Review

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

Alexandros Gazis * , Stylianos Pappas , Theodoros Vavouras , Asim Ali , Nikos E. Mastorakis

Posted Date: 23 March 2026

doi: 10.20944/preprints202603.1662.v1

Keywords: deepfake detection; computer vision; generative adversarial networks; adversarial deep-fakes; physiological signal-based detection; military information security; military and hybrid threats; privacy-preserving deepfake detection

 doi:10.20944/preprints202603.1662.v1

Review

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 [3,4], Asim Ali [5] and Nikos E. Mastorakis [2,6]

[1] Department of Electrical and Computer Engineering, Democritus University of Thrace, Xanthi, 67100, Greece
[2] Electrical Engineering and Computer Science, Hellenic Naval Academy, Terma Chatzikyriakou, 18539, Piraeus, Greece
[3] School of Italian Language and Literature, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece
[4] School of Humanities, Hellenic Open University, 26335 Patras, Greece
[5] Department of Artificial Intelligence, Cecoss University, Hayatabad, Peshawar 25000, Pakistan
[6] English Language Faculty of Engineering, Technical University of Sofia, 1756, Sofia, Bulgaria
* Correspondence: agazis@ee.duth.gr
† Member of Technical Chamber of Greece (TEE – TCG)

Abstract

Deepfake technology has advanced swiftly, assisting the development of new technology trends such as the rapid production of hyper-realistic synthetic media that present considerable threats to digital security, privacy, military operations, and information integrity. This paper offers an extensive examination of visual intelligence and computer vision methodologies for deepfake detection, encompassing recent developments in deep learning, adversarial strategies, and feature extraction techniques. Specifically, it examines prevalent deepfake generation architectures, such as GANs, autoencoders, neural rendering, and diffusion models, in conjunction with novel adversarial strategies aimed at improving the realism of synthetic media while circumventing detection, particularly in military and intelligence-driven scenarios. Additionally, we investigate visual artifacts and manipulation traces, scrutinizing physical discrepancies, digital fingerprints, and physiological signals that function as critical indications for detection models. As such, the article offers a comprehensive analysis of CNN-based, transformer-based, and frequency-domain deepfake detection methodologies, highlighting their advantages, drawbacks, and practical relevance. Furthermore, we examine assessment measures, and the generalization difficulties encountered by detection methods and emphasize prospective research avenues, including explainable AI, self-supervised learning, and federated learning. This study is 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; computer vision; generative adversarial networks; adversarial deepfakes; physiological signal-based detection; military information security; military and hybrid threats; privacy-preserving deepfake detection

1. Introduction

Over the last five years, there has been tremendous discussion around Artificial Intelligence (AI) and the broader concept of the upcoming Fourth Industrial Revolution, [1]. In collaboration with the Internet of Things (IoT) and Big Data, this revolution is expected to transform both our everyday lives and the industrial and military landscape, [2]. The term "Fourth Industrial Revolution" refers to the emerging wave of technological advancement conceived in the 21st century, which

 doi:10.20944/preprints202603.1662.v1

focuses on the integration of digital, biological, and physical technologies to usher in a new era of innovation, [3]. This is the era we currently live in, and what lies ahead is Industry 5.0, the next phase of industrialization. Industry 5.0 introduces new and emerging concepts where humans, machines, and technological entities with memory (feedback) and action mechanisms (actors) will collaborate to advance technology in industrial and military ecosystems. This fusion brings together human initiative and creativity with the processing power and capabilities of AI, [4]. As a result, the ever-increasing number of interconnected devices, whether it be a new laptop, a smartwatch, or most notably, a smartphone, is continuously generating data points. These data points are not only growing in number but also contributing to what is known as Big Data, characterized by high volume, velocity, and variety. Furthermore, the global network formed by these interconnected devices, interacting within a shared environment, is referred to as the Internet of Things (IoT), [5]. The IoT enables seamless communication between devices, including military sensing, surveillance, and command systems, acting as the foundational infrastructure for machines that generate data and enable interactions. The term "Big Data" and the so called "Three Vs" (Volume, Velocity, and Variety) are intrinsically linked to the processing methods required to handle such complex datasets in military and high-security environments methods which often rely on deep learning techniques or, more broadly, Artificial Intelligence (AI), [6]. In this paper, we do not delve into Artificial Intelligence (AI) in general, but rather focus on one of its most recent and controversial developments: deepfakes. 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, [7]. 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, [8]. There are also applications of deepfake technology in educational contexts, [9], such as simulating famous scientists or historical figures, or enhancing interactive learning through realistic virtual personas and dialogues, [10]. However, the term is more commonly associated with negative connotations, including misinformation, [11], identity theft, [12], reputational harm, [13], and, in its simplest form, satirical or humorous misuse, [14]. Modern AI techniques now produce hyper-realistic media, making it increasingly difficult to detect n military intelligence, industrial applications or in general and defense analysis and complex system 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, [15–17]. 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, [18]. 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, [19]. A recent generative image modeling framework jointly models low-level VAE latents and VFM features for improved generative quality and training convergence speed, [20]. 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, [21]. The primary technical challenges in building convincing deepfakes are the availability of

 doi:10.20944/preprints202603.1662.v1

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, [22]. As a result, deepfakes are becoming increasingly sophisticated, making it harder, even for digital analysts and observers, to distinguish authentic content from fabricated material, [23,24]. 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, [25,26]. 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, [27,28]. 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, [29]. Table 1 summarizes key review studies that form the foundational literature for deepfake detection in visual media.

Table 1. Selected survey papers on audio deepfake detection, focusing on speech synthesis, adversarial audio attacks, and robustness evaluation.

No.	Title	Subject / IT domain	Year	Doi
1.	Deepfake Detection Using Deep Learning Methods	Audio Forensics, Speech Synthesis	2023	[30]
2.	A Comprehensive Review of DeepFake Detection Using Advanced ML and Fusion	Speech AI, Multimodal Detection	2023	[31]
3.	Deepfake Video Detection: Challenges and Opportunities	Audio-visual Fusion, Deep Learning	2024	[32]
4.	Deepfake: Definitions, Metrics, Datasets, and Methods	Speech Processing, Detection Standards	2024	[33]
5.	Effectiveness of Deepfake Detection Methods	Digital Forensics, AI Evaluation	2023	[34]
6.	Effect of Deep Learning on Audio Deepfake Detection	Forensic AI, Neural Networks	2023	[35]
7.	Modern Audio Deepfake Detection Methods: Challenges and Directions	Speech AI, Adversarial Learning	2022	[36]
8.	Deepfake Audio Detection in Group Conversations	Real-time Detection, Signal Processing	2020	[37]
9.	Does Audio Deepfake Detection Generalize?	Generalization, Robustness Testing	2022	[38]
10.	Audio Deepfake Detection: Adversarial Attacks and Countermeasures	AI Security, Adversarial Defense	2024	[39]

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, which typically occur in less than 10–20% of flagged content, require human intervention for final validation and interpretation, [40]. 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, [41]. 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, [42,43]. 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, [33]. This paper provides a brief but comprehensive survey and

 doi:10.20944/preprints202603.1662.v1

explanation of the state of the art regarding visual intelligence and computer vision techniques used for deepfake detection. The following sections present a general overview of deepfake generation architectures, including GANs, autoencoders, neural rendering techniques, and recent advances in diffusion models. We then explore adversarial strategies used to enhance the realism of synthetic content and the resulting challenges for detection systems. A detailed analysis is provided of visual artifacts and manipulation traces, including facial texture distortions, compression patterns, physiological signals, and image inconsistencies, all of which serve as valuable indicators for identifying tampered media. The study then presents deep learning–based detection methods through three main groups: CNN-based models, transformer-based approaches, and frequency-domain techniques, and uses this structure to discuss their performance, robustness, and limitations. We also review benchmark datasets, commonly used evaluation metrics, and the major challenges faced when attempting to generalize across diverse types of deepfake content. Finally, we outline key research directions, such as explainable AI and self-supervised learning, that are likely to shape future advances in visual deepfake detection. 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.

