standardized evaluation protocols. To ensure a meaningful comparison, the reported metrics for each model correspond to testing on the high-quality (c23) or raw versions of the FaceForensics++ (FF++) dataset, alongside the test splits of Celeb-DF and the Deepfake Detection Challenge (DFDC). While individual study settings for frame sampling may vary slightly, the consistency of these benchmarks allows for a reliable assessment of how different architectures, such as CNNs and Transformers, generalize across diverse synthetic media distributions.

Finally, as for recent advancements in late 2025 and early 2026, researchers have shifted the focus toward more specialized detection challenges. Specifically, emerging research has begun to address the unique artifacts found in diffusion-based generative models, which often lack the traditional upsampling patterns found in GANs. Furthermore, the introduction of 3D-aware benchmarks and multi-disciplinary tools, such as those from the vera.ai project, represents a move toward more integrated, real-world forensic toolkits. These latest developments emphasize the transition from simple binary classification to more nuanced, interpretation-heavy systems that combine pixel-level forensics with broad contextual analysis.

Notwithstanding these advances, several critical open challenges remain. Robustness against adversarial deepfakes represents one of the most pressing concerns [118, 119]: increasingly sophisticated generation methods are designed not only to improve visual realism but also to evade both human judgment and automated detection systems [120, 121], a phenomenon empirically confirmed by evaluations showing that even publicly accessible detection tools fail under realistic deployment conditions [122]. Adversarial tactics increasingly overlap with evasion strategies that exploit visual identity manipulation, against which identity- and context-aware detectors have been proposed [123, 124], a challenge further amplified by the broader societal disruptions introduced by generative AI [121]. Many existing detectors also suffer from limited cross-dataset generalization [125], often performing well under controlled benchmark conditions but degrading significantly when exposed to unseen manipulation techniques [126–128], compression artifacts [129], noise [130], real-world recording variability [131], or the multimodal forensic identification challenges encountered across diverse real-world deployment scenarios and social media platforms [132]. The practical deployment of detection systems in operational environments [133] further demands

computational efficiency, explainability, scalability, and near-real-time decision-making, compounded by the need to move beyond binary classification toward precise localization of manipulated regions within images and videos [134]—a capability essential for forensic analysis and accountability in domains such as journalism [128], digital forensics [135], and public communication, where synthetic political video has been empirically shown to erode trust and amplify deception [45]. These challenges are especially consequential in education [120] and security-sensitive environments, where false positives, false negatives, and poor interpretability may have serious consequences, motivating the development of hardened detection architectures incorporating both classical and quantum-resilient neural network designs [136]. Future directions therefore include robust multimodal detection, self-supervised and transferable learning strategies, adversarially resilient architectures, explainable AI [137], cross-domain benchmarking [138], privacy-preserving detection frameworks [139], and human-in-the-loop validation mechanisms. An adjacent challenge warranting future attention concerns the converging risks of visual deepfakes and large language model-generated content [140], whose joint exploitation increasingly demands unified detection frameworks to support trustworthy and deployable systems.

6 Conclusion

Deepfake technology has advanced rapidly, enabling the creation of hyper-realistic synthetic media that pose substantial risks to digital security, privacy, and information integrity. This paper presents a detailed examination of visual intelligence and computer vision methodologies for deepfake detection, incorporating recent developments in deep learning, adversarial strategies, and feature extraction techniques. It reviews major deepfake generation architectures, including GANs, autoencoders, neural rendering, and diffusion models, alongside adversarial approaches designed to increase realism while evading detection. The study also explores visual artifacts and manipulation traces, assessing physical inconsistencies, digital fingerprints, and physiological signals that serve as key indicators for detection systems. Building on this, we provide a comprehensive analysis of CNN-based, transformer-based, and frequency-domain detection methods, outlining their strengths, limitations, and practical applicability. We further assess benchmark datasets, evaluation metrics, and the generalization challenges faced by current detectors, and highlight emerging research directions such as multimodal fusion, explainable AI, self-supervised learning, and decentralized privacy-preserving frameworks based on blockchain and federated learning. Overall, this work serves as a valuable resource for researchers and practitioners addressing deepfake-driven misinformation, offering insights into current advances and future challenges in adversarial AI.

Despite the substantial progress achieved in visual deepfake detection, several important research challenges remain open, as outlined in the preceding section. One of the most critical issues is robustness against adversarial deepfakes, as increasingly sophisticated generation methods are designed not only to improve visual realism but also to evade both human judgment and automated detection systems—a phenomenon empirically confirmed by comparative evaluations showing that even publicly accessible detection tools fail under realistic deployment conditions. Adversarial tactics increasingly overlap with evasion strategies that exploit visual identity manipulation, a challenge further amplified by the broader societal disruptions introduced by generative AI. In addition, many existing detectors still suffer from limited cross-dataset generalization, often performing well under controlled benchmark conditions but degrading significantly when exposed to unseen manipulation techniques, compression artifacts, noise, real-world recording variability, or the multimodal forensic identification challenges encountered across diverse real-world deployment scenarios and social media platforms. Another major challenge concerns the practical deployment of deepfake detection systems in operational environments, where methods must be not only accurate, but also computationally efficient, explainable, scalable, and capable of supporting near-real-time decision-making. This challenge is further compounded by the need to move beyond binary classification toward precise localization of manipulated regions within images and videos—a capability essential for forensic analysis and accountability. This is especially important in domains such as journalism, digital forensics, public communication, education, and security-sensitive environments, where false positives, false negatives, and poor interpretability may have serious consequences, motivating the development of hardened detection architectures incorporating both classical and quantum-resilient neural network designs. Future research should therefore focus

