Journal of
Sensor and Actuator Networks

MDPI

Systematic Review

Generative Artificial Intelligence and the Evolving Challenge of Deepfake Detection: A Systematic Analysis

Reza Babaei [1], Samuel Cheng [1,*], Rui Duan [2] and Shangqing Zhao [3]

[1] School of Electrical and Computer Engineering, University of Oklahoma, Norman, OK 73019, USA; rezababaei@ou.edu
[2] School of Computer Science, University of Missouri-Kansas City, Kansas City, MO 64110, USA; ruiduan@umkc.edu
[3] School of Computer Science, University of Oklahoma, Norman, OK 73019, USA; shangqing@ou.edu
* Correspondence: samuel.cheng@ou.edu

Abstract: Deepfake technology, which employs advanced generative artificial intelligence to create hyper-realistic synthetic media, poses significant challenges across various sectors, including security, entertainment, and education. This literature review explores the evolution of deepfake generation methods, ranging from traditional techniques to state-of-the-art models such as generative adversarial networks and diffusion models. We navigate through the effectiveness and limitations of various detection approaches, including machine learning, forensic analysis, and hybrid techniques, while highlighting the critical importance of interpretability and real-time performance in detection systems. Furthermore, we discuss the ethical implications and regulatory considerations surrounding deepfake technology, emphasizing the need for comprehensive frameworks to mitigate risks associated with misinformation and manipulation. Through a systematic review of the existing literature, our aim is to identify research gaps and future directions for the development of robust, adaptable detection systems that can keep pace with rapid advancements in deepfake generation.

Keywords: deepfake detection; generative artificial intelligence; digital forensics; media security

check for updates

Academic Editor: Adnan M. Abu-Mahfouz

Received: 25 December 2024
Revised: 26 January 2025
Accepted: 30 January 2025
Published: 6 February 2025

Citation: Babaei, R.; Cheng, S.; Duan, R.; Zhao, S. Generative Artificial Intelligence and the Evolving Challenge of Deepfake Detection: A Systematic Analysis. *J. Sens. Actuator Netw.* **2025**, *14*, 17. https://doi.org/10.3390/jsan14010017

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Deepfake technology, powered by advancements in generative artificial intelligence (AI), has revolutionized the creation of realistic synthetic media. Deepfake technologies entail the manipulation of existing content or the generation of entirely novel media, making it possible to alter facial attributes, swap faces, or synthesize scenarios where individuals appear to say or do things they never actually did [1]. Broadly, deepfakes can be categorized into two types: those that transform existing media and those generated entirely by AI models, such as image-generation systems, without reliance on a pre-existing source.

Conventional deepfakes rely on techniques like face swapping [2], where one individual's face is replaced with another, and puppet-master techniques [3], which animate static images to mimic the movements of a source. These methods, while capable of producing convincing results, often depend on identifiable and traceable source materials. In contrast, modern generative AI models, including transformer and diffusion-based architectures, have transcended these limitations. For instance, models like DALL·E generate intricate images from textual prompts, producing outputs that can be indistinguishable from real photographs [4]. These innovations extend to cross-modal generation, enabling seamless

transitions from text to images, videos, or audio, thus expanding their practical and creative applications. However, the enhanced capabilities of the generative models raise concerns regarding their potential for misuse.

The dual-edged nature of deepfake technology underscores its transformative potential and the challenges it poses. On the one hand, deepfakes offer opportunities for creative and practical innovation. For example, they enable realistic film dubbing, immersive storytelling, and educational reanimations of historical figures [5]. On the other hand, they present significant risks, such as facilitating misinformation, manipulating public opinion, and causing reputational harm through malicious applications like revenge pornography and political sabotage [6–9].

As the sophistication of deepfakes increases, detection algorithms face escalating challenges, prompting an ongoing arms race between their creators and defenders [5]. The rapid evolution of generative AI further highlights the need for responsible development and deployment of deepfake technologies [10]. By addressing these challenges proactively, researchers and policymakers can mitigate risks while fostering innovations that benefit society.

1.1. Review Objectives

This systematic review adheres to PRISMA 2020 guidelines, aiming to provide a comprehensive exploration of the evolution, applications, and challenges of deepfake technology, with a focused lens on the transformative role of generative AI in both creation and detection methods. The objectives include critically analyzing state-of-the-art deepfake generation techniques, such as generative adversarial networks (GANs) and diffusion models, and evaluating their capabilities to produce highly realistic and complex synthetic media. For example, techniques such as StyleGAN3 and Stable Diffusion are reviewed to illustrate advances in image and video manipulation.

Additionally, the review assesses the effectiveness of detection methods, including forensic analysis, machine learning models, and hybrid approaches, in countering the sophistication of modern deepfakes. Studies implementing convolutional neural networks (CNNs) and recurrent neural networks (RNNs) for detecting artifacts in GAN-generated videos are compared against multi-modal approaches combining visual and audio cues. The review also identifies gaps in the research landscape, including the need for real-time, robust detection systems capable of addressing cross-modal and adversarially crafted deepfakes. A gap analysis highlights the challenges of detecting synthetic voices in video-based deepfakes. Furthermore, actionable directions for future research and policy-making are proposed, such as fostering adaptive detection algorithms, promoting interdisciplinary collaborations, and enhancing public awareness to manage the dual-edged nature of deepfake technologies. Insights into legislative efforts and ethical AI design principles are discussed to balance innovation and regulation.

1.2. Methodology for Literature Collection

This review followed a systematic approach to literature collection and evaluation. The search strategy involved using databases such as Semantic Scholar, Google Scholar, IEEE Xplore Digital Library, arXiv, and other relevant research-sharing platforms to identify studies from diverse academic disciplines, including computer vision, deep learning, and digital forensics. Keywords such as “deepfake detection”, “deepfake generation”, and “deepfake generative AI” were utilized. Studies published between January 2016 and 22 October 2024, were included. Automation tools like Research Rabbit, Litmaps, and Connected Papers assisted in tracing connections between foundational studies and derivative works.

The inclusion criteria for this review prioritized peer-reviewed journal articles (Q1 priority) and open-access papers based on their contributions to the field. Studies reporting experimental results, novel methodologies, or theoretical insights into deepfake generation and detection were included, provided their full-text PDFs were available. Papers lacking accessible full texts, studies not directly contributing to the deepfake field, and those with unverifiable claims or insufficient methodological detail were excluded. Articles that did not directly address deepfakes were excluded from the review, such as the generative AI work by Fui et al. [11].

The screening process began with an initial pool of over 300 papers. Data extraction was facilitated by tools like Docanalyzer.ai, with further validation against the original text performed by the author. Abstracts and contribution sections were prioritized for preliminary relevance assessment. For example, papers providing performance metrics of deepfake detection models were shortlisted. A detailed review was then conducted on 156 papers, which were manually screened by the author. The study selection process is illustrated in the PRISMA flow diagram presented in Figure 1. Numerical results were extracted primarily from tables; if unavailable, the main text was reviewed. Bias was managed by assuming peer-reviewed studies to be unbiased, while non-peer-reviewed papers were manually checked for potential conflicts of interest or methodological biases.

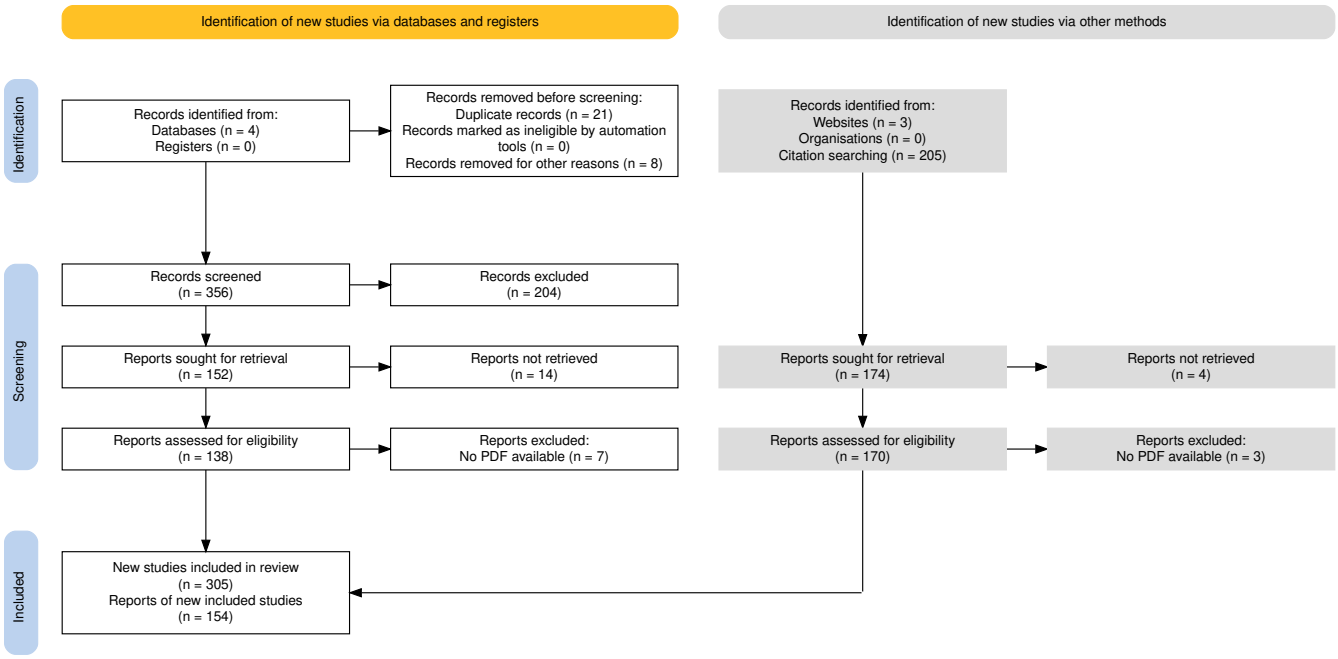

Figure 1. PRISMA flow diagram for literature selection.

The metrics collected in this review included dataset usage, performance measures such as accuracy and area under curve (AUC) score, and mean squared error (MSE). Detection accuracy on datasets like FaceForensics++ (FF++) and CelebDF was systematically compared. Studies proposing novel models were emphasized, while reviews without experimental contributions provided interpretative insights.

This review acknowledges several limitations. First, the reliance on human screening introduces potential subjectivity and oversight. Second, the search scope was not exhaustive, as seed papers served as the primary protocol for identifying related works. Finally, the review protocol was not pre-registered, which limits reproducibility, although the methodology followed a structured framework consistent with PRISMA 2020 standards. Future systematic reviews will adopt pre-registration to ensure methodological transparency and adherence to best practices.

By detailing the review objectives and methodology, this protocol ensures transparency and rigor, offering a reproducible framework for systematic exploration of deepfake technologies.

2. Deepfake Technology Overview

This section provides a comprehensive overview of the development and current landscape of deepfake technology, setting the stage for a deeper exploration of its applications and the implications in various domains.

2.1. History of Deepfake Technology

The history of deepfake technology can be traced through its evolution through distinct stages, each marked by significant advancements in AI and computational techniques. Early research from 2014 to 2016, spurred by breakthroughs in deep learning and neural networks, focused on enhancing the ability to edit and synthesize images and videos. These advancements laid the groundwork for more sophisticated manipulations, gradually shifting the focus from traditional photo editing to automated, AI-driven methods. The term "deepfake" gained widespread attention in 2017 when celebrity face-swapping videos surfaced on Reddit, highlighting the profound potential of this technology to create misleading content [12,13].

By 2018, the accessibility of deepfake technology reached new heights. High-profile examples, such as a video impersonating former U.S. President Barack Obama, showcased the startling realism that could be achieved through deepfake manipulations [14,15]. This period also saw the release of user-friendly software like FakeApp and DeepFaceLab, popularizing the creation of deepfakes for individuals with minimal technical expertise. While these tools enabled creative applications, they also raised significant ethical and security concerns, particularly regarding privacy and potential misuse [16–18].

From 2019 to 2020, growing awareness of the risks posed by deepfakes led to a global response from governments, organizations, and the research community. Privacy, misinformation, and security concerns prompted initiatives like Facebook's Deepfake Detection Challenge (DFDC) [19], aimed at fostering the development of robust detection tools [20]. Such efforts underscored the urgent need to mitigate the risks associated with increasingly sophisticated deepfake content.

Since 2021, deepfake technology has undergone a transformation in quality and realism, driven by advancements in machine learning architectures such as vision transformers and diffusion models. These innovative methods enable the creation of hyper-realistic deepfakes, including synthetic media entirely detached from real-world data, pushing the boundaries of generative AI [21–23]. This progression highlights both the immense creative potential of deepfakes and the escalating challenge of distinguishing authentic content from sophisticated forgeries.

2.2. Evolution of Deepfake Generation Techniques

The evolution of deepfake technology has progressed through several stages, each marking significant advancements in the realism and capabilities of synthetic media. Early methods primarily used autoencoders, which utilized an encoder to compress input data and a decoder to reconstruct them [2]. These models, often involving two autoencoders, one for the source face and one for the target face, shared a common encoder to facilitate face swapping. While this approach demonstrated the feasibility of face manipulation, the outputs often had smooth textures and lacked fine details, giving them an artificial appearance [15,17].

The introduction of GANs in 2014 revolutionized synthetic media generation. GANs use a generator to create synthetic images and a discriminator to evaluate their authenticity. This adversarial process allows the generator to produce highly realistic outputs [24]. Notable GAN architectures, such as StyleGAN and StarGAN, have enabled the creation of high-resolution, photorealistic images, with techniques like puppet-master methods and facial expression transfer enhancing realism by mapping expressions from one individual to another [24–27]. However, despite these advancements, GAN-based deepfakes often suffer from artifacts, such as mismatched lip movements, inconsistencies in blinking, or lighting variations, which detection systems have been able to exploit [2,9].