2. Background and Motivation

Firstly, 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, [44]. 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, [45]. These fabricated media files challenge our ability to distinguish real from fake and erode digital trust across multiple domains such as civilian, governmental, and military, i.e. from political misinformation to military deception and digital fraud, identity theft, and public safety, [46,47]. For instance, in the United States, several deepfake videos of former President Biden have circulated on social media, urging voters to stay home during recent elections, [48]. In Europe, synthetic voice imitations have been used to impersonate well-known CEOs and authorize fraudulent wire transfers, [49]. Similarly, in Arab nations, manipulated videos of public figures have emerged, inciting political tensions and spreading false, often religious, statements, [50]. 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, [31]. 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, [51]. Additionally, physiological signal analysis is an emerging field that involves studying involuntary human

 doi:10.20944/preprints202603.1662.v1

signals such as heartbeat-induced skin tone changes, facial micro-expressions, and gaze dynamics, [52]. 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. 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, [53]. 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.

3. Definition and Characteristics of Deepfakes

As such, 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, [54]. They can be automatically generated from diverse input sources, namely, datasets, without requiring manual animation or scripting. 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, the most commonly recognized form of deepfake by the public is facial manipulation targeting military personnel or officials, [55]. 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 [56]. 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, [57], enables the creation of real-time deepfakes. 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 breifings. 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:

 doi:10.20944/preprints202603.1662.v1

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, [58].

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, [59].

3. **Frequency domain anomalies**: By applying techniques like the Discrete Cosine Transform (DCT), analysts can detect unnatural signal patterns introduced during the synthesis process that are not visible in the spatial domain, [60].

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, [61].

5. **Anatomical irregularities**: Distorted or misaligned features, such as fingers, ears, or limbs, are often visible in deepfakes. These physical inaccuracies are common when body parts are rendered in motion or at unusual angles, [62].

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, [63].

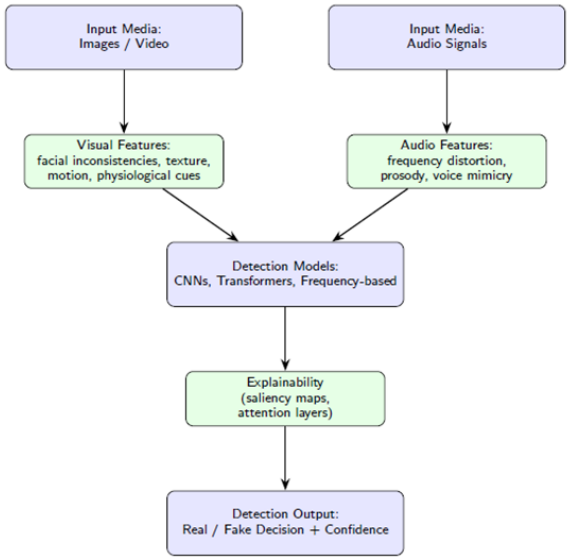

Figure 1. Conceptual framework of visual intelligence-based deepfake detection systems based on recent bibliography.

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, [64]. 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 by hyper-realistic synthetic data is its weaponization, particularly in the

 doi:10.20944/preprints202603.1662.v1

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, [65,66].

Additionally, there is a growing concern about the erosion of accountability, [67]. 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, [68].

Another critical issue is the erosion of public trust, [69,70]. 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 create a cultural bias, [71], where misuse is the first thing that comes to mind, undermining user confidence in otherwise legitimate and beneficial applications. Table 2 provides a curated overview of recent survey papers in the field of audio deepfake detection, reflecting advances in speech synthesis, adversarial audio attacks, and voice-based manipulation forensics.

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.

Table 2. Selected studies on Audio and Video Deepfake Analysis.

No.	Subject / IT domain	Year	Doi
1.	Speech AI, Urdu Audio Analysis	2025	[72]
2.	Audio Deepfake Analysis, Neural Networks	2025	[73]
3.	Multimedia AI, Detection Tools	2024	[74]
4.	Generative Models, Biometrics	2025	[75]
5.	Video Deepfake, Systematic Review	2025	[76]
6.	Deepfake Survey, Detection Techniques	2025	[77]
7.	Robustness Testing, Audio Forensics	2025	[78]
8.	AI Security, Adversarial Defense	2025	[79]

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, [80], and FaceSwap GAN, [81]. More recent

 doi:10.20944/preprints202603.1662.v1

face-swapping approaches incorporate attention mechanisms to preserve motion constraints and improve realism, [82]. 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, [83]. Another important technique is face reenactment, [84], 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, [85]. A widely known system in this area is Face2Face, [86,87]. Other advanced approaches employ 3D Convolutional Neural Networks (3D-CNNs), [88], or mesh-based deep learning frameworks, [89], to model facial dynamics, often using facial action units or dense optical flow across video frames, [90]. Applications of this technology include film dubbing, [91], virtual avatars, [92], and the digital resurrection of historical figures, [93,94], as well as more concerning uses such as journalistic manipulation, [17], or forensic falsification, [95]. A third category is full body synthesis, [96], where the aim is to 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, [97], and DensePose, [98], 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, [99]. Recent innovations include Neural Radiance Fields (NeRF), [100], and volumetric human models, which allow for 3D-aware synthesis of full-body sequences, [101]. Finally, audiovisual synthesis is a more traditional deepfake generation method that focuses on synchronizing synthetic speech with matching facial and lip movements, [102]. 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, [102], and SyncNet, [103], 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. The primary technologies that underpin deepfake synthesis include Generative Adversarial Networks (GANs), [104], autoencoder-based architectures, [105], and neural rendering systems, [106], with diffusion models emerging as one of the most promising recent developments.