on robust multimodal detection, self-supervised and transferable learning strategies, adversarially resilient architectures, and more realistic evaluation settings that better reflect practical deployment conditions. Greater emphasis should also be placed on explainable AI, cross-domain benchmarking, privacy-preserving detection frameworks, and human-in-the-loop validation mechanisms. While the present survey focuses on visual intelligence, an adjacent challenge warranting future attention concerns the converging risks of visual deepfakes and large language model-generated content, whose joint exploitation increasingly demands unified detection frameworks to support trustworthy and deployable systems.

Data Availability Statement

Not applicable.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

This study is a review article and does not involve human participants, clinical trials, or animal experiments. The face image used in Figure 2 for illustrative purposes is that of the corresponding author, included with explicit consent.

References

[1] Whittaker, L., Mulcahy, R., Letheren, K., Kietzmann, J., & Russell-Bennett, R. (2023). Mapping the deepfake landscape for innovation: A multidisciplinary systematic review and future research agenda. *Technovation, 125*, 102784. [CrossRef]

[2] Chadha, A., Kumar, V., Kashyap, S., & Gupta, M. (2021, May). Deepfake: an overview. In *Proceedings of second international conference on computing, communications, and cyber-security: IC4S 2020* (pp. 557-566). Singapore: Springer Singapore. [CrossRef]

[3] Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM computing surveys (CSUR), 54*(1), 1-41. [CrossRef]

[4] Chesney, B., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev., 107*, 1753. https://access.heinonline.com/HOL/LandingPage?handle=hein.journals/calr107&div=51

[5] Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh-The, T., Nahavandi, S., ... & Nguyen, C. M. (2022). Deep learning for deepfakes creation and detection: A survey. *Computer Vision and Image Understanding, 223*, 103525. [CrossRef]

[6] Agarwal, A., & Ratha, N. (2023). Manipulating faces for identity theft via morphing and deepfake: Digital privacy. In *Handbook of statistics* (Vol. 48, pp. 223-241). Elsevier. [CrossRef]

[7] Wang, T., Liao, X., Chow, K. P., Lin, X., & Wang, Y. (2024). Deepfake detection: A comprehensive survey from the reliability perspective. *ACM Computing Surveys, 57*(3), 1-35. [CrossRef]

[8] Pant, P., Mishra, K. K., & Rajaram, R. (2025). Risks to Individual Privacy and Security in the World of Deepfake Technology. In *Mastering Deepfake Technology: Strategies for Ethical Management and Security* (pp. 237-253). River Publishers. https://www.taylorfrancis.com/chapters/edit/10.1201/9788743801146-14

[9] George, A. S., & George, A. H. (2023). Deepfakes: the evolution of hyper realistic media manipulation. *Partners Universal Innovative Research Publication, 1*(2), 58-74. [CrossRef]

[10] Boediman, E. P. (2025). Exploring the impact of deepfake technology on public trust and media manipulation: A scoping review. *Jurnal Komunikasi, 19*(2), 313-334. [CrossRef]

[11] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM, 63*(11), 139-144. [CrossRef]

[12] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022, June). High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 10674-10685). ieee. [CrossRef]

[13] Amerini, I., Galteri, L., & Caldelli, R. (2019). Deepfake video detection through optical flow based cnn. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)* (pp. 1205-1207). IEEE Computer Society. [CrossRef]

[14] Das, M. K., Kumar, M., Kapil, I. K., & Yadav, R. K. (2023, May). Deepfake creation using GANs and autoencoder and Deepfake detection. In *2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN)* (pp. 1-6). IEEE. [CrossRef]

[15] Gambín, Á. F., Yazidi, A., Vasilakos, A., Haugerud, H., & Djenouri, Y. (2024). Deepfakes: current and future trends. *Artificial Intelligence Review, 57*(3). [CrossRef]

[16] Hoque, M. A., Ferdous, M. S., Khan, M., & Tarkoma, S. (2021). Real, forged or deep fake? Enabling the ground truth on the internet. *IEEE Access, 9*, 160471-160484. [CrossRef]

[17] Ghiurău, D., & Popescu, D. E. (2024). Distinguishing reality from AI: approaches for detecting synthetic content. *Computers, 14*(1), 1. [CrossRef]

[18] Daukantas, P. (2025). Generating and detecting deepfakes: A 21st-century arms race. *Optics and Photonics News, 36*(2), 24-31. [CrossRef]

[19] Patel, Y., Tanwar, S., Gupta, R., Bhattacharya, P., Davidson, I. E., Nyameko, R., ... & Vimal, V. (2023). Deepfake generation and detection: Case study and challenges. *IEEE Access, 11*, 143296-143323. [CrossRef]