Further sophistication has been achieved in synchronizing audio and visual content, resulting in highly convincing audio-visual deepfakes. By integrating synthetic voices with visual content, these techniques achieve seamless synchronization, but challenges like mismatched lip movements and disjointed appearances still persist. Neural textures, which use deferred neural rendering to create photorealistic media, have further pushed the boundaries of deepfake realism [28]. Despite their success, they struggle with inconsistencies in skin tones and visible splicing boundaries, particularly under complex lighting conditions [29–31].

The most recent innovation in deepfake technology is the adoption of diffusion models, which transform simple distributions, like Gaussian noise, into complex data that closely resemble real-world images or videos [32]. However, they are not without their limitations, as they can still produce artifacts such as residual noise and temporal inconsistencies in video outputs, indicating areas for further improvement [32,33]. Figure 2 depicts an overview of the key generative deepfake methods.

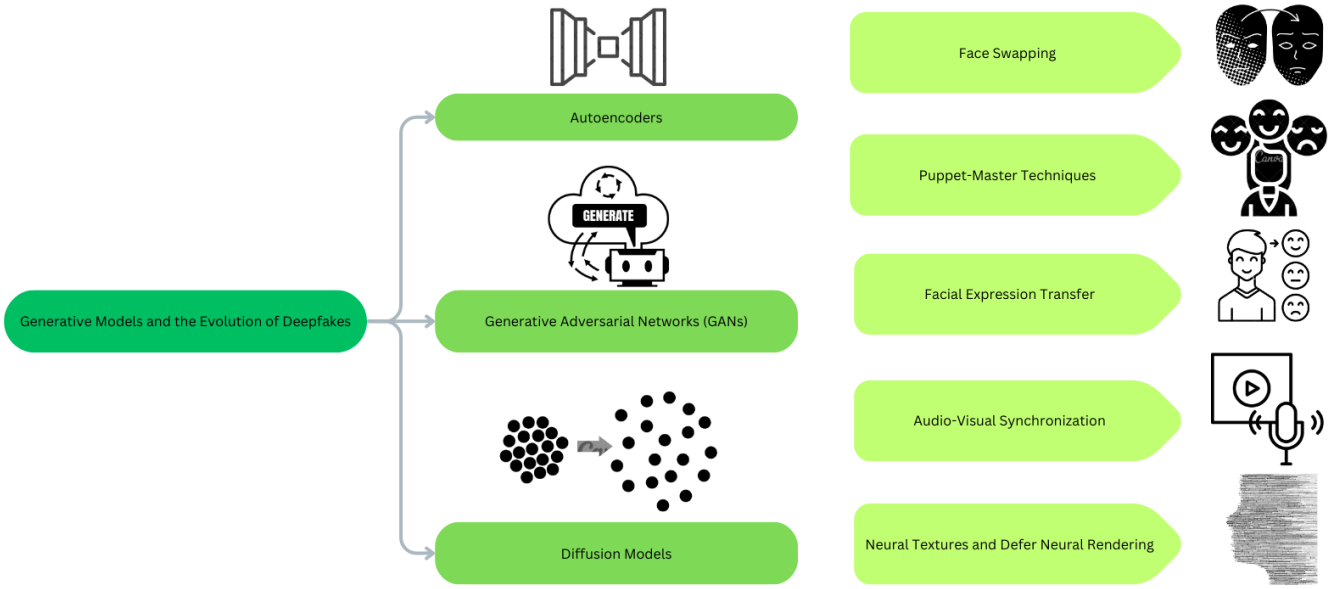

Figure 2. Deepfake generative models and key methods.

In addition to the visual-grade improvements, evaluating deepfake generation models' quality requires several key metrics. The MSE metric quantifies the pixel-wise difference between generated and ground truth images, with lower values indicating better quality. The Structural Similarity Index (SSIM) measures perceptual similarity between two images, while verification scores assess identity preservation through face verification. Additional metrics, such as pose and expression errors, evaluate the accuracy of facial pose and expression replication, providing a comprehensive evaluation of model performance [28,34]. Table 1 presents a comparative analysis of literature on deepfake generation techniques.

Table 1. Comparative overview of techniques for deepfake generation.

Study	Year	Approach	Performance	Dataset	Limitations
			Face Swap		
Bitouk et al. [35]	2008	Seamless integration of faces across varying poses, lighting conditions, and skin tones without requiring 3D models	Validated by a user study where 58% of replaced face images were misidentified as real	33,000 faces collected from public sources, filtered for quality and pose alignment	Dependency on the accuracy of face detection, potential mismatches in gender and age, and challenges in handling occlusions and extreme poses
Lin et al. [36]	2012	3D head model from a single frontal image	Subjective evaluations, outperformed or matched a commercial software solution in 80% of cases	Specific datasets were not detailed	Requirement for a single frontal image, the subjective nature of performance evaluation, the need for offline labeling of target face masks, and potential variability in color and illumination adjustments
Korshunova et al. [2]	2017	Style transfer via CNN, face alignment, multi-image style loss, integration of lighting conditions	Qualitative evaluation, focusing on visual fidelity and the preservation of key facial features	Nicolas Cage Images, CelebA	Lower performance with profile views compared to frontal views
Suwajanakorn et al. [14]	2017	RNN to synthesize photorealistic lip-sync video from audio	Phoneme overlaps averaged across the four targets were 99.9% (diphones), 82.9% (triphones), 35.7% (tetraphones), 12.1% (pentaphones), and 4.9% for five consecutive phonemes	17 h of Pres. Obama's weekly address videos	Requiring teeth proxy selection step as a manual process for each target video
Zhu et al. [37]	2017	Leveraging cycle consistency to learn mappings between two domains without the need for paired training examples	Per-pixel accuracy of 0.52, compared to 0.40 for CoGAN and 0.71 for the fully supervised pix2pix method	Cityscapes dataset, Imagenet, Flickr, WikiArt	High perceptual realism, but limited in new domains
Wiles et al. [38]	2018	Using various modalities, such as images, audio, and pose codes, without requiring explicit facial representations or annotations	Generating realistic facial expressions and poses while maintaining the identity of the source face	VoxCeleb dataset	Potential for lower generation quality compared to methods specifically designed for face transformation, dealing with unseen identities or significant variations in pose and expression

Table 1. *Cont.*

Study	Year	Approach	Performance	Dataset	Limitations
Wang et al. [39]	2018	Conditional GANs, new adversarial loss and multi-scale generator and discriminator architectures	Generating images at a resolution of 2048 × 1024, high pixel accuracy and mean intersection-over-union scores on the Cityscapes dataset	Cityscapes, NYU Indoor RGBD, ADE20K, and Helen Face	Reliance on high-quality training data, generating diverse outputs while maintaining realism
Nirking et al. [40]	2018	Standard fully convolutional network (FCN) for fast and accurate face segmentation	Evaluated on the Labeled Faces in the Wild (LFW) benchmark, demonstrating that intra-subject face swapping maintains high recognition accuracy while inter-subject swapping results in decreased recognizability	The IARPA Janus CS2 dataset	Reliance on accurate facial landmark localization
Choi et al. [41]	2018	Simultaneously train across multiple datasets, flexible domain translation, and mask vector introduction	Classification error of 2.12% with facial expression synthesis on the RaFD dataset	CelebA, RaFD Dataset	Performance limited by domain size
Natsume et al. [42]	2018	Variationally learning separate latent spaces for face and hair regions	Multi-scale structural similarity index (MS-SSIM) score of 0.760,	CelebA	Max. resolution of 128x128 pixels
He et al. [3]	2019	Attribute classification constraint, reconstruction learning, and adversarial learning	Average of 90.89% accuracy per attribute on CelebA testing set	CelebA, LFW (Labeled Faces in the Wild)	Attribute-Independent Constraint, Complex Attribute Changes
Nirkin et al. [34]	2019	Subject Agnostic Methodology, RNN for pose and expression adjustments, face completion network	SSIM of 0.51	IJB-C, VGGFace2, CelebA, FF++, LFW Parts Labels Dataset,	May struggle with large angular differences in facial poses
Natsume et al. [43]	2019	Deep generative model that disentangles face appearance as a latent variable, independent of face geometry and non-face regions	Achieving high scores in identity preservation and image quality metrics during evaluation on the CelebA dataset	CelebA	Handling occluded faces and the potential for inaccuracies in face region segmentation

Table 1. *Cont.*

Study	Year	Approach	Performance	Dataset	Limitations
Karras et al. [26]	2020	Redesign of generator normalization, path length regularization, progressive growing revisited	Improving the Fréchet Inception Distance (FID) score from 4.40 for the baseline StyleGAN to 2.84 for StyleGAN2	FFHQ (Flickr-Faces-HQ), LSUN (Large-Scale Scene Understanding)	Reliance on the perceptual path length (PPL) metric for assessing image quality
			Face Reenactment		
Thies et al. [44]	2016	Real-time facial reenactment using only monocular RGB video input	Live facial reenactment at an average frame rate of approximately 28.4 Hz	Monocular video sequences sourced from YouTube	Limited to specific reenactment applications
Siarohin et al. [45]	2019	Animate objects in still images based on driving videos using self-supervised learning, employing learned keypoints and local affine transformations	Better video reconstruction and user preference ratings	The Tai-Chi-HD, VoxCeleb, UvA-Nemo, and BAIR datasets	Assumption of similar poses between source and driving objects and reliance on large datasets for training
Thies et al. [28]	2019	Integration of traditional graphics rendering techniques with learnable components	MSE of 0.38 at a resolution of 2048 × 2048	Custom real and synthetic sequences, facial reenactment	Dependent on quality of 3D reconstructions
Wang et al. [46]	2020	Decomposing appearance and motion through a spatiotemporal fusion mechanism and a transposed (1+2)D convolution	Better quantitative and qualitative results compared to methods like VGAN and MoCoGAN	MUG, UvA-NEMO, NATOPS, and Weizmann	Reliance on high-quality input images and the challenge of generating videos with complex motions, or actions that were not well represented in the training datasets
Ha et al. [47]	2020	Addressing the identity preservation problem in face reenactment through components such as an image attention block, target feature alignment, and a landmark transformer	High scores in metrics like cosine similarity (CSIM) and masked peak signal-to-noise ratio (M-PSNR) across various experiments	VoxCeleb1 and CelebV datasets	Difficulty in properly disentangling identity and expression, especially for large pose variations and in one-shot settings

Table 1. *Cont.*

Study	Year	Approach	Performance	Dataset	Limitations
Lahiri et al. [48]	2021	Animating personalized 3D talking faces from audio, utilizing pose and lighting normalization	Achieving high realism, lip-sync accuracy, and visual quality, as evidenced by human ratings and objective metrics	GRID, CREMA-D, and TCD-TIMIT	Inability to explicitly handle facial expressions and potential issues with strong movements in target videos, the processing speed is slightly slower than real time
Zhou et al. [49]	2021	Lip synchronization and free pose control without relying on structural intermediate information	Excelling in lip-sync accuracy and robustness under extreme conditions	VoxCeleb2 and LRW datasets	Reliance on high-quality input data for effective pose control and challenges in scenarios with significant head pose variations or low-light conditions
Lu et al. [50]	2021	Generating personalized, photorealistic talking-head animations in real time, driven solely by audio input, achieving over 30 frames per second	User studies, preserving individual talking styles and generating high-fidelity facial details	Various video sequences of different subjects, totaling approximately 3 to 5 min each, with a focus on Mandarin Chinese audio	Challenges in accurately capturing plosive and nasal consonants, potential issues with fast speech, and dependency on the quality of the training corpus
			From Scratch		
Rombach et al. [32]	2022	Latent space utilization, separation of learning phases, cross-attention mechanism	FID score of 5.11, outperforming StyleGAN, which scored 4.16 on the CelebA-HQ dataset	CelebA-HQ, FFHQ, LSUN-Churches and LSUN-Bedrooms, MS-COCO	Slow computation compared to GANs
Bregler et al. [51]	2023	Automatically generate new video footage of a person mouthing words that were not originally spoken	Mean spatial registration error of just 1.0 pixels	Dataset comprising approximately 8 min of video containing 109 sentences, historic footage of John F. Kennedy	Accurately modeling large head rotations and the potential for lip fluttering due to mismatched triphone sequences

The computational sophistication required to create deepfake content is equally important, as the demands of generative modeling techniques can vary significantly, reflecting their architectural complexities and operational requirements. Autoencoders, particularly Variational Autoencoders (VAEs), generally require a low-to-moderate number of GPU resources, which makes them suitable for tasks such as basic image generation and representation learning [52]. These models are efficient for tasks where computational power is limited. In contrast, GANs demand significantly higher computational power due to their adversarial training process, which involves the simultaneous optimization of generator and discriminator networks. This process not only increases resource consumption but also introduces challenges like mode collapse, further compromising efficiency [52,53]. Diffusion models surpass both VAEs and GANs in terms of resource requirements, necessitating extensive computational power for training and inference. These models are capable of generating high-fidelity images with exceptional semantic coherence, but this performance comes at the cost of significantly higher GPU usage [52,53]. Despite their demanding nature, diffusion models have become increasingly prominent due to their ability to produce state-of-the-art results. While the resource-intensive nature of these generative modeling techniques can be a barrier to widespread adoption, ongoing advancements in optimization strategies and hardware efficiency aim to alleviate these challenges. These developments hold the potential to broaden the accessibility and application scope of generative models in various domains [54].

2.3. Applications of Deepfake Technology

Deepfake technology has found numerous positive applications across various fields. In entertainment and media, it is used for realistic video dubbing, enabling actors to appear as though they are speaking different languages [5]. Deepfakes also help create special effects, such as reanimating historical figures or deceased actors for new roles [18]. Applications like *Reface* allow users to swap their faces with celebrities, enhancing social media engagement [55]. In education, deepfakes bring historical figures to life, making learning more interactive and engaging. They can generate realistic video lectures featuring notable personalities [17], and aid language learning by exposing students to native speakers through video dubbing [18]. Deepfake technology also enhances interactive learning, allowing students to explore various roles or scenarios to understand complex topics better [56].