Generative Adversarial Networks (GANs) are among the most widely used and studied methods in deepfake creation, [107]. A typical GAN consists of two competing neural networks: a generator, which produces synthetic data, and a discriminator, which attempts to distinguish between real and fake data. Through adversarial training, the generator learns to produce increasingly realistic outputs by attempting to deceive the discriminator. This mechanism has proven highly effective for tasks such as face generation, style transfer, and domain adaptation. Variants like StyleGAN, [108], CycleGAN, [109], and StarGAN, [110], allow for enhanced control over facial attributes, expressions, lighting conditions, and pose. These models can also be trained with conditional inputs to enable synchronized image or video generation. However, GANs are known to suffer from training instabilities, [111], and issues such as mode collapse, [112], especially when producing high-resolution outputs, making them less reliable in some applications. Another important architectural category is autoencoder based methods, [29], which serve as a cornerstone of many deepfake generation systems. Autoencoders

 doi:10.20944/preprints202603.1662.v1

work by compressing input data into a latent representation using an encoder and then reconstructing it with a decoder. In the context of face-swapping, dual autoencoders are often employed, one for the source and one for the target identity, sharing a common encoder but using identity-specific decoders. This approach enables the projection of different facial identities while preserving similar expressions and facial geometry. Autoencoders are particularly valued for their training stability and lower computational requirements compared to GANs, [113]. However, they tend to produce blurrier outputs with less fine detail, particularly in textures like skin or under dynamic lighting conditions, which can reduce their realism. Neural rendering techniques, [114], represent another major advancement. These methods combine classical graphics pipelines with deep neural networks to generate content, offering greater control over geometry, lighting, and camera parameters. This makes neural rendering particularly well suited for 3D-aware generation. Notable examples include deferred neural rendering, [115], neural textures, [116], and volumetric scene representations, all of which leverage learned representations of faces and scenes to ensure viewpoint consistency, enable expression manipulation, and support multi-angle synthesis. More recently, diffusion models have emerged as a powerful alternative for generative modeling, [117]. These models are inspired by statistical thermodynamics and work by gradually adding noise to training data, then learning to reverse the diffusion process in order to reconstruct the original clean content. Once trained, diffusion models can generate new data by iteratively denoising random noise samples through probabilistic transitions. Models such as Denoising Diffusion Probabilistic Models (DDPM) and Stable Diffusion, [99,117], have demonstrated state of the art performance in terms of visual quality, controllability, and interpretability. Compared to GANs, diffusion models offer more stable training and are less prone to mode collapse, resulting in broader diversity and higher fidelity outputs. However, they are computationally intensive and require many inference steps, making them less suitable for real-time applications. Nevertheless, advancements such as DDIMs (Denoising Diffusion Implicit Models), [118] and improvements in GPU processing power, in line with Huang's Law, [119] are helping to close this gap and enable more practical deployment.

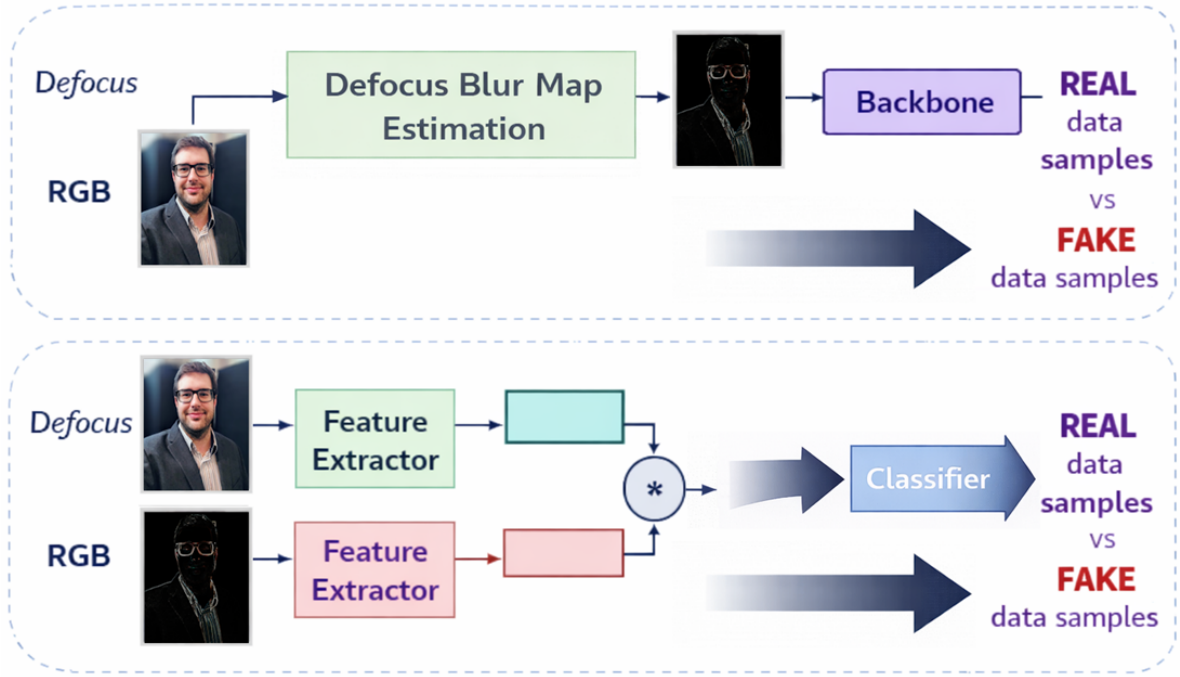

Figure 2. Defocus-based deepfake detection pipeline. Top: RGB image → defocus blur map estimation → backbone for real/fake classification. Bottom: defocus extractor + adder/classifier for real/fake artifacts. As such, the illustration exploits generation models' (GANs, autoencoders, etc.) failure to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

 doi:10.20944/preprints202603.1662.v1

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, as they introduce intelligent actors designed to evade both human and algorithmic scrutiny, [120]. One key application is adversarial robust deepfake synthesis, [121], 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, [122]. These may include perceptual loss, [123], identity preservation loss, [124], and feature matching loss, [125], often derived from intermediate layers of convolutional neural networks (CNNs). 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, [126,127]. These adversarial perturbations can be applied post hoc to already generated deepfakes, making them resistant to classifiers without significantly affecting their visual integrity. Such perturbations are particularly effective at deceiving CNN-based classifiers or frame-level analysis tools. A recent study, [128], has proposed the use of transferable adversarial attacks, where perturbations crafted to fool one detection model are also effective against others, even when those models use different architectures or are trained on different datasets. 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 adopts ensemble-based uncertainty estimation, [129]. This approach allows models to better assess the confidence of detection systems and exploit their weaknesses more effectively. 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

Figures 3 and 4 compare several leading deepfake detection methods using three benchmarks that are commonly reported in recent studies: FF++, Celeb-DF, and DFDC. Drawing on the comparative results presented in works such as [130], Figure 3 reports AUC values, where higher scores indicate a clearer separation between real and fake video samples. Similarly, Figure 4 presents EER values, of tests where lower scores reflect better accuracy at the point where false positive and false negative rates are equal. For all datasets and for all evaluation metrics, the method proposed and further analyzed recent work, [130], shows consistently better performance than the competing approaches, with higher AUC scores and lower EER values. This consistent behavior suggests that the method can handle different types of deepfake content and adapt well to multiple datasets. The observed improvements are mainly linked to effective training choices, including stronger feature extraction, focused data handling, and an architecture designed specifically for deepfake detection tasks. In addition, the method remains stable across different testing conditions and appears less affected by dataset-specific characteristics compared to other models. Overall, these figures demonstrate that the

 doi:10.20944/preprints202603.1662.v1

proposed network performs reliably in cross-dataset settings and supports robust decision-making, addressing several of the limitations discussed earlier in this paper.