[20] Mubarak, R., Alsboui, T., Alshaikh, O., Inuwa-Dutse, I., Khan, S., & Parkinson, S. (2023). A survey on the detection and impacts of deepfakes in visual, audio, and textual formats. *IEEE Access, 11*, 144497-144529. [CrossRef]

[21] Ahmed, N. U. R., Badshah, A., Adeel, H., Tajammul, A., Daud, A., & Alsahfi, T. (2024). Visual deepfake detection: Review of techniques, tools, limitations, and future prospects. *IEEE Access, 13*, 1923-1961. [CrossRef]

[22] Rana, M. S., Nobi, M. N., Murali, B., & Sung, A. H. (2022). Deepfake detection: A systematic literature review. *IEEE Access, 10*, 25494-25513. [CrossRef]

[23] Masood, M., Nawaz, M., Malik, K. M., Javed, A., Irtaza, A., & Malik, H. (2023). Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward: Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. *Applied intelligence, 53*(4), 3974-4026. [CrossRef]

[24] Heidari, A., Jafari Navimipour, N., Dag, H., & Unal, M. (2024). Deepfake detection using deep learning methods: A systematic and comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 14*(2), e1520. [CrossRef]

[25] Gupta, G., Raja, K., Gupta, M., Jan, T., Whiteside, S. T., & Prasad, M. (2023). A comprehensive review of deepfake detection using advanced machine learning and fusion methods. *Electronics, 13*(1), 95. [CrossRef]

[26] Kaur, A., Noori Hoshyar, A., Saikrishna, V., Firmin, S., & Xia, F. (2024). Deepfake video detection: challenges and opportunities. *Artificial Intelligence Review, 57*(6), 159. [CrossRef]

[27] Altuncu, E., Franqueira, V. N., & Li, S. (2024). Deepfake: definitions, performance metrics and standards, datasets, and a meta-review. *Frontiers in Big Data, 7*, 1400024. [CrossRef]

[28] Stroebel, L., Llewellyn, M., Hartley, T., Ip, T. S., & Ahmed, M. (2023). A systematic literature review on the effectiveness of deepfake detection techniques. *Journal of Cyber Security Technology, 7*(2), 83-113. [CrossRef]

[29] Verdoliva, L. (2020). Media forensics and deepfakes: an overview. *IEEE journal of selected topics in signal processing, 14*(5), 910-932. [CrossRef]

[30] Zhao, C., Wang, C., Hu, G., Chen, H., Liu, C., & Tang, J. (2023). ISTVT: interpretable spatial-temporal video transformer for deepfake detection. *IEEE Transactions on Information Forensics and Security, 18*, 1335-1348. [CrossRef]

[31] Luo, A., Kong, C., Huang, J., Hu, Y., Kang, X., & Kot, A. C. (2023). Beyond the prior forgery knowledge: Mining critical clues for general face forgery detection. *IEEE Transactions on Information Forensics and Security, 19*, 1168-1182. [CrossRef]

[32] Wang, Y., Yu, K., Chen, C., Hu, X., & Peng, S. (2023, June). Dynamic graph learning with content-guided spatial-frequency relation reasoning for deepfake detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 7278-7287). IEEE. [CrossRef]

[33] Miao, C., Tan, Z., Chu, Q., Liu, H., Hu, H., & Yu, N. (2023). F2trans: High-frequency fine-grained transformer for face forgery detection. *IEEE Transactions on Information Forensics and Security, 18*, 1039-1051. [CrossRef]

[34] Groh, M., Epstein, Z., Firestone, C., & Picard, R. (2022). Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences, 119*(1), e2110013119. [CrossRef]

[35] Coccomini, D. A., Caldelli, R., Falchi, F., & Gennaro, C. (2023). On the generalization of deep learning models in video deepfake detection. *Journal of Imaging, 9*(5), 89. [CrossRef]

[36] Chamot, F., Geradts, Z., & Haasdijk, E. (2022). Deepfake forensics: Cross-manipulation robustness of feedforward-and recurrent convolutional forgery detection methods. *Forensic Science International: Digital Investigation, 40*, 301374. [CrossRef]

[37] Abbasi, M., Váz, P., Silva, J., & Martins, P. (2025). Comprehensive evaluation of deepfake detection models: Accuracy, generalization, and resilience to adversarial attacks. *Applied Sciences, 15*(3), 1225. [CrossRef]

[38] Mahmud, B. U., & Sharmin, A. (2021). Deep insights of deepfake technology: A review. *arXiv preprint arXiv:2105.00192*. [CrossRef]

[39] Alanazi, S., Asif, S., & Moulitsas, I. (2024). Examining the societal impact and legislative requirements of deepfake technology: A comprehensive study. *International Journal of Social Science and Humanity, 14*(2), 59–65. [CrossRef]

[40] Yan, Z., Yao, T., Chen, S., Zhao, Y., Fu, X., Zhu, J., ... & Yuan, L. (2024). Df40: Toward next-generation deepfake detection. *Advances in Neural Information Processing Systems, 37*, 29387-29434.