In healthcare, deepfakes create realistic simulations for medical training, allowing students to practice procedures on lifelike avatars. They also offer personalized avatars to help patients visualize treatment outcomes or provide therapeutic support. Telemedicine benefits from deepfakes through more interactive consultations with digital representations of healthcare professionals [18,57]. For Alzheimer's patients, animated representations of loved ones can provide comfort and emotional connection [58]. In marketing and advertising, brands use deepfakes to create personalized ads, enabling customers to visualize products, such as clothing, by superimposing their faces onto models, enhancing engagement and boosting sales [59].

Despite these positive uses, deepfake technology is often associated with harmful applications, particularly in spreading misinformation and fake news. It allows the creation of misleading videos that distort reality, contributing to false narratives [10,18]. Deepfakes of political figures can manipulate public perception and potentially influence elections [60]. They have also been used to create non-consensual pornographic content, causing severe emotional distress and reputational damage [7,60]. Additionally, deepfakes are exploited for harassment, blackmail, and other malicious purposes [10,18].

In politics, deepfakes can sabotage campaigns, mislead voters, or harm opponents' reputations. For instance, a fabricated video of a political leader calling for surrender could have serious national security implications [10]. Deepfakes also facilitate identity theft and financial fraud by generating fake identities. Scammers have used deepfake voices to impersonate individuals in financial schemes, such as a case where a deepfake CEO's voice led to a USD 243,000 scam [17,18,60,61]. Figure 3 exemplifies some positive and negative applications of deepfake technology.

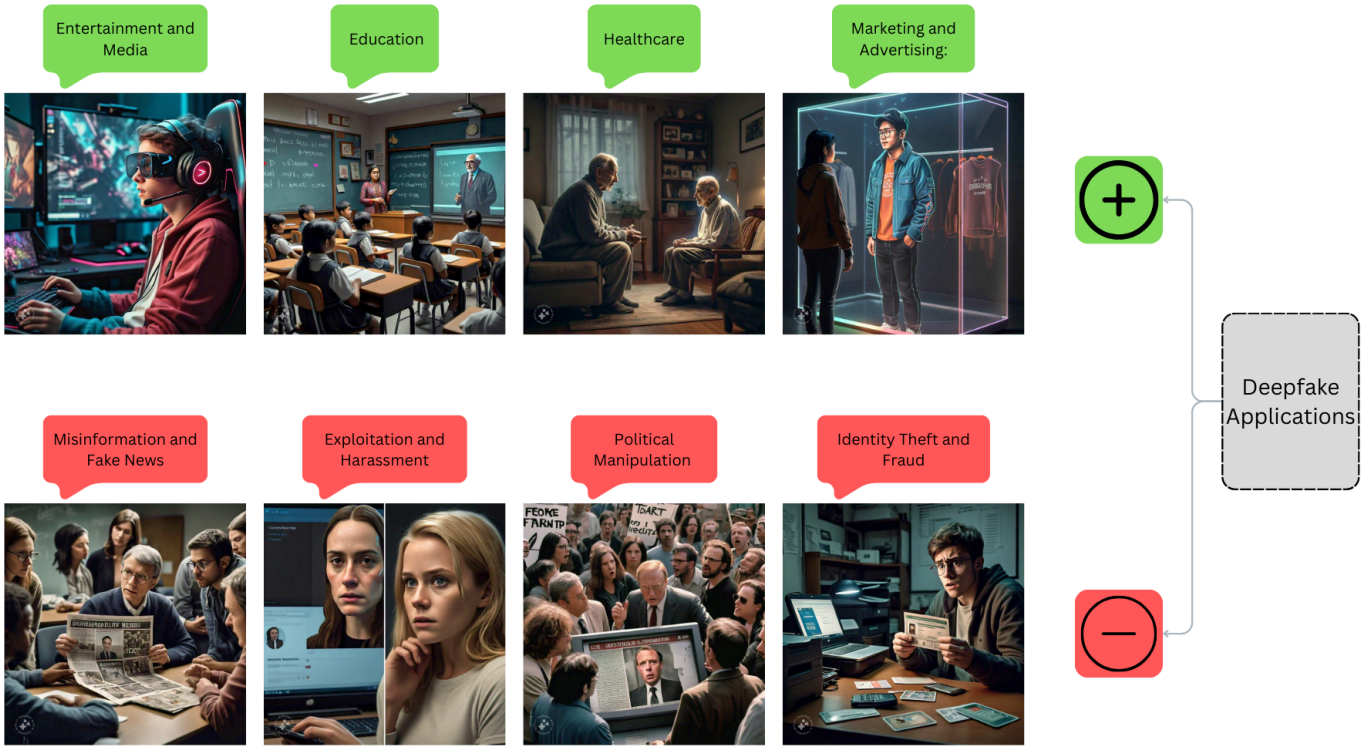

Figure 3. Beneficial and harmful applications of deepfakes.

2.4. Implications with Deepfake Technology

Addressing the challenges posed by harmful deepfakes primarily involves the timely and accurate detection of such content. Detection models often focus on identifying specific artifacts or patterns, but, as deepfake technology advances, these systems face increasing difficulties. Modern deepfakes can closely mimic authentic media while avoiding the indicators that detection algorithms typically rely on [62]. This is particularly problematic when adversarial attacks or AI-generated content are designed to evade detection [63]. Additional challenges include achieving high detection accuracy across diverse datasets, scalability for real-time applications, and adapting to evolving techniques [64]. Many models, particularly those using complex architectures like CNNs and GANs, suffer from overfitting on specific datasets, which limits their effectiveness for new types of deepfakes. Moreover, the lack of rigorous testing against adversarially crafted deepfakes can create a false sense of security about their robustness [65]. Ongoing research is essential to enhance the reliability and generalizability of detection methods.

Beyond technical challenges, deepfakes present significant ethical and societal concerns, particularly regarding privacy and consent. The ability to create unauthorized and potentially harmful representations of individuals, including public figures, infringes on privacy and raises concerns about consent and exploitation [66]. On a societal level, deepfakes threaten public trust in media and information sources. As deepfakes become more sophisticated, distinguishing between authentic and manipulated content becomes

increasingly difficult, potentially leading to a broader crisis of trust in media [67,68]. The normalization of deepfakes may desensitize individuals to media manipulation, altering perceptions of reality and authenticity. This could have long-term cultural and psychological consequences for social interactions and media consumption [55].

The ethical concerns mentioned highlight the need for strong guidelines to govern the responsible use of deepfake technology [69]. Comprehensive policies are necessary to mitigate risks while balancing innovation and free expression [69]. Governments and regulatory bodies can establish guidelines mandating transparency in AI systems used for deepfake detection. This includes ensuring that algorithms are explainable, helping users understand how decisions are made [20]. While some regions have begun implementing regulations, many areas, including India, lack comprehensive legislation [70,71]. Policies should also prioritize public education about deepfakes and detection technologies [72]. Increasing media literacy can empower individuals to critically evaluate content authenticity, reducing the impact of malicious deepfakes [73]. As noted in [10], enhancing public understanding of deepfake technology strengthens societal resilience against misinformation.

In conclusion, deepfake technology presents a dual-edged sword. While it offers creative and innovative opportunities, it also carries significant risks related to misinformation, exploitation, and security. As the technology evolves, effective detection methods and robust regulatory frameworks are crucial to mitigating its negative impacts [18,20].

3. Deepfake Detection Approaches, Countermeasures, and Evaluations

The evolution of deepfake technology, particularly with the rise of generative AI, has significantly reshaped the landscape of digital media manipulation. Generative methods have become increasingly sophisticated, producing highly realistic synthetic media that pose significant challenges for detection [74,75]. Despite these advancements, certain telltale signs can still reveal AI-generated content. For instance, visual artifacts, such as inconsistent lighting, unnatural textures, or blurriness around edges, and statistical irregularities in images can indicate tampering. Statistical patterns in real and synthetic images differ, serving as key indicators [76]. In video content, deepfakes may exhibit temporal inconsistencies, such as unnatural movements or discrepancies in frame transitions. Techniques like motion pattern analysis and shadows and lighting analysis can help identify these issues [62,77].

Detection techniques often exploit artifacts from the manipulation process by analyzing convolutional traces or combining visual and audio cues to keep up with the evolution of deepfake technology [61,78,79]. Tools such as Deeptrace and Amber Video leverage machine learning algorithms for real-time detection, while FooSpidy's Fake Finder utilizes image forensics and deep neural networks to identify manipulations in images and videos [57]. Recent advances integrate sophisticated deep learning models, including attention mechanisms and temporal analysis, to enhance detection capabilities. For instance, the Dual-Branch 3D Transformer (DuB3D), proposed by Ji et al. [80], combines spatial-temporal data with optical flow information to effectively distinguish genuine content from deepfakes.

However, newer generative AI models increasingly produce deepfakes that lack detectable manipulation traces, presenting a significant challenge. Liang [76] introduces the Natural Trace Forensics (NTF) approach, which focuses on the stable and homogeneous features inherent in real images rather than relying solely on differences between real and fake images. This method enhances the ability of detectors to generalize across diverse generative models, thereby improving performance on a wider range of deepfakes [76]. Future research may prioritize identifying inconsistencies arising directly from the generative process, shifting focus from artifacts, which might be absent in high-quality synthetic content [76].

3.1. Deepfake Detection Methods and Countermeasures

Broadly, deepfake detection methods can be categorized into three main approaches: forensic-based approaches, machine learning approaches, and hybrid approaches. In the following section, we discuss each strategy along with its effectiveness and limitations, supported by examples from the literature. Additionally, we explore potential countermeasures that have been employed to evade these detection strategies.

3.1.1. Forensic-Based Approaches

Forensic techniques in digital media analysis rely on signal processing methods to detect anomalies and manipulations, such as examining frequency components and biological signals. Traditional forensic methods often focus on statistical characteristics, such as frequency domain analysis, to detect tampering [81–83]. These approaches identify artifacts from the deepfake generation process, including resolution inconsistencies, warping artifacts, and visual anomalies such as lighting mismatches or unnatural facial movements [84–86]. Agarwal et al. [87] introduced a method that analyzes the static and dynamic properties of the human ear and mandible movements, which are mismatched in manipulated facial or audio elements, to detect deepfake videos effectively.

Statistical-based detection methods are particularly robust against pixel-level artifacts and lossy compression but face challenges with video data when advanced techniques obscure visible artifacts [56]. Attackers can manipulate frequency characteristics, introduce noise to hide anomalies (e.g., unnatural blinking), and use post-processing techniques like compression or filters to mask artifacts further [5,16]. Real-time facial expression modifications enhance deepfake realism, while "social media laundering" obscures traces of manipulation through compression and metadata alteration [88].

A promising forensic approach analyzes biological signals, such as heart rate variations, eye blinking patterns, and other physiological responses that are often poorly replicated in deepfakes [89–91]. Tools like FakeCatcher leverage spatial and temporal coherence in biological signals for detection [92]. Fernandez et al. [89] introduced a Neural Ordinary Differential Equations (Neural-ODE) model to predict heart rate variations, distinguishing deepfakes from real videos through discrepancies in heart rates. Hernandez et al. [90] utilized remote photoplethysmography (rPPG) to detect skin color changes, achieving high accuracy by combining spatial and temporal data. Recent work by Çiftçi et al. [91] analyzed spatiotemporal patterns in facial photoplethysmography (PPG) signals to not only detect deepfakes but also identify the generative models used.

Although biological-signal-based methods are effective, they are prone to errors in low-quality or compressed videos where signal fidelity is degraded. Furthermore, advancements in deepfake models may enable them to mimic natural physiological signals, including realistic eye blinking and facial expressions, complicating detection [56,66]. These ongoing advancements underscore the need for adaptive detection methods capable of addressing increasingly sophisticated manipulations.

3.1.2. Machine-Learning-Based Approaches

Machine learning techniques are advanced computational methods capable of autonomously extracting and analyzing features from data. Support Vector Machines (SVMs) are widely employed to classify high-dimensional data by finding the optimal hyperplane that separates classes. They effectively distinguish real from fake content in images and videos by analyzing features like facial landmarks and texture patterns, leveraging kernel functions to handle non-linear relationships [5,93]. RNNs, which specialize in processing sequential data, are adept at analyzing video content by capturing temporal dependencies, enabling detection of manipulations such as unnatural facial movements or irregular blink-

ing patterns across frames [57]. CNNs excel in detecting spatial inconsistencies and artifacts in images, achieving high accuracy on deepfake videos. For instance, XceptionNet achieved up to 98% prediction accuracy [18]. However, CNNs struggle with subtle manipulations in high-quality deepfakes. Capsule networks, with their dynamic routing mechanisms [94], address this limitation by preserving spatial relationships, effectively identifying anomalies like lighting mismatches and facial expression irregularities. Nguyen et al. [12,95] demonstrated a 10% performance improvement over CNNs on datasets like FF++. Despite their strengths, capsule networks are computationally intensive, limiting real-time applicability.

Recent advancements in attention mechanisms and spatiotemporal feature incorporation have enhanced deepfake detection capabilities by focusing on regions of interest in images or videos [83,96–98]. Zhao et al. [83] introduced spatial attention heads to enhance textural features and integrate low- and high-level features using attention maps. Zhu et al. [97] highlighted manipulated regions through supervised attention, improving the detection of subtle forgery patterns. Wang et al. [98] integrated convolutional pooling with re-attention mechanisms to enhance local and global feature extraction, significantly improving video deepfake detection. Other models like Two-Branch Recurrent Networks analyze temporal inconsistencies across frames [99], and Siamese Networks [60] compare camera noise inconsistencies between original and manipulated frames, offering novel detection approaches.