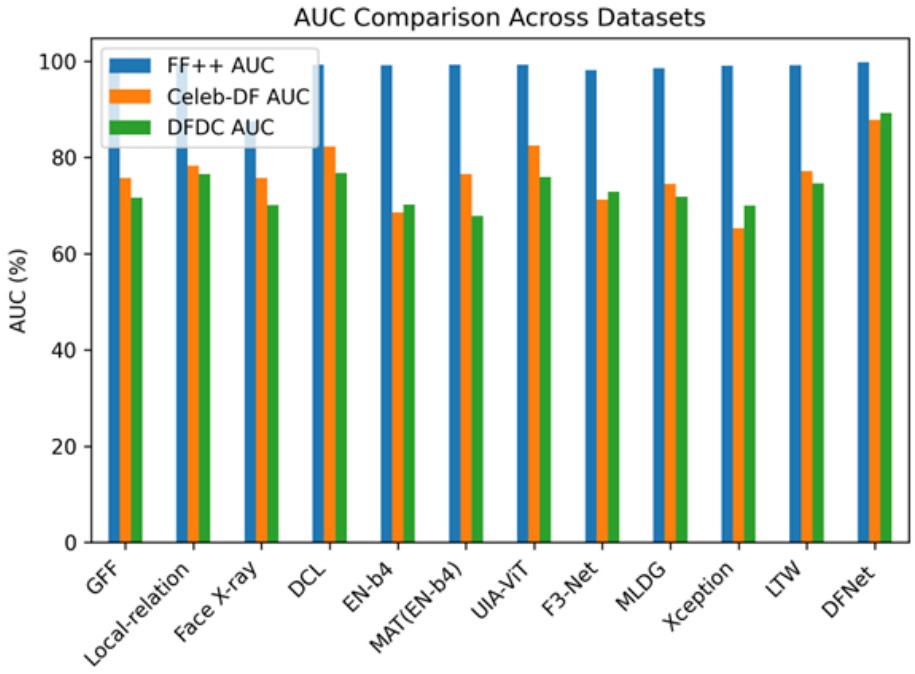

Figure 3. AUC comparison of top deepfake detectors on FF++, Celeb-DF, and DFDC. Higher AUC shows better separation of real vs. fake videos.

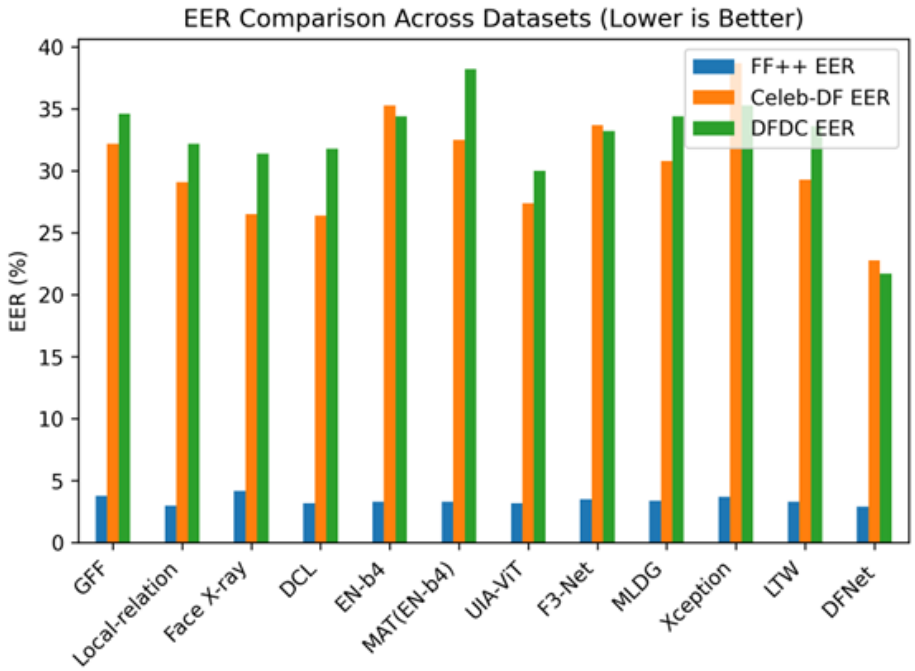

Figure 4. EER comparison of top deepfake detectors on FF++, Celeb-DF, and DFDC. Lower EER means better accuracy where false positives equal false negatives.

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

 doi:10.20944/preprints202603.1662.v1

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.

Data Availability Statement: The dataset will be made available on request. The data used to support the findings of this study are publicly available. The data used to support the findings of this study are available from the corresponding author upon request.

Acknowledgments: This work was supported without any funding.

Conflicts of Interest: The authors declare that they have no conflicts of interest.

References

1. Flood, J., & Lachlan, R. (2025). Beyond traditional expertise: How AI and blockchain are reshaping legal and accounting professional identity and practice in the fourth industrial revolution. SSRN. https://dx.doi.org/10.2139/ssrn.5207502
2. Gazis, A., & Gazi, T. (2021). Big data applications in industry fields. ITNOW, 63(2), 50–51. https://doi.org/10.1093/itnow/bwab056
3. Kruger, S., & Steyn, A. A. (2025). Navigating the fourth industrial revolution: A systematic review of technology adoption model trends. Journal of Science and Technology Policy Management, 16(10), 24–56. https://doi.org/10.1108/JSTPM-11-2022-0188
4. Akundi, A., Euresti, D., Luna, S., Ankobiah, W., Lopes, A., & Edinbarough, I. (2022). State of industry 5.0: Analysis and identification of current research trends. Applied System Innovation, 5(1), 27. https://doi.org/10.3390/asi5010027
5. Ystgaard, K. F., Atzori, L., Palma, D., Heegaard, P. E., Bertheussen, L. E., Jensen, M. R., & De Moor, K. (2023). Review of the theory, principles, and design requirements of human-centric Internet of Things (IoT). Journal of Ambient Intelligence and Humanized Computing, 14(3), 2827–2859. https://doi.org/10.1007/s12652-023-04539-3
6. Naeem, M., Jamal, T., Diaz-Martinez, J., Butt, S. A., Montesano, N., Tariq, M. I., Dela Hoz Franco, E., & De La Hoz Valdiris, E. (2022). Trends and future perspective challenges in big data. In J.-S. Pan, V. E. Balas, & C.-M. Chen (Eds.), Advances in intelligent data analysis and applications, 309–325. Springer. https://doi.org/10.1007/978-981-16-5036-9_30
7. Whittaker, L., Mulcahy, R., Letheren, K., Kietzmann, J., & Russell-Bennett, J. (2023). Mapping the deepfake landscape for innovation: A multidisciplinary systematic review and future research agenda. Technovation, 125, 102784. https://doi.org/10.1016/j.technovation.2023.102784
8. Chadha, A., Kumar, V., Kashyap, S., & Gupta, M. (2021). Deepfake: An overview. In P. K. Singh, S. T. Wierzchoń, S. Tanwar, M. Ganzha, & J. J. P. C. Rodrigues (Eds.), Proceedings of the second international conference on computing, communications, and cybersecurity, 557–566. Springer. https://doi.org/10.1007/978-981-16-0733-2_39
9. Roe, J., Perkins, M., Somoray, K., Miller, D., & Furze, L. (2025). To deepfake or not to deepfake: Higher education stakeholders' perceptions and intentions towards synthetic media. arXiv. https://doi.org/10.48550/arXiv.2502.18066
10. Slabin, U. (2024). Deepfake in science education: Why not? Journal of Baltic Science Education, 23(3), 416–420. https://dx.doi.org/10.33225/jbse/24.23.416