[41] Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review, 9*(11), 39-52. [CrossRef]

[42] Meneses, J. P. (2021). Deepfakes and the 2020 US elections: What (did not) happen. *arXiv preprint arXiv:2101.09092*. [CrossRef]

[43] Bateman, J. (2022). *Deepfakes and synthetic media in the financial system: Assessing threat scenarios*. Carnegie Endowment for International Peace. https://assets.carnegieendowment.org/static/files/Bateman_FinCyber_Deepfakes_final.pdf

[44] Pei, G., Zhang, J., Hu, M., Zhang, Z., Wang, C., Wu, Y., ... & Tao, D. (2026). Deepfake generation and detection: A benchmark and survey. *ACM Computing Surveys, 58*(11), 1-41. [CrossRef]

[45] Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social media+ society, 6*(1), 2056305120903408. [CrossRef]

[46] Tan, C., Zhao, Y., Wei, S., Gu, G., Liu, P., & Wei, Y. (2024, March). Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 5, pp. 5052-5060). [CrossRef]

[47] Chakraborty, R., & Naskar, R. (2024). Role of human physiology and facial biomechanics towards building robust deepfake detectors: A comprehensive survey and analysis. *Computer Science Review, 54*, 100677. [CrossRef]

[48] Ciftci, U. A., Demir, I., & Yin, L. (2020). Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE transactions on pattern analysis and machine intelligence*. [CrossRef]

[49] Mansoor, N., & Iliev, A. I. (2025). Explainable ai for deepfake detection. *Applied Sciences, 15*(2), 725. [CrossRef]

[50] Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., & Guo, B. (2020, June). Face x-ray for more general face forgery detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 5000-5009). IEEE. [CrossRef]

[51] Shao, R., Wu, T., & Liu, Z. (2022, October). Detecting and recovering sequential deepfake manipulation. In *European Conference on Computer Vision* (pp. 712-728). Cham: Springer Nature Switzerland. [CrossRef]

[52] Yang, W., Zhou, X., Chen, Z., Guo, B., Ba, Z., Xia, Z., Cao, X., & Ren, K. (2023). AVoiD-DF: Audio-visual joint learning for detecting deepfake. *IEEE Transactions on Information Forensics and Security, 18*, 2015-2029. [CrossRef]

[53] Yoon, J., Panizo-LLedot, A., Camacho, D., & Choi, C. (2024). Triple-modality interaction for deepfake detection on zero-shot identity. *Information Fusion, 109*, 102424. [CrossRef]

[54] Matern, F., Riess, C., & Stamminger, M. (2019, January). Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)* (pp. 83-92). IEEE. [CrossRef]

[55] Lewis, J. K., Toubal, I. E., Chen, H., Sandesera, V., Lomnitz, M., Hampel-Arias, Z., ... & Palaniappan, K. (2020, October). Deepfake video detection based on spatial, spectral, and temporal inconsistencies using multimodal deep learning. In *2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)* (pp. 1-9). IEEE. [CrossRef]

[56] Luo, X., & Wang, Y. (2025). Frequency-domain masking and spatial interaction for generalizable deepfake detection. *Electronics, 14*(7), 1302. [CrossRef]

[57] Ciftci, U. A., Demir, I., & Yin, L. (2020, September). How do the hearts of deep fakes beat? Deep fake source detection via interpreting residuals with biological signals. In *2020 IEEE international joint conference on biometrics (IJCB)* (pp. 1-10). IEEE. [CrossRef]

[58] Zi, B., Chang, M., Chen, J., Ma, X., & Jiang, Y. G. (2020). WildDeepfake: A challenging real-world dataset for deepfake detection. In *Proceedings of the 28th ACM International Conference on Multimedia* (pp. 2382-2390). [CrossRef]

[59] Karathanasis, A., Violos, J., & Kompatsiaris, I. (2025). A comparative analysis of compression and transfer learning techniques in deepfake detection models. *Mathematics, 13*(5), 887. [CrossRef]

[60] Gazis, A., Karypidis, E., Santamouri, K., Vavouras, T., Mastorakis, N. E., & Pappas, S. (2026). Deepfakes and Synthetic Media: Generation, Detection, and Governance. *Encyclopedia, 6*(8), 165. [CrossRef]

[61] Byman, D. L., Gao, C., Meserole, C., & Subrahmanian, V. S. (2023). *Deepfakes and international conflict* (Vol. 8). Washington, DC: Brookings Institution. https://www.brookings.edu/wp-content/uploads/2023/01/FP_20230105_deepfakes_international_conflict.pdf

[62] Michałkiewicz-Kądziela, E. (2024). The impact of deepfakes on elections and methods of combating disinformation in the virtual world. *Teka Komisji Prawniczej PAN Oddział w Lublinie, 17*(1), 151-161. [CrossRef]

[63] Bhandari, R., & Bhandari, S. (2025). Artificial intelligence: understanding deepfakes. *EDPACS, 70*(9), 21-31. [CrossRef]

[64] Ölvecký, M., Huraj, L., & Brlej, I. (2025). Evaluating the effectiveness of deepfake video detection tools: A comparative study. *TEM Journal, 14*(1), 64. [CrossRef]