The effectiveness of machine learning models heavily depends on the quality and diversity of training datasets. Data-driven approaches rely on large datasets of real and manipulated media to recognize anomalies such as facial irregularities and temporal inconsistencies. Synthetic training data and transfer learning enable models to adapt to novel deepfake types. Augmenting datasets with transformations like rotations, scaling, and color adjustments enhances detection robustness. Using generative models like diffusion models instead of GANs strengthens detection systems against emerging deepfake techniques [76,100]. Zero-shot learning techniques allow detection models to identify deepfakes without prior examples, proving valuable in rapidly evolving contexts [101]. Approaches like commonality learning improve generalization by identifying intrinsic features distinguishing forgeries, enhancing detection across new deepfake types without retraining [102,103]. Transfer learning, leveraging pre-trained models like ResNet and DenseNet, has further improved detection performance [104].

Despite these advancements, machine learning models for deepfake detection face scalability and computational challenges. Two-Branch Recurrent Networks can be hindered by rapid motion or lighting changes, and Siamese Networks require extensive training data for generalization [60,99]. Feature-based methods often struggle to differentiate subtle manipulations in sophisticated deepfakes [63]. Generative models using adversarial techniques, like imperceptible noise addition, can produce deepfakes that evade detection by machine learning models [29,76,105].

3.1.3. Hybrid Approaches

Hybrid methods in deepfake detection combine multiple techniques to enhance accuracy and reliability. These approaches integrate deep learning models, such as CNNs and RNNs, with traditional signal processing techniques. CNNs are employed to capture spatial features in images, while RNNs focus on temporal aspects in videos [30,106,107]. These are often coupled with forensic techniques that detect inconsistencies in lighting, shadows, and other physical cues. This hybrid approach blends data-driven and rule-based insights to improve detection accuracy [93]. For instance, Gallagher et al. [108] proposed a dual-input architecture analyzing both images and their Fourier frequency decompositions, which outperformed traditional methods. Chen et al. [96] introduced an

attention-based face manipulation detection model, combining spatial domain features from facial semantic segmentation with frequency domain features via Discrete Fourier Transform to enhance generalization.

Multi-modal detection is another emerging approach that analyzes visual and audio components to identify inconsistencies, such as mismatches in lip movements and speech [72,109,110]. Mittal et al. [109] developed a method leveraging audio-visual modalities and emotional cues for deepfake detection, while Oorlof et al. [110] introduced a two-stage cross-modal learning technique using self-supervised representation learning and complementary masking strategies to improve accuracy. Integrating multiple data streams enables a holistic understanding of manipulated content, which is essential as deepfakes grow more sophisticated.

Combining spatial, temporal, and frequency-based methods can significantly enhance detection accuracy and robustness [111]. These diverse techniques allow for comprehensive analysis by addressing different aspects of media—spatial irregularities, temporal inconsistencies, and frequency anomalies—thereby improving the system's resilience against evolving deepfake methods. For instance, analyzing audio alongside visual content provides additional context, aiding in detection [64]. Integrating spatiotemporal features and physiological signals offers a more detailed understanding of the content [13]. Furthermore, advancements in natural language processing (NLP) can help analyze textual content associated with deepfake videos, linking visual and textual cues to enhance interpretability and detection [61].

Despite their benefits, hybrid approaches are complex and computationally demanding, limiting scalability for real-time applications [112]. Deepfake creators continuously develop countermeasures, such as synchronizing audio and video using sophisticated voice synthesis and high-quality video generation, making detection more challenging [16,18,64]. Additionally, attackers may exploit weaknesses in hybrid algorithms by introducing undetectable inconsistencies in video frames [86]. Large Language Models (LLMs) further complicate detection by generating coherent and contextually relevant text narratives for deepfake videos, creating more sophisticated content that evades detection algorithms [63].

Figure 4 presents the methodologies employed by each approach to effectively identify manipulated media, along with their respective classes, while Table 2 summarizes the studies for each outlined technique, including their year, approaches, performance, datasets, and limitations.

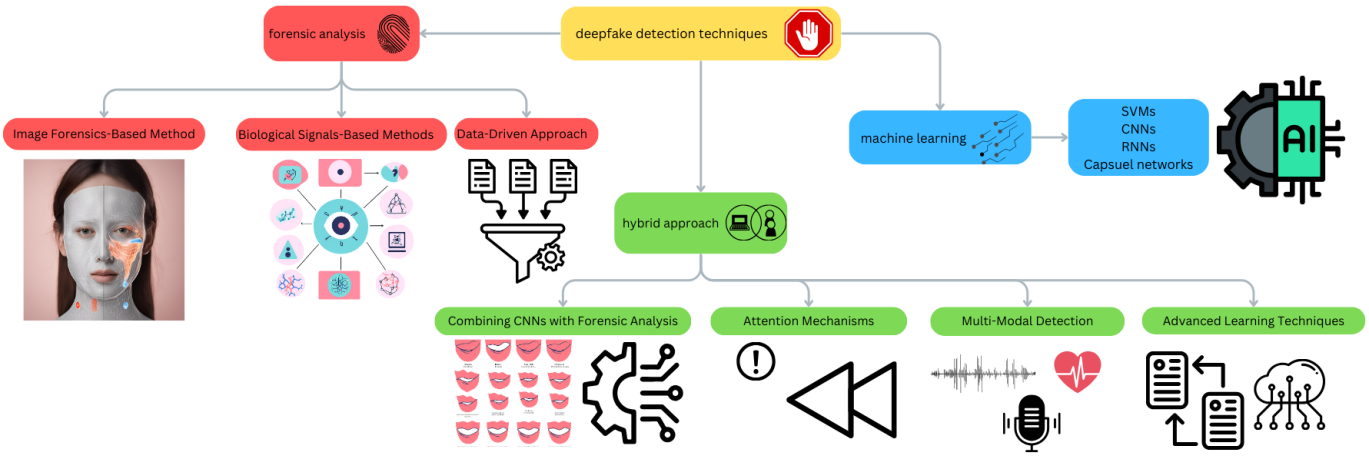

Figure 4. Overview of deepfake detection techniques.

Table 2. Comparative analysis of deepfake detection techniques.

Study	Year	Performance	Dataset	Approach	Limitations
			Forensic Techniques		
Matern et al. [113]	2019	AUC of up to 0.866 in classifying manipulated videos	CelebA, Deepfakes, Face2Face	Handcrafted features for classification	Varies across different datasets
Fernandes et al. [89]	2019	Neural-ODE model achieving average training losses of 0.010927 (original) and 0.010041 (donor)	COHFACE, DeepfakeTIMI, VidTIMIT	Heart rate extraction based on facial color variation	Requires precise video quality for accurate results
Agarwal et al. [114]	2020	Accuracies of 99.4% and 99.6% on original dataset; 83.7% and 71.1% on T2V-L dataset	Audio-to-Video (A2V), Text-to-Video (T2V), In-the-wild deep fakes	Phoneme–viseme mismatch analysis	Performance varies across datasets
Hernandez et al. [90]	2020	98.7% accuracy on Celeb-DF v2; 94.4% accuracy on DFDC Preview	Celeb-DF (v2), DFDC Preview	Heart rate estimation through rPPG	Dependence on video quality, susceptible to motion artifacts
Ciftci et al. [92]	2020	99.39% accuracy on Face Forensics; 91.07% on Deep Fakes Dataset	Face Forensics (FF), FF++, Celeb-DF, Deep Fakes	Leveraging biological signals as implicit descriptors of authenticity	Generalization and biological signal variability
Li et al. [102]	2020	AUC of 99.17 on unseen manipulations; 95.40 on FF dataset	FF++, DFD. DFDC, Celeb-DF	Face X-ray framework revealing the blending boundaries	Susceptible to low-resolution images and deepfakes created from scratch
Guarnera et al. [86]	2020	Up to 95% accuracy on image detection; 97% on video-based detection	Celen-A, Oxford-102, Cross-Age Celebrity Dataset (CACD), HOHA	Frequency analysis post Fourier transform	Robust to JPEG compression, blurring, and scaling
Luo et al. [111]	2021	The model trained on FF++ (HQ) achieved an AUC of 0.919 on DFD, 0.797 on DFDC, 0.794 on CelebDF, and 0.738 on DF1.0	FF++, DFD, CelebDF, DFDC, DF1.0	High-frequency noise features extraction	Lower accuracy than the F3Net [115] on FF++ dataset
Qian et al. [115]	2020	Accuracy of 90.43% and an AUC of 0.933 on LQ settings of FF++ with Xception backbone	FF++	Introducing two novel frequency-aware components: Frequency-Aware Decomposition (FAD) and Local Frequency Statistics (LFS)	Evaluations limited to FF++ dataset

Table 2. *Cont.*

Study	Year	Performance	Dataset	Approach	Limitations
Machine Learning Techniques					
Afchar et al. [116]	2018	Exceeds 98% detection for Deepfake, 95% for Face2Face	Deepfake, Face2Face	CNN-based network, Meso-4 and MesoInception-4	Performance affected by compression at low rates
Guera et al. [107]	2018	Classification accuracies of 99.5% on 20 frames	300 deepfake videos from multiple video-hosting websites + HOHA dataset [117]	Temporal-aware pipeline (CNN + RNN)	Limited to 2 s of video data
Li et al. [81]	2018	Achieving AUC scores of 97.4% for ResNet50 on the UADFV dataset and 99.9% on the low-quality set of DeepfakeTIMIT	UADFV, DeepfakeTIMIT	Targeting affine face warping artifacts	Varies by model; specific metrics provided
Nguyen et al. [29]	2019	93.63% accuracy with 7.20% EER for Deeper-FT; 92.77% accuracy with 8.18% EER for Proposed-New	FF and FF++	CNN with Y-shaped autoencoder	FT methods recorded lower accuracy and higher EER
Nguyen et al. [95]	2019	Accuracies of 89.57% for Real, 92.17% for Deepfakes, 90.00% for Face2Face, 92.79% for FaceSwap	FF++, CGI, and PI Dataset, Idiap's Replay-Attack Database	Capsule network for detecting deepfake videos	Applying capsule networks directly to time-series data (video) rather than just aggregating frames
De et al. [106]	2020	Significantly outperformed classical methods (66.8% accuracy) on Celeb-DF dataset with 98.26% accuracy	Kinetics dataset, Celeb-DF (v2)	Spatiotemporal convolution	Evaluated only on the Celeb-DF dataset
Guarnera et al. [78]	2020	Accuracies of 88.40% against GDWCT, 99.81% against STYLEGAN2 on CELEBA	CELEBA	Expectation Maximization for forensic trace detection	Adaptation to "wild" situations without prior knowledge of the image generation process
Nirkin et al. [118]	2020	99.7 and 66.0 AUC scores on FF-Deepfakes subset and Celeb-DF (v2), respectively	FF++, Celeb-DF (v2), DFDC	Analyzing discrepancies between the manipulated face and its surrounding context	Complexity of the contextual features around the face
Masi et al. [99]	2020	Video-level AUC of 76.65% on Celeb-DF; recall of 0.943 on DFDC preview set	FF++, Celeb-DF, DFDC preview	Two-Branch Recurrent Network with multi-scale LoG	Low performance on very low false alarm rates for practical, web-scale applications

Table 2. *Cont.*

Study	Year	Performance	Dataset	Approach	Limitations
Khalid et al. [15]	2020	95.30% accuracy on NeuralTextures; 98.00% accuracy on DFD	FF++, Deepfake Detection (DFD)	One-class classification using VAE	Relying on the RMSE function to compute the reconstruction score of images
Ismail et al. [119]	2021	90.73% accuracy, 90.62% AUC on CelebDF-FF++	CelebDF-FF++ (c23)	YOLO for face detection, InceptionResNetV2 for feature extraction, XGBoost for classification	Imbalance dataset, real-time applications
Zhu et al. [97]	2021	AUC of 98.73% with two-stream structure with halfway fusion on FF++; AUC of 66.09% when using the two-stream structure with halfway fusion on DFDC	FF++, DFD, DFDC	Two-stream network with 3D decomposition and supervised attention	The model's interpretability
Zheng et al. [21]	2021	AUC of 89.6% across four unseen datasets; video-level AUC of 99.7% on FF++ dataset	FaceShifter (FSh), FF++, DeeperForensics (DFo), DFDC, Celeb-DF (V2)	Fully Temporal Convolution Network (FTCN)	Low performance scores on Celeb-DF (V2) and DFDC datasets
Zhao et al. [83]	2021	97.60% accuracy on FF++ HQ; 67.44 AUC(%) cross-dataset evaluation on Celeb-DF by training on FF++	Celeb-DF, FF++	Multi-attentional framework for texture features	Sensitive to high compression rate
Czzolino et al. [120]	2021	80.4% accuracy and 0.91 AUC on high-quality videos DFDCp	VoxCeleb2, FF++, DFD, DFDCp	Low-dimensional 3D morphable model (3DMM)	Reliance on the availability of pristine reference videos of the target person
Das et al. [100]	2021	LogLoss of 0.178 and an AUC of 98.77% on FF++ with EfficientNet-B4 backbone	DFDC, FF++, Celeb-DF	Dynamic cutting of image regions based on landmarks	Lower performance benchmarks on Celeb-DF compared to Random Erase on EfficientNet-B4 backbone
Khormali et al. [22]	2022	Accuracies of 99.41% (FF), 99.31% (Celeb-DF V2), 81.35% (WildDeepfake)	FF, Celeb-DF (V2), WildDeepfake	End-to-end deepfake detection using transformers	Focused on facial area
Yu et al. [103]	2022	Outperformed SFFExtractors in cross-dataset evaluations	FF++, DFDC, Celeb-DF	Leveraging commonality learning to enhance generalization across various forgery methods	Assumption of common traces