 doi:10.20944/preprints202603.1662.v1

11. Lim, W. M. (2023). Fact or fake? The search for truth in an infodemic of disinformation, misinformation, and malinformation with deepfake and fake news. Journal of Strategic Marketing. https://doi.org/10.1080/0965254X.2023.2253805
12. Agarwal, A., & Ratha, N. (2023). Manipulating faces for identity theft via morphing and deepfake: Digital privacy. In Handbook of statistics, Vol. 48, 223–241. Elsevier. https://doi.org/10.1016/bs.host.2022.12.003
13. de Rancourt-Raymond, A., & Smaili, N. (2023). The unethical use of deepfakes. Journal of Financial Crime, 30(4), 1066–1077. https://doi.org/10.1108/JFC-04-2022-0090
14. Lu, H., & Yuan, S. (2024). "I know it's a deepfake": The role of AI disclaimers and comprehension in the processing of deepfake parodies. Journal of Communication, 74(5), 359–373. https://doi.org/10.1093/joc/jqae022
15. Pawan, 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, 237–253. River Publishers. https://www.taylorfrancis.com/chapters/edit/10.1201/9788743801146-14/risks-individual-privacy-security-world-deepfake-technology-pawan-pant-mishra-rajendra-rajaram [Access Date: 09/02/2026]
16. Shaji, A. G., & Hovan, A. S. G. (2023). Deepfakes: The evolution of hyper realistic media manipulation. Partners Universal Innovative Research Publication, 1(2), 58–74. https://doi.org/10.5281/zenodo.10148558
17. Boediman, E. P. (2025). Exploring the impact of deepfake technology on public trust and media manipulation: A scoping review. Jurnal Komunikasi, 19(2), 131–152. https://doi.org/10.20885/komunikasi.vol19.iss2.art8
18. Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial nets. Communications of the ACM, 63, 11, 139-144. https://doi.org/10.1145/3422622
19. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10684-10695. https://doi.org/10.48550/arXiv.2112.10752
20. Kouzelis, T., Karypidis, E., Kakogeorgiou, I., Gidaris, S., & Komodakis, N. (2025). Boosting generative image modeling via joint image feature synthesis. arXiv. https://doi.org/10.48550/arXiv.2504.16064
21. Kumar, D. M., Kumar, M., Kapil, I. K., & Yadav, R. K. (2023). Deepfake creation using GANs and autoencoders and deepfake detection. In 2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN), 1–6. IEEE. https://doi.org/10.1109/ViTECoN58111.2023.10157962
22. Hatfield, A. R., & Badawy, A. H. A. (2025). Moore's law: What comes next? In H. R. Arabnia & L. Deligiannidis (Eds.), Software engineering research and practice and e-learning, e-business, enterprise information systems, and e-government (Communications in Computer and Information Science, Vol. 2263, 195–206. Springer. https://doi.org/10.1007/978-3-031-86644-9
23. 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. https://doi.org/10.1109/ACCESS.2021.3131517
24. Ghiurău, D., & Popescu, D. E. (2024). Distinguishing reality from AI: Approaches for detecting synthetic content. Computers, 14(1), 1. https://doi.org/10.3390/computers14010001
25. Daukantas, P. (2025). Generating and detecting deepfakes: A 21st century arms race. Optics and Photonics News, 36(2), 24–31. https://www.optica-opn.org/home/articles/volume_36/february_2025/features/generating_and_detecting_deepfakes_a_21st-century_arms_race/ [Access Date: 08/02/2026]
26. Pantserev, K. A. (2020). The malicious use of AI-based deepfake technology as the new threat to psychological security and political stability. In H. Jahankhani, S. Kendzierskyj, N. Chelvachandran, & J. Ibarra (Eds.), Cyber defence in the age of AI, smart societies and augmented humanity, 37–55. Springer. https://doi.org/10.1007/978-3-030-35746-7_3
27. 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. https://doi.org/10.1109/ACCESS.2023.3344653
28. Ahmed, N. U. R., Badshah, A., Adeel, H., Tajammul, A., Duad, A., & Alsahfi, V. (2024). Visual deepfake detection: Review of techniques, tools, limitations, and future prospects. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3523288
29. Rana, M. S., Nobi, M. N., Murali, B., & Sung, A. H. (2022). Deepfake detection: A systematic literature review. IEEE Access, 10, 25494–25513. https://doi.org/10.1109/ACCESS.2022.3154404

 doi:10.20944/preprints202603.1662.v1

30. Heidari, A., Navimipour, N. G., 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. https://doi.org/10.1002/widm.1520
31. Gourav, G., Kiran, R., Manish, G., Tony, J., Whiteside, S. T., & Prasad, M. (2023). A comprehensive review of deepfake detection using advanced machine learning and fusion methods. Electronics, 13(1), 95. https://doi.org/10.3390/electronics13010095
32. Kaur, A., Noori, H. A., Saikrishna, V., et al. (2024). Deepfake video detection: Challenges and opportunities. Artificial Intelligence Review, 57, 159. https://doi.org/10.1007/s10462-024-10810-6
33. Altuncu, E., Virginia, N. L., & Franqueira, S. L. (2024). Deepfake: Definitions, performance metrics and standards, datasets, and a meta-review. Frontiers in Big Data, 7, 1400024. https://doi.org/10.3389/fdata.2024.1400024
34. Stroebel, L., Llewellyn, M., Hartley, T., & Ahmed, M. (2023). A systematic literature review on the effectiveness of deepfake detection techniques. Journal of Cyber Security Technology, 7(2), 83–113. https://doi.org/10.1080/23742917.2023.2192888
35. Mcuba, M., Singh, A., Ikuesan, R. A., & Venter, H. (2023). The effect of deep learning methods on deepfake audio detection for digital investigation. Procedia Computer Science, 219, 211–219. https://doi.org/10.1016/j.procs.2023.01.283
36. Almutairi, Z., & Elgibreen, H. (2022). A review of modern audio deepfake detection methods: Challenges and future directions. Algorithms, 15(5), 155. https://doi.org/10.3390/a15050155
37. Wijethunga, R. L. M. A. P. C., Matheesha, D. M. K., Noman, A. A., De Silva, K. H. V. T. A., Tissera, M., & Rupasinghe, L. (2020). Deepfake audio detection: A deep learning-based solution for group conversations. In 2020 2nd International Conference on Advancements in Computing (ICAC), Vol. 1, 192–197. IEEE. https://doi.org/10.1109/ICAC51239.2020.9357161
38. Müller, N. M., Czempin, P., Dieckmann, F., Froghyar, A., & Böttinger, K. (2022). Does audio deepfake detection generalize? arXiv. https://doi.org/10.48550/arXiv.2203.16263
39. Rabhi, M., Bakiras, S., & Di Pietro, R. (2024). Audio-deepfake detection: Adversarial attacks and countermeasures. Expert Systems with Applications, 250, 123941. https://doi.org/10.1016/j.eswa.2024.123941
40. 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. https://doi.org/10.1073/pnas.2110013119
41. 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. https://doi.org/10.3390/jimaging9050089
42. 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. https://doi.org/10.1016/j.fsidi.2022.301374
43. 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. https://doi.org/10.3390/app15031225
44. Mahmud, B. U., & Sharmin, A. (2021). Deep insights of deepfake technology: A review. arXiv. https://doi.org/10.48550/arXiv.2105.00192
45. 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. http://doi.org/10.18178/ijssh.2024.14.2.1194
46. Chapagain, D., Kshetri, N., & Aryal, B. (2024). Deepfake disasters: A comprehensive review of technology, ethical concerns, countermeasures, and societal implications. In 2024 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC), 1–9. IEEE. https://doi.org/10.1109/ETNCC63262.2024.10767452
47. Alanazi, S., Asif, S., Caird-Daley, A., et al. (2025). Unmasking deepfakes: A multidisciplinary examination of social impacts and regulatory responses. Human-Centric Intelligent Systems Integration. https://doi.org/10.1007/s42454-025-00060-4
48. Meneses, J. P. (2021). Deepfakes and the 2020 US elections: What (did not) happen. arXiv. https://doi.org/10.48550/arXiv.2101.09092
49. Yi, J., Wang, C., Tao, J., Zhang, X., Chengyang, Y., & Zhao, Y. (2023). Audio deepfake detection: A survey. Sensors, 25(7), 1989. https://doi.org/10.3390/s25071989