[65] Nightingale, S. J., & Farid, H. (2022). AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences, 119*(8), e2120481119. [CrossRef]

[66] Köbis, N. C., Doležalová, B., & Soraperra, I. (2021). Fooled twice: People cannot detect deepfakes but think they can. *iScience, 24*(11), 103364. [CrossRef]

[67] Groh, M., Sankaranarayanan, A., Singh, N., Kim, D. Y., Lippman, A., & Picard, R. (2024). Human detection of political speech deepfakes across transcripts, audio, and video. *Nature communications, 15*(1), 7629. [CrossRef]

[68] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020, June). Analyzing and improving the image quality of stylegan. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 8107-8116). IEEE. [CrossRef]

[69] Shaoanlu. (2022). *faceswap-GAN: A denoising autoencoder + adversarial losses and attention mechanisms for face swapping* (Version 2.2) [Computer software]. GitHub. Retrieved from https://github.com/shaoanlu/faceswap-GAN

[70] Chen, R., Chen, X., Ni, B., & Ge, Y. (2020, October). Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 2003-2011). [CrossRef]

[71] Xu, Z., Hong, Z., Ding, C., Zhu, Z., Han, J., Liu, J., & Ding, E. (2022, June). Mobilefaceswap: A lightweight framework for video face swapping. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 3, pp. 2973-2981). [CrossRef]

[72] Dhanyalakshmi, R., Popirlan, C. I., & Hemanth, D. J. (2024). A survey on deep learning-based reenactment methods for deepfake applications. *IET Image Processing, 18*(14), 4433–4460. [CrossRef]

[73] Waseem, S., Bakar, S. A. R. S. A., Ahmed, B. A., Omar, Z., Eisa, T. A. E., & Dalam, M. E. E. (2023). DeepFake on face and expression swap: A review. *IEEE Access, 11*, 117865-117906. [CrossRef]

[74] Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., & Nießner, M. (2016). Face2Face: Real-time face capture and reenactment of RGB videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2387-2395). [CrossRef]

[75] Suratkar, S., & Sharma, P. (2022). A simple and effective way to detect DeepFakes: using 2D and 3D CNN. In *Security, Privacy and Data Analytics: Select Proceedings of ISPDA 2021* (pp. 227-238). Singapore: Springer Singapore. [CrossRef]

[76] Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Niessner, M., ... & Theobalt, C. (2018). Deep video portraits. *ACM transactions on graphics (TOG), 37*(4), 1-14. [CrossRef]

[77] Boháček, M., & Farid, H. (2022). Protecting world leaders against deep fakes using facial, gestural, and vocal mannerisms. *Proceedings of the National Academy of Sciences, 119*(48), e2216035119. [CrossRef]

[78] Agarwal, M., Mukhopadhyay, R., Namboodiri, V. P., & Jawahar, C. V. (2023). Audio-visual face reenactment. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 5178-5187). [CrossRef]

[79] Shiohara, K., & Yamasaki, T. (2022, June). Detecting deepfakes with self-blended images. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 18699-18708). IEEE. [CrossRef]

[80] Lin, H., Huang, W., Luo, W., & Lu, W. (2023). DeepFake detection with multi-scale convolution and vision transformer. *Digital Signal Processing, 134*, 103895. [CrossRef]

[81] Zhao, Y., Liu, B., Ding, M., Liu, B., Zhu, T., & Yu, X. (2023, January). Proactive deepfake defence via identity watermarking. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (pp. 4591-4600). IEEE. [CrossRef]

[82] Amerini, I., Barni, M., Battiato, S., Bestagini, P., Boato, G., Bruni, V., ... & Vitulano, D. (2025). Deepfake media forensics: Status and future challenges. *Journal of Imaging, 11*(3), 73. [CrossRef]

[83] Dagar, D., & Vishwakarma, D. K. (2022). A literature review and perspectives in deepfakes: generation, detection, and applications. *International journal of multimedia information retrieval, 11*(3), 219-289. [CrossRef]

[84] Cao, Z., Simon, T., Wei, S. E., & Sheikh, Y. (2017, July). Realtime multi-person 2d pose estimation using part affinity fields. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 7291-7299). IEEE. [CrossRef]

[85] Güler, R. A., Neverova, N., & Kokkinos, I. (2018). Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7297-7306). [CrossRef]

[86] Liu, B., Liu, B., Zhu, T., & Ding, M. (2025). A review of deepfake and its detection: From generative adversarial networks to diffusion models. *International Journal of Intelligent Systems, 2025*(1), 9987535. [CrossRef]

[87] Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2021). Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM, 65*(1), 99-106. [CrossRef]

[88] Peng, C., Xu, T., Liu, D., Wang, N., & Gao, X. (2025). Within 3DMM space: Exploring inherent 3D artifact for video forgery detection. *IEEE Transactions on Information Forensics and Security, 20*, 7954-7965. [CrossRef]

[89] Prajwal, K. R., Mukhopadhyay, R., Namboodiri, V. P.,

& Jawahar, C. V. (2020, October). A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 484-492). [CrossRef]