Table 2. *Cont.*

Study	Year	Performance	Dataset	Approach	Limitations
Baraheem et al. [74]	2023	100% accuracy on Real or Synthetic Images (RSI) dataset	RSI, ADE20K, Sketchy, CUB-200-2011 Dataset	CNNs with EfficientNetB4 for GAN image detection	The model struggles with image classification errors, particularly with blurry, vintage, low-quality images, and those containing motion
Alnaim et al. [121]	2023	99.81% accuracy detecting face-mask deepfakes on Inception-ResNet-v2 model	Deepfake Face Mask Dataset (DFFMD)	Inception-ResNet-v2 architecture	The lack of video resources of humans wearing masks
Kingra et al. [60]	2023	99.7% accuracy on frame-levl FF++; 96.08% on DFD	FF++, DFD, Celeb-DF, DeeperForensics	Two-stream Siamese-like network	Degraded performance on lip-sync deepfakes
Xi et al. [4]	2023	93.1% accuracy on DALL-E2; 97.8% on DreamStudio	DALL-E2, DreamStudio	Text-to-image (T2I) detection	Limited to two T2I dataset benchmarks
Sawant et al. [65]	2023	Good sensitivity in distinguishing machine-generated content	Fake News Net, Machine Generated Fake News Dataset	Linear classifiers with TFIDF vectorization	Tracing the source of machine-generated fake news
Nadimpalli et al. [122]	2023	Enhances detection by human observers and state-of-the-art detectors	FF++, CelebA, Celeb-DF, RaDF	GAN-based visible watermarking	Vulnerability to cropping operations for watermark removal
Tang et al. [123]	2024	AUC of 1.0 for face replacement and 0.9999 for lip reenactment when the video quality was high (CRF = 23)	FF++, Celeb-DF	Embedding essential visual features into the DeepMark Meta (DMM) structure	Security relies on the robustness of the watermarking technology, inability to detect manipulations of features not explicitly protected
Jiang et al. [124]	2024	Near-perfect true detection rates (TDR) and attribution rates (TAR) above 0.94 for unprocessed images	Midjourney, DALLE 2	Watermark-based detection and attribution	Vulnerable to adversarial post-processing
Combs et al. [125]	2024	65% cosine similarity with true labels; 22.5% average classification accuracy	Caltech-256	Generative AI version of IRTARA	Dependence on pre-trained models, term frequency list limitations
Nanabala et al. [77]	2024	Over 95% accuracy in “GeneratedBy” classifier; 99.82% for Politics	Various domains, dividing into human or AI generated	Fake news detection	Difficulties in ethically creating fake news articles while adhering to ethical guidelines

Table 2. *Cont.*

Study	Year	Performance	Dataset	Approach	Limitations
Cao et al. [126]	2024	97.53% accuracy with 16.43% ACER (EfNet-b4)	FF-DFDC, FF-Celeb-DF	Dual space reconstruction learning framework	UniAttack benchmark's reliance on the existing datasets; susceptible to novel forgery types
Li et al. [127]	2024	Over 80 correct judgments on Kaggle true–false news dataset	Kaggle dataset	GPT-4 with web plugins for content authenticity assessment	The potential for biases and inaccuracies in the LLM assessments
Chakraborty et al. [128]	2024	High accuracy in real-time text analysis	Various datasets	Fine-tuned BERT model with preprocessing steps	Lack of detailed evaluations and comparison with other methods
Liang et al. [76]	2024	78.4% accuracy on Midjourney images	Midjourney, Self-Built Dataset	Natural Trace Forensics for fake image detection	Complexity of implementation
Wang et al. [104]	2024	DenseNet 97.74%, VGGNet 95.99%, ResNet 94.95%	CIFAKE	Machine learning to differentiate AI-generated from genuine images	Transfer learning dependence
Ji et al. [80]	2024	96.77% accuracy on GenVidDet dataset	GenVidDet	Dual-Branch 3D Transformer for motion and visual data	Motion modeling challenges, computational costs
Monkam et al. [62]	2024	94.14% accuracy on CelebA images	FFHQ, CelebA, and FF++	GAN for generating realistic images and identifying manipulated ones	Susceptible to adversarial attacks, incomplete comparison to state-of-the-art methods
Gallagher et al. [108]	2024	94% accuracy on CIFAKE dataset	CIFAKE	Dual-branch neural network using color images and DFT	Falling short in all metrics compared to VGGNet and DenseNet
Bai et al. [129]	2024	95.1% accuracy on Moonvalley; 91.1% on various datasets	Moonvalley generator, Large-scale generated video dataset (GVD), YouTube vos2 dataset	Two-branch spatiotemporal CNN	Dependence on quality of generated videos
Sun et al. [130]	2024	Boosting the AUC score on the unseen Celeb-DF dataset by 11% when integrated with the EfficientNet-B4	FF++, DeepFake Detection, DFDC Preview, WildDeepfake	Integrating a frozen pre-trained Stable Diffusion model to guide the forgery detector	Reliance on paired source and target images for training, which may limit the real-world applicability

Table 2. *Cont.*

Study	Year	Performance	Dataset	Approach	Limitations
			Hybrid Techniques		
Li et al. [75]	2018	Accuracy of over 90% for DCGANs (10,000 samples); over 94% for WGANs (nearing 99% with 100,000 pairs)	CelebA	Combines intrusive and non-intrusive approaches	Accuracy drops by 10% with mismatched datasets
Li et al. [131]	2018	AUC of 0.99, surpassing CNN's 0.98 and EAR's 0.79	CEW dataset, Eye Blinking Video (EBV) dataset	Long-Term Recurrent CNN (LRCN)	Sophisticated forgers can generate realistic blinking effects
Wang et al. [132]	2019	Over 90% average detection accuracy on four types of fake faces; AUC score of 66.8 on Celeb-DF (v2)	DFDC, FF++, Flicker-Faces-HQ (FFHQ), Celeb-DF (v2)	Neuron behavior monitoring with MNC criterion	Deteriorated performance on the DFDC deepfakes
Qi et al. [30]	2020	Top accuracies: 0.987 on DFD, 1.0 on DF, 0.995 on F2F	FF++, DFDC-Preview	Heartbeat rhythm analysis	Susceptible against specific adversarial attacks, worse performance than MesoNet on the DFDC dataset
Mittal et al. [109]	2020	Per-video AUC score of 96.6% on the DF-TIMIT dataset and 84.4% on the DFDC dataset	DFDC, DF-TIMIT	Audio-visual modality analysis	Misclassification due to different expressed perceived emotions, limited handling of multiple subjects in a video
Chintha et al. [27]	2020	100% accuracy on FF++; new benchmarks on DFDC-mini	FF++, DFDC-mini	Integration of edge and dense optical flow maps	Risk of adversarial attacks
Chen et al. [96]	2020	99.94% accuracy on Whole Face Forgery; 99.93% on FF	Whole Face Forgery, FF++	Joint spatial and frequency domain features	Limited region focus, model complexity
Rafique et al. [133]	2021	89.5% accuracy using ResNet18 feature vector and SVM classifier	Custom dataset compiled by Yonsei University's Computational Intelligence and Photography Lab	Integrated Error Level Analysis (ELA) with CNNs	Limited evaluation dataset

Table 2. *Cont.*

Study	Year	Performance	Dataset	Approach	Limitations
Sun et al. [59]	2021	90.50% accuracy on Meso-4Deepfake Database dataset; AUCs of 0.97 on DP23, 0.84 on F2F23	Meso4, DP23, F2F23	Edge geometric features and high-level semantic features	Performs less effectively on other datasets that require attention to subtle features like lips, eyes, and nose
Zhou et al. [134]	2021	69.4% accuracy on FFIW10K; AUC of 70.9%; and AUC of 99.3% on FF++	FFIW10K, FF++, Celeb-DF, DFDC	Multi-temporal-scale instance feature aggregation	Trained solely with video-level labels
Haliassos et al. [82]	2021	AUC of 95.1 on FaceSwap; outperforming existing methods on FaceShifter	DeeperForensics, FaceShifter, Celeb-DF (v2), DFDC	Spatiotemporal network pre-trained on lipreading	Lower scores on DFDC due to domain shifts
Khalil et al. [135]	2021	Frame-level AUC of 76.8 on DFDC-P; 91.70% video-level accuracy on Celeb-DF dataset	DFDC-P, Celeb-DF	Local Binary Patterns (LBP) and HRNet integration	Focus on visual manipulations
Kang et al. [9]	2022	Accuracies of 95.51% (face swap), 94.32% (puppet-master), 95.64% (attribute change), and 94.96% overall	DFDC, Celeb-DF, MegaFace, FF, ICFace, Glow, CelebA-HQ	SRNet for noise capture, landmark patch extraction	Prediction failures in cases of rapid facial movements or image overlaps during scene changes in video frames
Groh et al. [136]	2022	65% accuracy on 4000 videos; human participants outperformed model on political leaders	DFDC, Custom dataset	Evaluating deepfake detection through human–machine collaboration	Struggles with inverted videos
Wang et al. [98]	2023	65.76%, 63.27%, and 62.46% accuracy on DFDC, Celeb-DF, DF-1.0, respectively	DFDC, Celeb-DF, DF-1.0	Deep convolutional transformer	Likely to be fooled if the opponents intentionally raise the authenticity of the deepfake upon the keyframes
Haq et al. [61]	2023	87.5% accuracy on Presidential Deepfakes Dataset (PDD); 75.34% on WLD	PDD, World Leaders Dataset (WLD)	Emotional reasoning integration	Complexity of emotion recognition
Guan et al. [137]	2023	ACC of 0.644 and an AUC of 0.703 on the FF++-DFDC dataset	FF-DFDC, FF-Celeb-DF, Google DFD	Multi-feature channel domain-weighted framework	Fine-tuning results vary by dataset

Table 2. *Cont.*

Study	Year	Performance	Dataset	Approach	Limitations
Guo et al. [138]	2023	HTERs: 1.33 (HTERfs), 0.84 (HTERfg), 2.62 (HTERls), 4.66 (HTERlg)	GRID dataset	Lip-based visual speaker authentication	Evaluations are limited to the GRID dataset
Vismay et al. [139]	2023	89.99% accuracy on ML Olympiad dataset; 93.55% on CIFAKE	ML Olympiad, CIFAKE	Multi-modal method using text and image techniques	The model exhibited a slightly higher number of false positives and false negatives
Cciftcci et al. [91]	2024	97.29% deepfake detection accuracy; 93.39% source detection accuracy on FF++	FF++, Celeb-DF, FakeAVCeleb Dataset	Interpreted generative residuals through biological signals	Dependency on biological signals, demographic variability
Huang et al. [140]	2024	75.04% (text), 94.74% (image), 100.00% (voice)	COCO (Common Objects in Context), Flickr8K, Places205	Unified classification model with multi-modal embeddings	Challenges in cross-modality detection

The ongoing "arms race" between deepfake generation and detection technologies means that, as detection methods improve, so too will the sophistication of deepfake generation techniques [141]. This dynamic is discussed in [20], where the need for continuous innovation in detection methods to keep pace with evolving deepfake technologies is emphasized. The countermeasures against deepfake detection methods involve a combination of advanced generative techniques, strategic manipulation of content, and the use of sophisticated training methods to create deepfakes that are increasingly difficult to detect.

Targeted attacks aim to mislead the model into producing a specific incorrect output, such as modifying an image of a cat to be classified as a dog. These attacks are particularly concerning in applications where specific misclassifications can lead to significant consequences [142]. Conversely, untargeted attacks seek to degrade the model's overall performance without focusing on a specific output, aiming to cause general misclassifications and undermine model reliability [143].

Adversaries can continuously update their deepfake generation techniques based on the latest detection methods and combining various evasion techniques, ensuring that their content remains undetectable [20,112], like adversarial attacks, which involve subtly modifying input data, such as images or videos, to deceive detection algorithms [5,16,105]. Adversarial attacks on machine learning models are broadly categorized based on their characteristics and objectives, with each type posing unique challenges and necessitating tailored defense mechanisms. White-box attacks occur when attackers have complete knowledge of the model, including its architecture and parameters. This access enables precise manipulation of inputs to generate adversarial examples using methods such as the Fast Gradient Sign Method (FGSM) and Carlini and Wagner (C&W) attacks, which exploit model vulnerabilities to produce misleading inputs [143]. In contrast, black-box attacks operate without knowledge of the model's internal workings, relying instead on querying the model to infer its behavior. Although typically less precise and more complex than white-box attacks, black-box methods can still generate effective adversarial examples [143].

For instance, techniques such as FGSM can be used to generate adversarial examples that evade detection [5]. Adversarial techniques, such as distortion-minimizing and loss-maximizing attacks, can effectively mislead detection models [144]. Also, using knowledge from one model to improve the performance of another, through transfer learning, can help in creating deepfakes that are more difficult to detect using existing classifiers [18]. Moreover, generating content that changes dynamically can help evade detection. For example, using real-time facial reenactment techniques can create videos that adapt to the viewer's perspective, making it difficult for static detection algorithms to identify inconsistencies [63]. These countermeasures highlight the ongoing arms race between deepfake generation and detection technologies, where advancements in one area often lead to the development of new strategies in the other. Table 3 outlines several studies concerning adversarial attacks and the security concerns of deepfake detection models.

To counter these threats, several defense mechanisms have been developed. Gradient masking obscures the gradients used to craft adversarial examples, making it harder for attackers to exploit the model's weaknesses [143]. Ensemble methods, which combine multiple models, enhance robustness by ensuring that adversarial examples are less likely to fool all models simultaneously. Certified defenses provide formal guarantees of a model's resilience to adversarial inputs, offering a quantifiable level of security [142]. Additionally, adaptive adversarial training improves resilience by exposing the model to adversarial examples during the learning process, enhancing its ability to recognize and mitigate such inputs. While these strategies bolster model robustness, the rapidly evolving nature of

adversarial techniques demands ongoing research and innovation to safeguard machine learning applications against emerging threats.