 doi:10.20944/preprints202603.1662.v1

50. Fahmi, B. M., & Farouk, M. A. (2024). The extensive impacts of deepfake technology in Arab societies: Ethical, legal, and social challenges. Arab Journal of Media & Communication Research, 44, 162–176. https://doi.org/10.21608/jkom.2024.354797
51. Tan, C., Zhao, Y., Wei, S., Gu, G., Liu, P., & Wei, Y. (2024). Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 38, 5052–5060. https://doi.org/10.1609/aaai.v38i5.28310
52. 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. https://doi.org/10.1016/j.cosrev.2024.100677
53. Mansoor, N., & Iliev, A. I. (2025). Explainable AI for deepfake detection. Applied Sciences, 15(2), 725. https://doi.org/10.3390/app15020725
54. Mittal, S., Joshi, M., Vats, P., Upadhayay, G. M., Vats, S. K., & Kumar, S. (2024). Virtual illusions: Unleashing deepfake expertise for enhanced visual effects in film production. In 2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 1–6. IEEE. https://doi.org/10.1109/ICRITO61523.2024.10522334
55. Shao, R., Wu, T., & Liu, Z. (2022). Detecting and recovering sequential deepfake manipulation. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, & T. Hassner (Eds.), Computer vision – ECCV 2022, Lecture Notes in Computer Science, Vol. 13673, 712–728. Springer. https://doi.org/10.1007/978-3-031-19778-9_41
56. Amezaga, N., & Hajek, J. (2022). Availability of voice deepfake technology and its impact for good and evil. In Proceedings of the 23rd Annual Conference on Information Technology Education, 23–28. https://doi.org/10.1145/3537674.3554742
57. Yoon, J., Lledot, A. P., Camacho, D., & Choi, C. (2024). Triple-modality interaction for deepfake detection on zero-shot identity. Information Fusion, 109, 102424. https://doi.org/10.1016/j.inffus.2024.102424
58. Matern, F., Riess, C., & Stamminger, M. (2019). Exploiting visual artifacts to expose deepfakes and face manipulations. In 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), 83–92. IEEE. https://doi.org/10.1109/WACVW.2019.00020
59. Lewis, J. K., Toubal, I. E., Chen, H., Sandesera, V., Lomnitz, M., Arias, Z. H., Prasad, C., & Palaniappan, K. (2020). Deepfake video detection based on spatial, spectral, and temporal inconsistencies using multimodal deep learning. In 2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), 1–9. IEEE. https://doi.org/10.1109/AIPR50011.2020.9425167
60. Luo, X., & Wang, Y. (2025). Frequency domain masking and spatial interaction for generalizable deepfake detection. Electronics, 14(7), 1302. https://doi.org/10.3390/electronics14071302
61. Ciftci, U. A., Demir, I., & Yin, L. (2020). 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), 1–10. IEEE. https://doi.org/10.1109/IJCB48548.2020.9304909
62. Google LLC. (2021). Methods and systems for detecting deepfakes (U.S. Patent Application No. US20210142065A1). https://patents.google.com/patent/US20210142065A1 [Access Date: 09/02/2026]
63. Karathanasis, A., Violos, J., & Kompatsiaris, I. (2025). A comparative analysis of compression and transfer learning techniques in deepfake detection models. Mathematics, 13(5), 887. https://doi.org/10.3390/math13050887
64. Sharma, M., & Kaur, M. (2023). Deep Insights of Deepfake Technology: A Review, arxiv. https://doi.org/10.48550/arXiv.2105.00192
65. Byman, D. L., Gao, C., Meserole, C., & Subrahmanian, V. S. (2023). Deep fakes and international conflict (Vol. 8). Brookings Institution. https://www.brookings.edu/articles/deepfakes-and-international-conflict/ [Access Date: 09/02/2026]
66. 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. https://doi.org/10.32084/tkp.8615
67. Bhandari, R., & Bhandari, S. (2025). Artificial intelligence: Understanding deepfakes. EDPACS, 21-31. https://doi.org/10.1080/07366981.2025.2484863
68. Ölvecký, M., Huraj, L., & Brlej, I. (2025). Evaluating the effectiveness of deepfake video detection tools: A comparative study. TEM Journal, 14(1), 64. http://doi.org/10.18421/TEM141-07
69. Neyazi, T. A., Nadaf, A. H., Tan, K. E., & Schroeder, R. (2024). Does trust in government moderate the perception towards deepfakes? Comparative perspectives from Asia on the risks of AI and misinformation for democracy. Government Information Quarterly, 41(4), 101980. https://doi.org/10.1016/j.giq.2024.101980