[90] Chung, J. S., & Zisserman, A. (2016, November). Out of time: automated lip sync in the wild. In *Asian conference on computer vision* (pp. 251-263). Cham: Springer International Publishing. [CrossRef]

[91] Kaushal, A., Kumar, S., & Kumar, R. (2024). A review on deepfake generation and detection: bibliometric analysis. *Multimedia Tools and Applications, 83*(40), 87579-87619. [CrossRef]

[92] Fernando, T., Priyasad, D., Sridharan, S., Ross, A., & Fookes, C. (2025). Face Deepfakes–A Comprehensive Review. *arXiv preprint arXiv:2502.09812*. [CrossRef]

[93] Wang, T. (2025). A Comprehensive Review of Image-to-Image Translation, Face Manipulation, and Neural Style Transfer Methods. [CrossRef]

[94] Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information fusion, 64*, 131-148. [CrossRef]

[95] Sghir, A. A., Amraoui, S. E., Elafi, I., Zrira, N., & Benmiloud, I. (2023, December). StyleGan for Deepfake Generation: A Comparative Study. In *International Conference on Advanced Technologies for Humanity* (pp. 103-110). Cham: Springer Nature Switzerland. [CrossRef]

[96] Cheng, X. (2024). Refining CycleGAN with attention mechanisms and age-Aware training for realistic Deepfakes. *Heliyon, 10*(16). [CrossRef]

[97] Choi, Y., Choi, M., Kim, M., Ha, J. W., Kim, S., & Choo, J. (2018, June). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *2018 IEEE/CVF conference on computer vision and pattern recognition* (pp. 8789-8797). IEEE. [CrossRef]

[98] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training gans. *Advances in neural information processing systems, 29*.

[99] Durall, R., Chatzimichailidis, A., Labus, P., & Keuper, J. (2020). Combating mode collapse in gan training: An empirical analysis using hessian eigenvalues. *arXiv preprint arXiv:2012.09673*. [CrossRef]

[100] Ghojogh, B., Ghodsi, A., Karray, F., & Crowley, M. (2021). Generative adversarial networks and adversarial autoencoders: Tutorial and survey. *arXiv preprint arXiv:2111.13282*. [CrossRef]

[101] Tewari, A., Fried, O., Thies, J., Sitzmann, V., Lombardi, S., Sunkavalli, K., ... & Zollhöfer, M. (2020, May). State of the art on neural rendering. In *Computer graphics forum* (Vol. 39, No. 2, pp. 701-727). [CrossRef]

[102] Thies, J., Zollhöfer, M., & Nießner, M. (2019). Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG), 38*(4), 1-12. [CrossRef]

[103] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems, 33*, 6840-6851.

[104] Song, J., Meng, C., & Ermon, S. (2020). Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*. [CrossRef]

[105] Sharma, S., & Singh, S. (2024, April). Spear or Shield: Mastering the Art of Gen-AI in Face Recognition. In *International Conference on Advanced Network Technologies and Computational Intelligence* (pp. 392-405). Cham: Springer Nature Switzerland. [CrossRef]

[106] Hussain, S., Neekhara, P., Jere, M., Koushanfar, F., & McAuley, J. (2021, January). Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples. In *2021 IEEE winter conference on applications of computer vision (WACV)* (pp. 3347-3356). IEEE. [CrossRef]

[107] Wang, T. C., Liu, M. Y., Zhu, J. Y., Tao, A., Kautz, J., & Catanzaro, B. (2018, June). High-resolution image synthesis and semantic manipulation with conditional gans. In *2018 IEEE/CVF conference on computer vision and pattern recognition* (pp. 8798-8807). Ieee. [CrossRef]

[108] Johnson, J., Alahi, A., & Fei-Fei, L. (2016, September). Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision* (pp. 694-711). Cham: Springer International Publishing. [CrossRef]

[109] Sohail, S., Sajjad, S. M., Zafar, A., Iqbal, Z., Muhammad, Z., & Kazim, M. (2025). Deepfake image forensics for privacy protection and authenticity using deep learning. *Information, 16*(4), 270. [CrossRef]

[110] Mmaduekwe, E. (2025). AI-driven detection and prevention of deepfakes in National security. *Asian Journal of Research in Computer Science, 18*(6), 361-368. [CrossRef]

[111] Jeong, Y., Kim, D., Ro, Y., & Choi, J. (2022, June). Frepgan: robust deepfake detection using frequency-level perturbations. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 36, No. 1, pp. 1060-1068). [CrossRef]

[112] Tsigos, K., Apostolidis, E., & Mezaris, V. (2025, February). Improving the perturbation-based explanation of deepfake detectors through the use of adversarially-generated samples. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)* (pp. 658-667). IEEE. [CrossRef]

[113] Hussain, S., Neekhara, P., Dolhansky, B., Bitton, J., Canton Ferrer, C., McAuley, J., & Koushanfar, F. (2022). Exposing vulnerabilities of deepfake

detection systems with robust attacks. *Digital Threats: Research and Practice, 3*(3), 1-23. [CrossRef]

[114] Dutta, H., Pandey, A., & Bilgaiyan, S. (2021). EnsembleDet: ensembling against adversarial attack on deepfake detection. *Journal of Electronic Imaging, 30*(6), 063030-063030. [CrossRef]

[115] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1-11).