Table 3. Overview of adversarial attacks and security measures in deepfake detection systems.

Study	Year	Performance	Dataset	Approach	Limitations
Ling et al. [145]	2019	Average 42.4% transferability rate of all attacks on the three models on CIFAR-10	MNIST, CIFAR-10	DEEPSEC incorporates 16 state-of-the-art attacks with 10 attack utility metrics and 13 state-of-the-art defenses with 5 defensive utility metrics	Employing one setting for each individual attack and defense, mainly focusing on non-adaptive and white-box attacks
Carlini et al. [146]	2020	Classifier performance drops significantly to AUC of 0.22 under the black-box attack	ProGAN images, real images	Vulnerability analysis under adversarial attacks	Harder to execute than real-world attacks
Hussain et al. [105]	2021	XceptionNet: 97.49%, MesoNet: 89.55% on FF++	FF++	Iterative gradient sign approach and Expectation over Transforms	Only focusing on deep neural networks
Cao et al. [23]	2021	Face classifiers' accuracies drop to nearly random guessing (i.e., 0.5) in cross-method settings	Six public datasets	Investigating vulnerabilities in detection systems	Only considered deepfake detection for a static face image
Aneja et al. [147]	2022	Takes only 77.89 ± 2.71 ms and 117.0 MB memory to compute the perturbation for a single image	CelebHQ, VGGFace2HQ	A novel attention-based fusion of manipulation-specific perturbations, only needing a single forward pass	Limited performance when evaluated with different compression qualities, while trained on a fixed quality
Panariello et al. [148]	2023	Increases EER significantly; highest EER 22.0 for RawNet2 CM	AASIST, RawNet2, and SSL countermeasures	Adversarial attack on ASV spoofing countermeasures	The performance of the integrated system that uses self-supervised learning countermeasures is reasonably robust against this attack
Zhong et al. [149]	2023	Underscoring the need for robust approaches to safeguard training sets and ensure provenance tracing	Various datasets	Evaluation framework for copyright protection measures	Solely focusing on GANs as generative models

3.2. Deepfake Detection Evaluations and Key Datasets

The effectiveness of deepfake detection, in contrast to deepfake creation, is heavily influenced by the availability of large-scale, high-quality training datasets. Existing datasets often face constraints such as resolution, compression, and post-processing artifacts, which limit their utility and impede the development of robust detection algorithms [88,144]. A significant challenge in advancing detection methods is the lack of comprehensive datasets that include both AI-generated and human-generated content [150]. Moreover, the scarcity of synthetic datasets worsens the difficulty of training models capable of distinguishing AI-generated content. Diverse datasets are essential for preparing models to address a wide range of deepfake scenarios, including hypothetical cases [77,112]. Efforts such as the introduction of datasets such as CIFAKE [104] and GenVidDet [80] have mitigated the limited resources available for training detection models. The release of a diffusion-generated deepfake speech dataset has facilitated further research and development in synthetic speech and deepfake detection [151]. These datasets simulate potential future deepfake techniques, enabling researchers to develop proactive models capable of identifying novel manipulation methods. Table 4 provides an overview of well-known video datasets used in deepfake detection, while Table 5 highlights notable image datasets. Additionally, Table 6 outlines relevant competitions in this domain.

Detection algorithms, while often excelling on known datasets, face significant challenges when confronted with unseen or novel types of deepfakes, raising concerns about their reliability in real-world scenarios [69]. For instance, models that achieve high accuracy rates (e.g., 94% on specific datasets) frequently exhibit significant performance variability when tested on deepfakes generated by different methods or under diverse conditions [108]. This variability highlights the critical need for standardized benchmarks and evaluation metrics to facilitate meaningful comparisons across detection techniques. As emphasized in [69], establishing common criteria for evaluating both the quality of deepfakes and the performance of detection systems is imperative. Such standardization would advance the field, ensuring the robustness and consistency of detection algorithms in practical applications.

Table 4. Overview of datasets utilized in deepfake detection research.

Dataset	Real Videos	Fake Videos	Year	Description
UADFV [152]	49	49	2018	Focus on head pose
EBV [131]	-	49	2018	Focus on eye blinking
Deepfake-TIMIT [153]	320	640	2018	GAN-based methods
DFFD [154]	1000	3000	2019	Multiple SOTA methods
DeepfakeDetection	363	3068	2019	Collected from actors with publicly available generation methods
Celeb-DF (v2) [31]	590	5639	2019	High quality
DFDC [19]	23,564	104,500	2019	DFDC competition on Kaggle
FF++ [155]	1000	5000	2019	Five different generation methods
FFIW-10K [134]	10,000	10,000	2019	Multiple faces in one frame
WLDR [156]	-	-	2019	Person of interest video from YouTube
DeeperForensics-1.0 [157]	50,000	10,000	2020	Add real-world perturbations
Wild-Deepfake [158]	3805	3509	2021	Collected from the Internet
ForgeryNet [159]	99,630	121,617	2021	8 video-level generation methods, added perturbations
FakeAVCeleb [160]	500	19,500	2021	Audio-visual multi-modal dataset
DeepSpeak [161]	6226	5958	2024	Lip-sync and face-swap deepfakes with audio manipulation

Table 5. Overview of image datasets employed in deepfake detection research.

Dataset	Real Images	Fake Images	Year	Description
DFFD [154]	58,703	240,336	2019	Multiple SOTA methods
iFakeFaceDB [162]	-	87,000 (StyleGAN)	2020	Generated by StyleGAN
100k Faces Website (accessed on 25 January 2025) https://generated.photos/datasets	-	100,000 (StyleGAN)	2021	Generated by StyleGAN
DFGC [163]	1000	N × 1000	2021	DFGC 2021 competition, fake images generated by users
ForgeryNet [159]	1,438,201	1,457,861	2021	7 image-level generation methods, added perturbations

Table 6. Overview of Competitions Focused on Deepfake Detection.

Name	Reference	Year	Description
Deepfake Detection Challenge	[164]	2019	1. Video-level detection. 2. The first worldwide competition. 3. More than 2000 teams participated.
DeepForensics Challenge	[165]	2020	1. Video-level detection. 2. Use DeeperForensics-1.0 datasets. 3. Simulates real-world scenarios.
Deepfake Game Competition	[166]	2021	1. Both image-level generation and video-level detection track. 2. Use Celeb-DF (v2) datasets.
Face Forgery Analysis Challenge	[167]	2021	1. Both image-level and video-level detection track. 2. Additional temporal localization track. 3. Use ForgeryNet dataset.

4. Discussion

The rapid advancement of deepfake generation techniques continually outpaces detection methods. As new deepfake generation methods emerge, the effectiveness of current detection algorithms diminishes, complicating the identification of manipulated content [144]. This trend is further underscored in a study by Firc et al. [10], which indicates that the quality of deepfake videos is approaching a level where they are nearly indistinguishable from real videos. The challenges facing detection systems are multifaceted, including the need for interpretability to build user trust, the necessity for real-time detection capabilities, the availability of diverse datasets for robust training, the establishment of effective evaluation protocols, and the development of regulatory frameworks. Together, these factors contribute to an escalating arms race between deepfake creation and detection technologies. Figure 5 highlights the key challenges and open research questions.

4.1. Computational Resource Requirements

The computational complexities of CNNs, RNNs, and transformers vary significantly, influencing their efficiency, scalability, and suitability for diverse applications. CNNs are particularly efficient for spatial feature extraction, leveraging convolutional and pooling layers to learn hierarchical spatial representations from image data. This results in relatively low computational complexity, as CNNs typically require fewer floating-point operations (FLOPs) for image classification tasks, making them well suited for real-time applications. In contrast, RNNs, including Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs), are designed to process sequential data but exhibit high memory consumption due to their recurrent nature, which demands storing long-term dependencies. This high memory footprint often leads to slower inference times, especially for applications involving lengthy sequences. Transformers, on the other hand, are computationally demanding, primarily due to their self-attention mechanisms, which scale quadratically with input size [168,169]. Despite their substantial resource requirements, transformers excel in handling large datasets and consistently achieve state-of-the-art performance in NLP tasks, making them the architecture of choice for complex, large-scale applications [170]. While CNNs and RNNs remain efficient for specific tasks, the unparalleled scalability and performance of transformers underscore the trade-offs between computational efficiency and model capability across different neural network architectures.

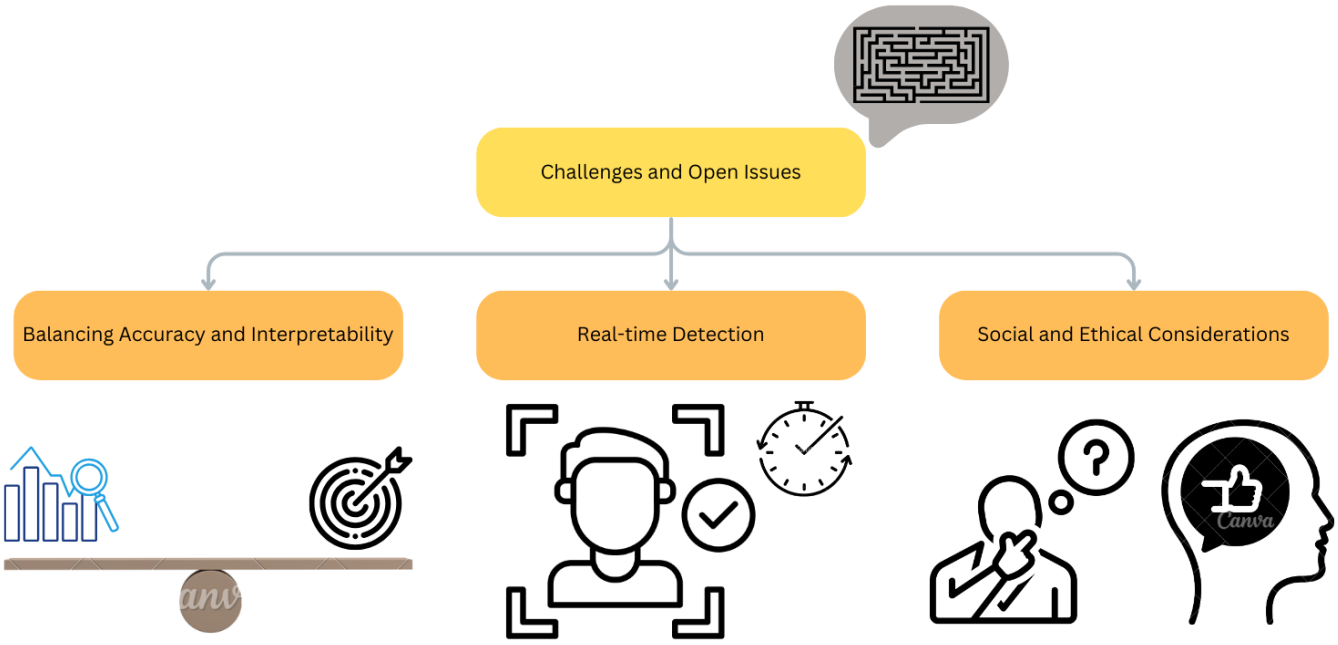

Figure 5. Key challenges and open research questions in deepfake detection.

4.2. Real-Time Detection Versus Post-Analysis

The development of real-time deepfake generation capabilities poses significant challenges to detection systems, as the rapid creation and dissemination of deepfakes can outpace the ability of existing methods to respond effectively [13]. Addressing this issue requires the implementation of efficient infrastructure and lightweight models that prioritize both interpretability and speed to manage the immense volume of content being uploaded, particularly on platforms such as social media and live broadcasting [64]. Timely detection in these contexts is crucial to curbing the spread of misinformation [119,122]. Leveraging technologies like edge computing and distributed systems can enable content analysis at the point of creation or sharing, enhancing the ability to intercept malicious content in real time [5].

Integrating real-time detection systems into platforms allows for a proactive approach to mitigating the impact of deepfakes [5,17]. However, these systems face limitations, including significant computational resource requirements that can lead to processing delays, making large-scale implementation challenging [171]. Furthermore, real-time detection methods often struggle with accuracy when confronted with high-quality deepfakes designed to evade detection. Environmental factors such as low lighting and visual obstructions can further degrade performance [16].

Conversely, post-analysis detection methods provide a reactive but detailed approach by examining content after its creation or dissemination. These methods benefit from more advanced algorithms and access to larger datasets, enabling techniques like pixel-level analysis and semantic feature extraction to identify even sophisticated deepfakes effectively [5,91]. While post-analysis methods tend to be more accurate, their reactive nature means that misinformation can spread before detection occurs [5]. Additionally, detection success often depends on the quality of the deepfake; lower-quality manipulations are generally easier to identify, while higher-quality ones may evade these systems [16].

Methods such as capsule networks and hybrid approaches achieve high accuracy in post-analysis scenarios but often require substantial computational resources, limiting their feasibility for real-time applications. For instance, RNNs are effective at analyzing temporal inconsistencies in videos but are computationally expensive, which hampers

their performance in environments demanding low latency, such as live streaming or social media [107].

Hybrid detection methods combine the strengths of both real-time and post-analysis approaches, offering a more robust and adaptive solution to the deepfake detection problem. These systems leverage continuous learning from new data to adapt to emerging types of deepfakes, improving their effectiveness over time [13,56]. Despite their promise, hybrid systems face challenges, including implementation complexity, significant resource requirements, and the risk of overfitting to specific datasets, which may reduce their generalizability to real-world scenarios with diverse deepfake characteristics [5,60]. Balancing computational efficiency with detection robustness remains a critical focus for future research.