 doi:10.20944/preprints202603.1662.v1

70. Obioha, V. O., Gbadebo, M. O., Olaniyi, O. O., Chinye, N. C., & Balogun, A. Y. (2025). Innovative regulation of open source intelligence and deepfakes AI in managing public trust, Journal of Engineering Research and Reports, 27(2), 137-156. https://doi.org/10.9734/jerr/2025/v27i21400
71. Haut, K., Wohn, C., Antony, V., Goldfarb, A., Welsh, M., Sumanthiran, D., Ali, M. D. R., & Hoque, E. (2022). Demographic feature isolation for bias research using deepfakes. In Proceedings of the 30th ACM International Conference on Multimedia, 6890–6897. https://doi.org/10.1145/3503161.3549204
72. Ahmad, O., Khan, M. S., Jan, S., & Khan, I. (2025). Deepfake audio detection for Urdu language using deep neural networks. 13, 97765-97778, IEEE Access. https://doi.org/10.1109/ACCESS.2025.3571293
73. Jalan, D., Jain, S., Tuli, A., Chaudhary, V., Sharma, A., & Kumar, A. (2025). Deepfake detection: A comprehensive analysis of modern techniques. In A. Sharma & R. Rani (Eds.), Artificial intelligence and speech technology, Communications in Computer and Information Science, Vol. 2390, 102–115. Springer. https://doi.org/10.1007/978-3-031-91340-2_7
74. Khan, A. A., Laghari, A. A., Inam, S. A., et al. (2025). A survey on multimedia-enabled deepfake detection: State-of-the-art tools and techniques, emerging trends, current challenges and limitations, and future directions. Discover Computing, 28, 48. https://doi.org/10.1007/s10791-025-09550-0
75. Khan, F. A., & Khan, M. K. (2025). Generative AI and deepfake detection in biometric systems. Cognitive Computation, 17, 112. https://doi.org/10.1007/s12559-025-10469-3
76. Alrawahneh, A. A. M., Abdullah, S. N. A. S., Abdullah, S. N. H. S., et al. (2025). Video authentication detection using deep learning: A systematic literature review. Applied Intelligence, 55, 239. https://doi.org/10.1007/s10489-024-05997-8
77. Sunil, R., Mer, P., Diwan, A., Mahadeva, R., & Sharma, A. (2025). Exploring autonomous methods for deepfake detection: A detailed survey on techniques and evaluation. Heliyon, e42273. https://doi.org/10.1016/j.heliyon.2025.e42273
78. Li, X., Chen, P. Y., & Wei, W. (2025). Measuring the robustness of audio deepfake detectors. arXiv. https://doi.org/10.48550/arXiv.2503.17577
79. Shaaban, O. A., & Yildirim, R. (2025). Audio deepfake detection using deep learning. Engineering Reports, 7(3), e70087. https://doi.org/10.1002/eng2.70087
80. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of StyleGAN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 8110–8119. IEEE. https://doi.org/10.1109/CVPR42600.2020.00813
81. Shaoanlu. (2018). Faceswap-GAN: A deepfakes using GAN framework in Keras [Computer software]. https://github.com/shaoanlu/faceswap-GAN [Access Date: 08/02/2026]
82. Das, A., Das, S., & Dantcheva, A. (2021). Demystifying attention mechanisms for deepfake detection. In 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), 1–7. IEEE. https://doi.org/10.1109/FG52635.2021.9667026
83. Yang, P., Cheung, N. M., & Ma, X. (2025). Text-to-image generation and editing: A survey. arXiv. https://doi.org/10.48550/arXiv.2505.02527
84. 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. https://doi.org/10.1049/ipr2.13201
85. Prashnani, E., Goebel, M., & Manjunath, B. S. (2025). Generalizable deepfake detection with phase-based motion analysis, 34, 100-112, IEEE Transactions on Image Processing. https://doi.org/10.1109/TIP.2024.3441821
86. 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 (CVPR), 2387–2395. https://doi.org/10.1145/3292039
87. Kohli, A., & Gupta, A. (2021). Detecting deepfake, FaceSwap and Face2Face facial forgeries using frequency CNN. Multimedia Tools and Applications, 80(12), 18461–18478. https://doi.org/10.1007/s11042-020-10420-8
88. Suratkar, S., & Sharma, P. (2022). A simple and effective way to detect deepfakes: Using 2D and 3D CNN. In U. P. Rao, S. J. Patel, P. Raj, & A. Visconti (Eds.), Security, privacy and data analytics, Lecture Notes in Electrical Engineering, Vol. 848, 265–277. Springer. https://doi.org/10.1007/978-981-16-9089-1_19
89. Bontcheva, K., Papadopoulos, S., Teyssou, D., Tsabouraki, D., Spangenberg, J., Srba, I., et al. (2026). Using AI to Tackle Disinformation: Methods and Tools from the vera. ai Project, In Disinformation: A Multi-Disciplinary Analysis, 465-487. Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-00480-2_24

 doi:10.20944/preprints202603.1662.v1

90. Agarwal, S., & Farid, H. (2021). Detecting deepfake videos from aural and oral dynamics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 981–989. IEEE. https://doi.org/10.1109/CVPRW53098.2021.00109
91. Holliday, C. (2024). Ghosts in the celluloid: AI video dubbing and TrueSync. JCMS: Journal of Cinema and Media Studies, 64(1), 175–182. https://doi.org/10.1353/cj.2024.a944431
92. Stanishevskii, G., Steczkiewicz, J., Szczepanik, T., Tadeja, S., Tabor, J., & Spurek, P. (2024). Deepfake for the good: Generating avatars through face-swapping with implicit deepfake generation. arXiv. https://doi.org/10.48550/arXiv.2402.06390
93. Lees, D., Bashford-Rogers, T., & Keppel-Palmer, M. (2021). The digital resurrection of Margaret Thatcher: Creative, technological and legal dilemmas in the use of deepfakes in screen drama. Convergence, 27(4), 954–973. https://doi.org/10.1177/13548565211030452
94. Ping, J. (2024). The resurrection will not be televised: Legal remedies for posthumous deepfakes. Georgetown Law Technology Review, 8, 338. https://georgetownlawtechreview.org/the-resurrection-will-not-be-televised-legal-remedies-for-posthumous-deepfakes/GLTR-05-2024/ [Access Date: 09/02/2026]
95. Amerini, I., Barni, M., Battiato, S., Bestagini, P., Boato, G., Bruni, V., Caldelli, R., De Natale, F., De Nicola, R., Guarnera, L., & Mandelli, S. (2025). Deepfake media forensics: Status and future challenges. Journal of Imaging, 11(3), 73. https://doi.org/10.3390/jimaging11030073
96. Liu, Z., & Zheng, Y. (2022). Virtual world under AI: Augmented reality and deep synthesis. In AI ethics and governance. Springer. https://www.springerprofessional.de/en/ai-ethics-and-governance/22167576 [Access Date: 09/02/2026]
97. Yasrab, R., Jiang, W., & Riaz, A. (2021). Fighting deepfakes using body language analysis. Forecasting, 3(2), 303–321. https://doi.org/10.3390/forecast3020020
98. Yang, W., Zhang, Z., Zhou, X., Duan, J., & Cao, J. (2025). TT-DF: A large-scale diffusion-based dataset and benchmark for human body forgery detection. In Z. Lin et al. (Eds.), Pattern recognition and computer vision, Lecture Notes in Computer Science, Vol. 15041. Springer. https://doi.org/10.1007/978-981-97-8795-1_29
99. 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. https://doi.org/10.1155/int/9987535
100. Dong, C., Bhagavatula, V., Zhou, Z., & Kumar, A. (2024). Towards more accurate fake detection on images generated from advanced generative and neural rendering models. arXiv. https://doi.org/10.48550/arXiv.2411.08642
101. Kim, J., Li, Y., & Shin, B. S. (2024). Volumetric imitation generative adversarial networks for anatomical human body modeling. Bioengineering, 11(2), 163. https://doi.org/10.3390/bioengineering11020163
102. Bohacek, M., & Farid, H. (2024). Lost in translation: Lip-sync deepfake detection from audio-video mismatch. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 4315–4323. IEEE. https://doi.org/10.1109/CVPRW63382.2024.00435
103. Gopinath, S., Rajasekar, A., Sharmiela, G. V., Monica, T., & Pooja, M. A. (2024). Deepfake and face swapping detection: A review. In 2024 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS), 1–6. IEEE. https://doi.org/10.1109/ICPECTS62210.2024.10780003
104. Kaushal, A., Kumar, S., & Kumar, R. (2024). A review on deepfake generation and detection: Bibliometric analysis. Multimedia Tools and Applications, 83, 87579–87619. https://doi.org/10.1007/s11042-024-18706-x
105. Fernando, T., Priyasad, D., Sridharan, S., Ross, A., & Fookes, C. (2025). Face deepfakes: A comprehensive review. arXiv. https://doi.org/10.48550/arXiv.2502.09812
106. Wang, T. (2025). A comprehensive review of image-to-image translation, face manipulation, and neural style transfer methods. Authorea Preprints. https://doi.org/10.36227/techrxiv.174431621.17773362/v1
107. Ben Aissa, F., Hamdi, M., Zaied, M., et al. (2024). An overview of GAN-deepfakes detection: Proposal, improvement, and evaluation. Multimedia Tools and Applications, 83, 32343–32365. https://doi.org/10.1007/s11042-023-16761-4
108. Ait Sghir, A., El Amraoui, S., Elafi, I., & Zrira, N. (2025). StyleGAN for deepfake generation: A comparative study. In Communication and information technologies through the lens of innovation: Proceedings of the 5th International Conference on Advanced Technologies and Humanity (ICATH 2023), 103–112. Springer Nature. https://doi.org/10.1007/978-3-031-74470-9_13
109. Cheng, X. (2024). Refining CycleGAN with attention mechanisms and age-aware training for realistic deepfakes. Heliyon, 10(16), e36665. https://doi.org/10.1016/j.heliyon.2024.e36665