[116] Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020, June). Celeb-df: A large-scale challenging dataset for deepfake forensics. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3204-3213). IEEE. [CrossRef]

[117] Usman, M. T., Khan, H., Singh, S. K., Lee, M. Y., & Koo, J. (2024). Efficient deepfake detection via layer-frozen assisted dual attention network for consumer imaging devices. *IEEE Transactions on Consumer Electronics, 71*(1), 281-291. [CrossRef]

[118] Serrano, A., Umlil, E., & Thomas, R. (2026). Deepfake detectors are DUMB: A benchmark to assess adversarial training robustness under transferability constraints. *arXiv preprint arXiv:2601.05986*. [CrossRef]

[119] Cao, Z., Wang, Z., Wang, R., Ai, H., & Wu, G. (2026). Beyond Global Perturbations: Focused Adversarial Attacks for Face Deepfake. *IEEE Transactions on Biometrics, Behavior, and Identity Science, 8*(4), 518-528. [CrossRef]

[120] Khan, A. A., Laghari, A. A., Inam, S. A., Ullah, S., Shahzad, M., & Syed, D. (2025). A survey on multimedia-enabled deepfake detection: state-of-the-art tools and techniques, emerging trends, current challenges & limitations, and future directions. *Discover Computing, 28*(1), 48. [CrossRef]

[121] Babaei, R., Cheng, S., Duan, R., & Zhao, S. (2025). Generative artificial intelligence and the evolving challenge of deepfake detection: A systematic analysis. *Journal of Sensor and Actuator Networks, 14*(1), 17. [CrossRef]

[122] Rettinger, M., Beaumont, B., Le-Khac, N. A., & Nguyen-Le, H. H. (2026). How Effective Are Publicly Accessible Deepfake Detection Tools? A Comparative Evaluation of Open-Source and Free-to-Use Platforms. *arXiv preprint arXiv:2603.04456*. [CrossRef]

[123] Juefei-Xu, F., Wang, R., Huang, Y., Guo, Q., Ma, L., & Liu, Y. (2022). Countering malicious deepfakes: Survey, battleground, and horizon. *International Journal of Computer Vision, 130*(7), 1678-1734. [CrossRef]

[124] Nirkin, Y., Wolf, L., Keller, Y., & Hassner, T. (2022). DeepFake detection based on discrepancies between faces and their context. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(10), 6111-6121. [CrossRef]

[125] Prabhu, O., Naik, S. S., Prabhu, S., & Chadaga, K. (2026). Countering Synthetic Realities: Advances, Challenges, and Future Directions in Deepfake Image Detection. *Systems and Soft Computing*, 200476. [CrossRef]

[126] Bukhari, M., Maqsood, M., Sattar, A., & Lee, M. Y. (2026). A texture-guided adversarial defense framework against deepfake generation. *IEEE Access, 14*, 45017-45037. [CrossRef]

[127] Sun, B., Han, D., Yang, G., Yang, X., & Wang, X. (2026). Deepfake detection via domain adversarial learning with strong-weak augmentation strategies. *Journal of Visual Communication and Image Representation*, 104761. [CrossRef]

[128] Abdel-Wahab, A., & Alkhatib, M. (2025). Toward Robust Deepfake Defense: A Review of Deepfake Detection and Prevention Techniques in Images. *Computers, Materials and Continua, 86*(2), 1-34. [CrossRef]

[129] Perelli, G., Micheletto, M., Concas, S., Puglisi, G., & Marcialis, G. L. (2026). Robust deepfake detection in compressed videos with scalable network strategies. *Expert Systems with Applications*, 131761. [CrossRef]

[130] Bunzel, N., Frick, R. A., Graner, L., Göller, N., & Steinebach, M. (2026). Statistical Feature-Based Detection of Adversarial Noise and Patch Attacks in Image and Deepfake Analysis. In *Adversarial Example Detection and Mitigation Using Machine Learning* (pp. 195-208). Cham: Springer Nature Switzerland. [CrossRef]

[131] Zhai, T., Lu, K., Li, J., Wang, Y., Zhang, W., Yu, P., & Xia, Z. (2024). Learning spatial-frequency interaction for generalizable deepfake detection. *IET Image Processing, 18*, 4666-4679. [CrossRef]

[132] Qureshi, S. M., Saeed, A., Almotiri, S. H., Ahmad, F., & Al Ghamdi, M. A. (2024). Deepfake forensics: a survey of digital forensic methods for multimodal deepfake identification on social media. *PeerJ Computer Science, 10*, e2037. [CrossRef]

[133] Sajeev, A., Hiremath, M., & Pohrmen, F. H. (2026, January). Defensive Mechanisms Against Deepfake Threats in Cybersecurity: Detection, Robustness, and Simulation-Based Evaluation. In *2026 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE)* (pp. 1-6). IEEE. [CrossRef]