4.3. Interpretability in Deep Fake Detection

In deepfake detection, achieving a balance between accuracy and interpretability is crucial to ensure effective performance while fostering user trust [93]. Interpretability refers to the ability of humans to understand an AI model's decisions, whereas explainability involves clarifying the mechanisms behind those decisions [20]. Highly accurate models, particularly those based on complex architectures like deep neural networks, often function as "black boxes", offering limited insights into their decision-making processes. This lack of transparency can hinder user trust and acceptance. Providing clear explanations of how detection systems identify manipulations is essential not only for maintaining effectiveness but also for gaining the confidence of users such as law enforcement personnel and social media moderators, who benefit from understanding why specific content is flagged as deepfake [18,68,85]. Techniques like layer-wise relevance propagation and fuzzy inference systems can help enhance the explainability of these frameworks [172].

Human experts, unlike machine learning models, often conduct a more holistic analysis when evaluating media. For example, they consider contextual factors such as the plausibility of the scenario, the subject's behavior, and the overall narrative, which can reveal inconsistencies that automated models may miss [147]. Additionally, incorporating psychological insights into human perception and emotion can lead to the development of detection systems that mimic human judgment, resulting in outputs that are more intuitive and interpretable for non-experts [61]. These approaches bridge the gap between machine and human interpretability, aligning automated systems with natural human reasoning.

Interpretability also plays a pivotal role in debugging and improving AI models. By enabling developers to identify weaknesses or biases, interpretability supports efforts to counter the rapidly evolving tactics used in deepfake creation [66]. Furthermore, detection systems are susceptible to adversarial attacks, where attackers exploit vulnerabilities in algorithms. For instance, Hussain et al. [105] demonstrated that adversarial attacks could reduce the accuracy of state-of-the-art models by up to 20%. To counteract these threats, techniques such as adversarial training and regularization are necessary to harden models against evasion attempts. Enhanced interpretability allows developers to pinpoint and address these vulnerabilities, thereby improving model robustness [105].

As deepfakes pose increasing risks, governments worldwide are implementing regulations to mitigate their negative impacts. For example, the U.S. has introduced laws criminalizing the malicious use of deepfake technology [66]. Ensuring compliance with these regulations necessitates explainable AI systems that promote accountability in automated decision-making [23]. Explainable models not only bolster transparency but also empower media professionals to understand the rationale behind detection results, thereby maintaining public trust. However, the enhanced transparency of these systems also exposes potential vulnerabilities, which malicious actors could exploit through adversarial

attacks. This dual-edged nature highlights the importance of striking a balance between interpretability, robustness, and security in deepfake detection frameworks.

4.4. The Impact of Generative AI on Deepfake Technology

Generative AI marks a revolutionary paradigm shift in computational intelligence, emerging from significant advances in machine learning, neural network architectures, and computational power. Unlike its predecessors, generative AI leverages sophisticated deep learning algorithms, particularly transformer models, which enable systems to learn from vast datasets through probabilistic modeling and contextual understanding [72]. The key distinguishing characteristics of generative AI lie in its dynamic learning capabilities, computational flexibility, and unprecedented content generation potential. Where traditional AI models were constrained to narrow, predefined tasks with minimal adaptability, generative AI demonstrates remarkable versatility across diverse domains. These advanced systems can analyze complex, unstructured datasets, continuously improve their performance through iterative learning, and generate novel content that reflects nuanced patterns learned from extensive training data. This technological evolution represents more than just an incremental improvement; it signifies a fundamental reimagining of AI's potential. Generative AI introduces a more adaptive, creative, and contextually responsive approach to problem-solving capable of generating intricate textual, visual, and auditory outputs that challenge previous limitations of computational creativity [63,173]. Generative AI opens unprecedented computational horizons. It embodies a learning-oriented model of artificial intelligence that closely mimics the human capacity for understanding, interpretation, and creative generation. This technological leap not only expands our conceptual understanding of machine intelligence but also raises profound philosophical and ethical questions about the nature of creativity, learning, and the potential symbiosis between human and artificial intelligence. As these technologies continue to evolve, they promise to reshape our understanding of intelligence, creativity, and the potential interfaces between human cognition and computational systems.

5. Conclusions

In conclusion, deepfake technology represents a double-edged sword, offering both innovative applications and significant risks to authenticity, privacy, and security. This literature review underscored the complexities inherent in developing effective detection methods that can combat the ever-evolving capabilities of deepfake generation. While advancements in machine learning and hybrid approaches have improved detection accuracy, challenges such as computational demands, dataset limitations, and susceptibility to adversarial attacks remain critical obstacles. Furthermore, the ethical and societal implications of deepfake technology necessitate urgent attention from policymakers and researchers alike. Table 7 summarizes the key findings and open research questions identified in this analysis. Future research must prioritize developing real-time detection approaches, enhancing the interpretability of detection models, and developing regulations that balance innovation with public safety. By addressing these challenges, the field can work toward more effective solutions that protect against the potential harms of deepfakes while harnessing their creative potential. Therefore the future of deepfake technology will be shaped by the ongoing advancements in deepfake generation techniques, the need for robust and adaptable detection methods, the establishment of regulatory frameworks, and the importance of public education. Addressing these challenges will require a collaborative effort among researchers, policymakers, and technology developers to ensure that detection systems remain effective and that society is protected from the potential harms of deepfakes.

Table 7. Summary of deepfake research findings and open questions.

Category	Main Findings	Open Research Questions
Deepfake Evolution	Generative AI models like GANs, transformers, and diffusion models have advanced deepfake realism and cross-modal capabilities.	How can detection systems keep pace with the rapid advancements in generative AI technologies?
Detection Methods	Detection approaches include forensic, machine learning, and hybrid techniques, leveraging artifacts, physiological signals, and multi-modal analysis.	What new techniques can enhance the adaptability and generalization of detection models for unseen deepfake types?
Challenges	Deepfake quality outpaces detection systems; adversarial attacks and limited generalization remain key issues.	How can detection algorithms remain robust to adversarial attacks and highly realistic deepfakes?
Applications	Positive uses include education, healthcare, and marketing; negative uses include misinformation, political sabotage, and privacy violations.	How can the benefits of deepfake technologies be promoted while minimizing harm in real-world applications?
Ethical and Regulatory	Strong ethical guidelines, public awareness, and robust policies are essential for balancing innovation and risk mitigation.	What are the most effective ways to implement global regulations and ethical frameworks for deepfake creation and detection?
Future Directions	Focus areas include real-time and cross-modal detection systems, adversarial robustness, standardized benchmarks, and interdisciplinary collaboration.	How can standardized datasets and evaluation benchmarks be designed to improve the reliability and scalability of detection algorithms across various applications?

Author Contributions: R.B. conducted the searches for relevant literature, synthesized key findings, analyzed the collected data, and wrote the initial drafts of the manuscript. S.C., R.D. and S.Z. provided essential feedback on the manuscript, contributed to the interpretation of findings, and assisted in revising the draft. All authors have read and agreed to the published version of the manuscript.

Funding: This work is partially supported by the Vice President for Research and Partnerships of the University of Oklahoma, the Data Institute for Societal Challenges, and the Stephenson Cancer Center through the DISC/SCC Seed Grant Award.

Data Availability Statement: All AI-assisted tools utilized in this review are publicly accessible, either as free resources or through subscription-based models. Additionally, the PDF versions of the studies and their corresponding results are publicly available through the publishers' online platforms.

Acknowledgments: During the preparation of this manuscript, the authors utilized ChatGPT to enhance content fluency and identify potential grammatical errors. The authors carefully reviewed and edited the content following the use of this tool, taking full responsibility for the final version of the manuscript. The figures presented in Figure 1 were generated using AI technology. These figures are computer-generated visualizations created solely for illustrative purposes and do not depict real-world entities or photographs. Additionally, automation tools such as Research Rabbit, Litmaps, and Connected Papers were employed to explore and trace connections between foundational studies and subsequent works. Data extraction was facilitated through the use of Docanalyzer.ai, with all extracted information rigorously validated against the original texts by the authors to ensure accuracy and reliability.

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflicts of interest.

References

1. Tolosana, R.; Vera-Rodriguez, R.; Fierrez, J.; Morales, A.; Ortega-Garcia, J. Deepfakes and beyond: A survey of face manipulation and fake detection. *Inf. Fusion* **2020**, *64*, 131–148. [CrossRef]
2. Korshunova, I.; Shi, W.; Dambre, J.; Theis, L. Fast face-swap using convolutional neural networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3677–3685.
3. He, Z.; Zuo, W.; Kan, M.; Shan, S.; Chen, X. Attgan: Facial attribute editing by only changing what you want. *IEEE Trans. Image Process.* **2019**, *28*, 5464–5478. [CrossRef]
4. Xi, Z.; Huang, W.; Wei, K.; Luo, W.; Zheng, P. Ai-generated image detection using a cross-attention enhanced dual-stream network. In Proceedings of the 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Taipei, Taiwan, 31 October–3 November 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1463–1470.
5. Mirsky, Y.; Lee, W. The creation and detection of deepfakes: A survey. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–41. [CrossRef]
6. Kane, T.B. Artificial intelligence in politics: Establishing ethics. *IEEE Technol. Soc. Mag.* **2019**, *38*, 72–80. [CrossRef]
7. Maras, M.H.; Alexandrou, A. Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos. *Int. J. Evid. Proof* **2019**, *23*, 255–262. [CrossRef]
8. Öhman, C. Introducing the pervert's dilemma: A contribution to the critique of Deepfake Pornography. *Ethics Inf. Technol.* **2020**, *22*, 133–140. [CrossRef]
9. Kang, J.; Ji, S.K.; Lee, S.; Jang, D.; Hou, J.U. Detection enhancement for various deepfake types based on residual noise and manipulation traces. *IEEE Access* **2022**, *10*, 69031–69040. [CrossRef]
10. Firc, A.; Malinka, K.; Hanáček, P. Deepfakes as a threat to a speaker and facial recognition: An overview of tools and attack vectors. *Heliyon* **2023**, *9*, e15090. [CrossRef]
11. Nah, F.-H.; Zheng, R.; Cai, J.; Siau, K.; Chen, L. Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *J. Inf. Technol. Case Appl. Res.* **2023**, *25*, 277–304.
12. Malik, A.; Kuribayashi, M.; Abdullahi, S.M.; Khan, A.N. DeepFake detection for human face images and videos: A survey. *IEEE Access* **2022**, *10*, 18757–18775. [CrossRef]
13. Masood, M.; Nawaz, M.; Malik, K.M.; Javed, A.; Irtaza, A.; Malik, H. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Appl. Intell.* **2023**, *53*, 3974–4026. [CrossRef]
14. Suwajanakorn, S.; Seitz, S.M.; Kemelmacher-Shlizerman, I. Synthesizing obama: Learning lip sync from audio. *ACM Trans. Graph. (ToG)* **2017**, *36*, 1–13. [CrossRef]
15. Khalid, H.; Woo, S.S. Oc-fakedect: Classifying deepfakes using one-class variational autoencoder. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 656–657.
16. Juefei-Xu, F.; Wang, R.; Huang, Y.; Guo, Q.; Ma, L.; Liu, Y. Countering malicious deepfakes: Survey, battleground, and horizon. *Int. J. Comput. Vis.* **2022**, *130*, 1678–1734. [CrossRef]
17. Nguyen, T.T.; Nguyen, Q.V.H.; Nguyen, D.T.; Nguyen, D.T.; Huynh-The, T.; Nahavandi, S.; Nguyen, T.T.; Pham, Q.V.; Nguyen, C.M. Deep learning for deepfakes creation and detection: A survey. *Comput. Vis. Image Underst.* **2022**, *223*, 103525. [CrossRef]
18. Passos, L.A.; Jodas, D.; Costa, K.A.; Souza Júnior, L.A.; Rodrigues, D.; Del Ser, J.; Camacho, D.; Papa, J.P. A review of deep learning-based approaches for deepfake content detection. *Expert Syst.* **2024**, *41*, e13570. [CrossRef]
19. Dolhansky, B.; Bitton, J.; Pflaum, B.; Lu, J.; Howes, R.; Wang, M.; Ferrer, C.C. The deepfake detection challenge (dfdc) dataset. *arXiv* **2020**, arXiv:2006.07397.
20. Jacobsen, B.N.; Simpson, J. The tensions of deepfakes. *Inf. Commun. Soc.* **2024**, *27*, 1095–1109. [CrossRef]
21. Zheng, Y.; Bao, J.; Chen, D.; Zeng, M.; Wen, F. Exploring temporal coherence for more general video face forgery detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 15044–15054.
22. Khormali, A.; Yuan, J.S. DFDT: An end-to-end deepfake detection framework using vision transformer. *Appl. Sci.* **2022**, *12*, 2953. [CrossRef]
23. Cao, X.; Gong, N.Z. Understanding the security of deepfake detection. In Proceedings of the International Conference on Digital Forensics and Cyber Crime, Virtual, 6–9 December 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 360–378.
24. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
25. Horvitz, E. On the horizon: Interactive and compositional deepfakes. In Proceedings of the 2022 International Conference on Multimodal Interaction, Bengaluru (Bangalore), India, 7–11 November 2022; pp. 653–661.
26. Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; Aila, T. Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8110–8119.