 doi:10.20944/preprints202603.1662.v1

110. Yang, W. C. (2021). StarGAN deepfake videos detection based on no-reference image quality assessments and support vector machine. Forensic Science Journal, 20(1), 1–12. https://doi.org/10.6593/FSJ.202112_20(1).0001
111. Kozak, J. P., Zhang, R., Porter, M., Song, Q., Liu, J., Wang, B., Wang, R., Saito, W., & Zhang, Y. (2023). Stability, reliability, and robustness of GAN power devices: A review. IEEE Transactions on Power Electronics, 38(7), 8442–8471. https://doi.org/10.1109/TPEL.2023.3266365
112. Durall, R., Chatzimichailidis, A., Labus, P., & Keuper, J. (2020). Combating mode collapse in GAN training: An empirical analysis using Hessian eigenvalues. arXiv. https://doi.org/10.48550/arXiv.2012.09673
113. Ghojogh, B., Ghodsi, A., Karray, F., & Crowley, M. (2021). Generative adversarial networks and adversarial autoencoders: Tutorial and survey. arXiv. https://doi.org/10.48550/arXiv.2111.13282
114. Tewari, A., Fried, O., Thies, J., Sitzmann, V., Lombardi, S., Sunkavalli, K., Brualla, R. M., Simon, T., Saragih, J., Nießner, M., & Pandey, R. (2020). State of the art on neural rendering. Computer Graphics Forum, 39, 701–727. https://doi.org/10.1111/cgf.14022
115. Thies, J., Zollhöfer, M., & Nießner, M. (2019). Deferred neural rendering: Image synthesis using neural textures. ACM Transactions on Graphics, 38(4), 1–12. https://doi.org/10.1145/3306346.3323035
116. Haseena, S., Saroja, S., Shri Dharshini, D., & Nivetha, A. (2023). TVN: Detect deepfake images using texture variation network. Inteligencia Artificial, 26(72), 1–14. https://doi.org/10.4114/intartif.vol26iss72pp1-14
117. Aghasanli, A., Kangin, D., & Angelov, P. (2023). Interpretable through prototypes deepfake detection for diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, 467–474. https://doi.org/10.1109/ICCVW60793.2023.00053
118. Ganguly, R., Bah, M.D., Dahmane, M. (2025). Diffusion Models as a Representation Learner for Deepfake Image Detection. In: Antonacopoulos, A., Chaudhuri, S., Chellappa, R., Liu, CL., Bhattacharya, S., Pal, U. (eds) Pattern Recognition. ICPR 2024. Lecture Notes in Computer Science, vol 15321. Springer, Cham. https://doi.org/10.1007/978-3-031-78305-0_15
119. Noura, H. N., et al. (2025). Advanced machine learning in smart grids: An overview. Internet of Things and Cyber-Physical Systems, 5, 95-142. https://doi.org/10.1016/j.iotcps.2025.05.002
120. Sharma, S., Singh, S. (2025). Spear or Shield: Mastering the Art of Gen-AI in Face Recognition. In: Singh, J., Goyal, S.B., Kumar, M., Mittal, R. (eds) Advanced Network Technologies and Computational Intelligence. ICANTCI 2024. Communications in Computer and Information Science, vol 2382. Springer, Cham. https://doi.org/10.1007/978-3-031-86069-0_31
121. Hussain, S., Neekhara, P., Jere, M., Koushanfar, F., & McAuley, J. (2021). Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 3347–3358. IEEE. https://doi.ieeecomputersociety.org/10.1109/WACV48630.2021.00339
122. Dang, Q., Zhang, G., Wang, L., Yang, S., & Zhan, Z. (2024). A generative adversarial networks model-based evolutionary algorithm for multimodal multi-objective optimization. IEEE Transactions on Emerging Topics in Computational Intelligence, 1-10. IEEE. https://doi.org/10.1109/TETCI.2024.3397996
123. Wang, R., Huang, Z., Chen, Z., Liu, L., Chen, J., & Wang, L. (2022). Anti-forgery: Towards a stealthy and robust deepfake disruption attack via adversarial perceptual-aware perturbations. arXiv. https://doi.org/10.48550/arXiv.2206.00477
124. 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. https://doi.org/10.3390/info16040270
125. Jiang, W., & Shu, L. (2025). A generative adversarial network face recognition algorithm based on mean feature matching. In International Conference on Mechatronics and Intelligent Control (ICMIC 2024), Vol. 13447, 1382–1388. SPIE. https://doi.org/10.1117/12.3045918
126. Jeong, W., Kim, D., Ro, Y., & Choi, J. (2022). FrepGAN: Robust deepfake detection using frequency-level perturbations. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 1060–1068. https://doi.org/10.1609/aaai.v36i1.19990
127. Tsigos, K., Apostolidis, E., & Mezaris, V. (2025). Improving the perturbation-based explanation of deepfake detectors through the use of adversarially generated samples. arXiv. https://doi.org/10.48550/arXiv.2502.03957
128. Farooq, M. U., Khan, A., Uddin, K., & Malik, K. M. (2025). Transferable adversarial attacks on audio deepfake detection. arXiv. https://doi.org/10.48550/arXiv.2501.11902

 doi:10.20944/preprints202603.1662.v1

129. Dutta, H., Pandey, A., & Bilgaiyan, S. (2021). EnsembleDet: Ensembling against adversarial attack on deepfake detection. Journal of Electronic Imaging, 30(6), 063030. https://doi.org/10.1117/1.JEI.30.6.063030
130. 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. https://doi.org/10.1109/TCE.2024.3476477

Short Biography of Authors

Alexandros Gazis is a software engineer and researcher with extensive experience in distributed systems, game-based learning, AI, middleware architectures, and intelligent data processing. His work focuses on cloud and IoT middleware, context-aware systems, serious games, and performance-aware architectures. 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 and smart environments. (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)

 doi:10.20944/preprints202603.1662.v1

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)