[134] Cai, L., Wang, H., Ji, J., Zhoumen, Y., Chen, S., Yao, T., & Sun, X. (2026, March). Zooming in on fakes: A novel dataset for localized AI-generated image detection with forgery amplification approach. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 40, No. 4, pp. 2534-2542). [CrossRef]

[135] Jia, L., Sun, H., Guo, Z., Diao, Y., Ma, D., & Yang, G. (2026, March). Uncovering and mitigating destructive multi-embedding attacks in deepfake

proactive forensics. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 40, No. 1, pp. 471-479). [CrossRef]

[136] Gupta, S., Hariprasad, Y., Iyengar, S. S., Gurappa, S., & Mohanty, P. (2026). Enhancing digital security: A novel dual-paradigm approach for robust deepfake detection using pre and post quantum-trained neural networks. *Digital Threats: Research and Practice*, *7*(1), 1-15. [CrossRef]

[137] Tasnim, N., Uddin, K., & Malik, K. (2026). A comprehensive survey, large-scale empirical study, and future insights on generalization, robustness, and explainability of AI-generated image detection. *Available at SSRN*. [CrossRef]

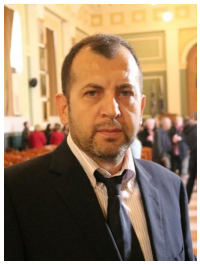
[138] Khan, N., Nguyen, T., Bermak, A., & Khalil, I. (2026). CAMME: Cross-Domain Adaptive Multi-Modal Embeddings with Attention Mechanisms for Adversarially Robust Deepfake Detection. *Available at SSRN 6067189*. [CrossRef]

[139] Zhang, L., Xiong, P., & Wang, J. (2026). Privacy-preserving and zero-shot detection strategies for multimodal AI-generated content. *Neurocomputing*, 133229. [CrossRef]

[140] Mitra, A., Mohanty, S. P., & Kougianos, E. (2026). Deepfakes and Large Language Models: Risks, Defenses, and the Future of Generative AI. [CrossRef]

Alexandros Gazis is a software engineer and researcher with extensive experience in distributed systems, game-based learning, artificial intelligence, middleware architectures, wireless sensor network (WSN) instrumentation, electrical load forecasting, and intelligent data processing. His research focuses on energy systems and applications, electrical and electronic systems, cloud and IoT middleware, context-aware systems, serious games, AI-driven optimization, and resilient and intelligent infrastructures. His work also extends to military and mission-critical applications, including military-resilient critical infrastructure and smart energy systems. He has contributed to several international research projects and peer-reviewed publications, combining applied engineering with academic research. His interests also include AI-driven system optimization, smart environments, and intelligent energy management. (Email: agazis@teemail.gr)

Stylianos Pappas is a researcher and academic contributor with expertise in computer science and information systems. His research interests include artificial intelligence applications, data analytics, intelligent software systems, high-voltage transmission, electric load forecasting (both long- and short-term), wind speed prediction and electrical insulation materials. He has participated in interdisciplinary research activities and has co-authored scientific publications in international venues. His work emphasizes practical implementations of emerging digital technologies. (Email: steliospappas@teemail.gr)



Theodoros Vavouras is a researcher and educator specializing in educational AI, digital technologies, digital humanities, game-based learning, serious games, distance learning education, computer science, and applied informatics. His academic interests include artificial intelligence, educational technologies, and intelligent information systems. He has been involved in research projects and publications focusing on the integration of emerging technologies in learning and professional environments. (Email: vavouras.theodoros@ac.eap.gr, vavouras@itl.auth.gr)

Asim Ali is an aspiring Artificial Intelligence researcher based in Peshawar, Pakistan. He completed his intermediate studies from Leeds College, Peshawar, and is currently pursuing a BS degree in Artificial Intelligence at Cecoss University, Hayatabad, Peshawar. His academic and research interests span artificial intelligence, deep learning, and machine learning, with a strong focus on visual intelligence. He is particularly interested in solving computer vision problems, including image segmentation, object detection, localization, and visual recognition. Through his studies and ongoing research, Asim aims to develop practical AI solutions that can understand and interpret visual data for real-world applications. (Email: asimbinyousaf@gmail.com)

Nikos Mastorakis is a Greek engineer, mathematician, and academic from Sitia, Crete. He is a Full Professor at the Technical University of Sofia, Bulgaria, and at the Naval Academy. He also holds honorary professorships at the University of Budapest; the Budapest University of Technology and Economics; the Technical University of Cluj, Romania; and the University of Salerno, Italy. In addition, he has served as a Visiting Professor at the University of California, Berkeley (USA), and the University of Exeter (UK). He is widely known for his contributions to Systems Theory, particularly in Multidimensional Systems, Mathematical Modeling, and Computational Mathematics. His research on the factorization of multivariate polynomials brought him international recognition, as he was among the first to address this previously unsolved problem. His work also includes the design of multidimensional filters, the analysis of stability and stability margins in multidimensional systems, and various problems in signal processing. He has been honored by the Romanian Academy of Sciences. He is considered one of the most prolific Greek authors and among the most prolific researchers worldwide in terms of number of publications. (Email: mastor@hna.gr, mastor@tu-sofia.bg)