27. Chintha, A.; Rao, A.; Sohrawardi, S.; Bhatt, K.; Wright, M.; Ptucha, R. Leveraging edges and optical flow on faces for deepfake detection. In Proceedings of the 2020 IEEE International Joint Conference on Biometrics (IJCB), Houston, TX, USA, 28 September–1 October 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–10.
28. Thies, J.; Zollhöfer, M.; Nießner, M. Deferred neural rendering: Image synthesis using neural textures. *ACM Trans. Graph. (TOG)* **2019**, *38*, 1–12. [CrossRef]
29. Nguyen, H.H.; Fang, F.; Yamagishi, J.; Echizen, I. Multi-task learning for detecting and segmenting manipulated facial images and videos. In Proceedings of the 2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS), Tampa, FL, USA, 23–26 September 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–8.
30. Qi, H.; Guo, Q.; Juefei-Xu, F.; Xie, X.; Ma, L.; Feng, W.; Liu, Y.; Zhao, J. Deeprhythm: Exposing deepfakes with attentional visual heartbeat rhythms. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 4318–4327.
31. Li, Y.; Yang, X.; Sun, P.; Qi, H.; Lyu, S. Celeb-df: A large-scale challenging dataset for deepfake forensics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3207–3216.
32. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10684–10695.
33. Pei, G.; Zhang, J.; Hu, M.; Zhai, G.; Wang, C.; Zhang, Z.; Yang, J.; Shen, C.; Tao, D. Deepfake generation and detection: A benchmark and survey. *arXiv* **2024**, arXiv:2403.17881.
34. Nirkin, Y.; Keller, Y.; Hassner, T. Fsgan: Subject agnostic face swapping and reenactment. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7184–7193.
35. Bitouk, D.; Kumar, N.; Dhillon, S.; Belhumeur, P.; Nayar, S.K. Face swapping: Automatically replacing faces in photographs. *ACM Trans. Graph. (TOG)* **2008**, *27*, 1–8. [CrossRef]
36. Lin, Y.; Lin, Q.; Tang, F.; Wang, S. Face replacement with large-pose differences. In Proceedings of the 20th ACM International Conference on Multimedia, Nara, Japan, 29 October–2 November 2012; pp. 1249–1250.
37. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
38. Wiles, O.; Koepke, A.; Zisserman, A. X2face: A network for controlling face generation using images, audio, and pose codes. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 670–686.
39. Wang, T.C.; Liu, M.Y.; Zhu, J.Y.; Tao, A.; Kautz, J.; Catanzaro, B. High-resolution image synthesis and semantic manipulation with conditional gans. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8798–8807.
40. Nirkin, Y.; Masi, I.; Tuan, A.T.; Hassner, T.; Medioni, G. On face segmentation, face swapping, and face perception. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 98–105.
41. Choi, Y.; Choi, M.; Kim, M.; Ha, J.W.; Kim, S.; Choo, J. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8789–8797.
42. Natsume, R.; Yatagawa, T.; Morishima, S. Rsgan: Face swapping and editing using face and hair representation in latent spaces. *arXiv* **2018**, arXiv:1804.03447.
43. Natsume, R.; Yatagawa, T.; Morishima, S. Fsnet: An identity-aware generative model for image-based face swapping. In *Computer Vision—ACCV 2018, Proceedings of the 14th Asian Conference on Computer Vision, Perth, Australia, 2–6 December 2018*; Revised Selected Papers, Part VI 14; Springer: Berlin/Heidelberg, Germany, 2019; pp. 117–132.
44. Thies, J.; Zollhofer, M.; Stamminger, M.; Theobalt, C.; Nießner, M. Face2face: Real-time face capture and reenactment of rgb videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2387–2395.
45. Siarohin, A.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; Sebe, N. First order motion model for image animation. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Volume 32.
46. Wang, Y.; Bilinski, P.; Bremond, F.; Dantcheva, A. Imaginator: Conditional spatio-temporal gan for video generation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 1160–1169.
47. Ha, S.; Kersner, M.; Kim, B.; Seo, S.; Kim, D. Marionette: Few-shot face reenactment preserving identity of unseen targets. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 10893–10900.
48. Lahiri, A.; Kwatra, V.; Frueh, C.; Lewis, J.; Bregler, C. Lipsync3d: Data-efficient learning of personalized 3d talking faces from video using pose and lighting normalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2755–2764.

49. Zhou, H.; Sun, Y.; Wu, W.; Loy, C.C.; Wang, X.; Liu, Z. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 4176–4186.
50. Lu, Y.; Chai, J.; Cao, X. Live speech portraits: Real-time photorealistic talking-head animation. *ACM Trans. Graph. (ToG)* **2021**, *40*, 1–17. [CrossRef]
51. Bregler, C.; Covell, M.; Slaney, M. Video rewrite: Driving visual speech with audio. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*; Association for Computing Machinery: New York, NY, USA, 2023; pp. 715–722.
52. Vivekananthan, S. Comparative analysis of generative models: Enhancing image synthesis with vaes, gans, and stable diffusion. *arXiv* **2024**, arXiv:2408.08751.
53. Deshmukh, P.; Ambulkar, P.; Sarjoshi, P.; Dabhade, H.; Shah, S.A. Advancements in Generative Modeling: A Comprehensive Survey of GANs and Diffusion Models for Text-to-Image Synthesis and Manipulation. In Proceedings of the 2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), Bhopal, India, 24–25 February 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 1–8.
54. Lee, Y.; Sun, A.; Hosmer, B.; Acun, B.; Balioglu, C.; Wang, C.; Hernandez, C.D.; Puhrsch, C.; Haziza, D.; Guessous, D.; et al. Characterizing and Efficiently Accelerating Multimodal Generation Model Inference. *arXiv* **2024**, arXiv:2410.00215.
55. Bode, L.; Lees, D.; Golding, D. The digital face and deepfakes on screen. *Convergence* **2021**, *27*, 849–854. [CrossRef]
56. Altuncu, E.; Franqueira, V.N.; Li, S. Deepfake: Definitions, performance metrics and standards, datasets and benchmarks, and a meta-review. *arXiv* **2022**, arXiv:2208.10913. [CrossRef]
57. Mukta, M.S.H.; Ahmad, J.; Raiaan, M.A.K.; Islam, S.; Azam, S.; Ali, M.E.; Jonkman, M. An investigation of the effectiveness of deepfake models and tools. *J. Sens. Actuator Netw.* **2023**, *12*, 61. [CrossRef]
58. Kaur, A.; Noori Hoshyar, A.; Saikrishna, V.; Firmin, S.; Xia, F. Deepfake video detection: Challenges and opportunities. *Artif. Intell. Rev.* **2024**, *57*, 1–47. [CrossRef]
59. Sun, F.; Zhang, N.; Xu, P.; Song, Z. Deepfake Detection Method Based on Cross-Domain Fusion. *Secur. Commun. Netw.* **2021**, *2021*, 2482942. [CrossRef]
60. Kingra, S.; Aggarwal, N.; Kaur, N. SiamNet: Exploiting source camera noise discrepancies using Siamese network for Deepfake detection. *Inf. Sci.* **2023**, *645*, 119341. [CrossRef]
61. Haq, I.U.; Malik, K.M.; Muhammad, K. Multimodal neurosymbolic approach for explainable deepfake detection. *ACM Trans. Multimed. Comput. Commun. Appl.* **2023**, *20*, 1–16. [CrossRef]
62. Monkam, G.; Yan, J. Digital image forensic analyzer to detect AI-generated fake images. In Proceedings of the 2023 8th International Conference on Automation, Control and Robotics Engineering (CACRE), Guangzhou, China, 13–15 July 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 366–373.
63. Shree, M.S.; Arya, R.; Roy, S.K. Investigating the Evolving Landscape of Deepfake Technology: Generative AI's Role in it's Generation and Detection. *Int. Res. J. Adv. Eng. Hub (IRJAEH)* **2024**, *2*, 1489–1511. [CrossRef]
64. Kingsley, M.S.; Adithya, S.; Babu, B. AI Simulated Media Detection for Social Media. *Int. Res. J. Adv. Eng. Hub (IRJAEH)* **2024**, *2*, 938–943. [CrossRef]
65. Sawant, P. Neural Fake Det Net-Detection and Classification of AI Generated Fake News. In Proceedings of the CS & IT Conference Proceedings, CS & IT Conference Proceedings, Turku, Finland, 7–12 July 2023; Volume 13.
66. Zobaed, S.; Rabby, F.; Hossain, I.; Hossain, E.; Hasan, S.; Karim, A.; Md Hasib, K. Deepfakes: Detecting forged and synthetic media content using machine learning. In *Artificial Intelligence in Cyber Security: Impact and Implications: Security Challenges, Technical and Ethical Issues, Forensic Investigative Challenges*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 177–201.
67. Vaccari, C.; Chadwick, A. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Soc. Media+ Soc.* **2020**, *6*, 2056305120903408. [CrossRef]
68. Eberl, A.; Kühn, J.; Wolbring, T. Using deepfakes for experiments in the social sciences-A pilot study. *Front. Sociol.* **2022**, *7*, 907199. [CrossRef] [PubMed]
69. Akhtar, Z.; Pendyala, T.L.; Athmakuri, V.S. Video and Audio Deepfake Datasets and Open Issues in Deepfake Technology: Being Ahead of the Curve. *Forensic Sci.* **2024**, *4*, 289–377. [CrossRef]
70. Maniyal, V.; Kumar, V. Unveiling the Deepfake Dilemma: Framework, Classification, and Future Trajectories. *IT Prof.* **2024**, *26*, 32–38. [CrossRef]
71. Narayan, K.; Agarwal, H.; Thakral, K.; Mittal, S.; Vatsa, M.; Singh, R. Deephy: On deepfake phylogeny. In Proceedings of the 2022 IEEE International Joint Conference on Biometrics (IJCB), Abu Dhabi, United Arab Emirates, 10–13 October 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–10.
72. Naitali, A.; Ridouani, M.; Salahdine, F.; Kaabouch, N. Deepfake attacks: Generation, detection, datasets, challenges, and research directions. *Computers* **2023**, *12*, 216. [CrossRef]
73. Shahzad, H.F.; Rustam, F.; Flores, E.S.; Luis Vidal Mazon, J.; de la Torre Diez, I.; Ashraf, I. A review of image processing techniques for deepfakes. *Sensors* **2022**, *22*, 4556. [CrossRef]

74. Baraheem, S.S.; Nguyen, T.V. AI vs. AI: Can AI Detect AI-Generated Images? *J. Imaging* **2023**, *9*, 199. [CrossRef]
75. Li, H.; Chen, H.; Li, B.; Tan, S. Can forensic detectors identify gan generated images? In Proceedings of the 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Honolulu, HI, USA, 12–15 November 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 722–727.
76. Liang, Z.; Wang, R.; Liu, W.; Zhang, Y.; Yang, W.; Wang, L.; Wang, X. Let Real Images be as a Judger, Spotting Fake Images Synthesized with Generative Models. *arXiv* **2024**, arXiv:2403.16513.
77. Nanabala, C.; Mohan, C.K.; Zafarani, R. Unmasking AI-Generated Fake News Across Multiple Domains. *Preprints* **2024**. [CrossRef]
78. Guarnera, L.; Giudice, O.; Battiato, S. Deepfake detection by analyzing convolutional traces. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 666–667.
79. Nguyen, H.H.; Yamagishi, J.; Echizen, I. Capsule-forensics: Using capsule networks to detect forged images and videos. In Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 2307–2311.
80. Ji, L.; Lin, Y.; Huang, Z.; Han, Y.; Xu, X.; Wu, J.; Wang, C.; Liu, Z. Distinguish Any Fake Videos: Unleashing the Power of Large-scale Data and Motion Features. *arXiv* **2024**, arXiv:2405.15343.
81. Li, Y.; Lyu, S. Exposing deepfake videos by detecting face warping artifacts. *arXiv* **2018**, arXiv:1811.00656.
82. Haliassos, A.; Vougioukas, K.; Petridis, S.; Pantic, M. Lips don't lie: A generalisable and robust approach to face forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 5039–5049.
83. Zhao, H.; Zhou, W.; Chen, D.; Wei, T.; Zhang, W.; Yu, N. Multi-attentional deepfake detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2185–2194.
84. Xia, F.; Akoglu, L.; Aggarwal, C.; Liu, H. Deep Anomaly Analytics: Advancing the Frontier of Anomaly Detection. *IEEE Intell. Syst.* **2023**, *38*, 32–35. [CrossRef]
85. Verdoliva, L. Media forensics and deepfakes: An overview. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 910–932. [CrossRef]
86. Guarnera, L.; Giudice, O.; Nastasi, C.; Battiato, S. Preliminary forensics analysis of deepfake images. In Proceedings of the 2020 AEIT International Annual Conference (AEIT), Catania, Italy, 23–25 September 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–6.
87. Agarwal, S.; Farid, H. Detecting deep-fake videos from aural and oral dynamics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 981–989.
88. Roy, M.; Raval, M.S. Unmasking DeepFake Visual Content with Generative AI. In Proceedings of the 2023 IEEE 11th Region 10 Humanitarian Technology Conference (R10-HTC), Rajkot, India, 15 October 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 169–176.
89. Fernandes, S.; Raj, S.; Ortiz, E.; Vintila, I.; Salter, M.; Urosevic, G.; Jha, S. Predicting heart rate variations of deepfake videos using neural ode. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019.
90. Hernandez-Ortega, J.; Tolosana, R.; Fierrez, J.; Morales, A. Deepfakeson-phys: Deepfakes detection based on heart rate estimation. *arXiv* **2020**, arXiv:2010.00400.
91. Çiftçi, U.A.; Demir, İ.; Yin, L. Deepfake source detection in a heart beat. *Vis. Comput.* **2024**, *40*, 2733–2750. [CrossRef]
92. Ciftci, U.A.; Demir, I.; Yin, L. Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE Trans. Pattern Anal. Mach. Intell.* 2020, *early access*. [CrossRef]
93. Rana, M.S.; Nobi, M.N.; Murali, B.; Sung, A.H. Deepfake detection: A systematic literature review. *IEEE Access* **2022**, *10*, 25494–25513. [CrossRef]
94. Sabour, S.; Frosst, N.; Hinton, G.E. Dynamic routing between capsules. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
95. Nguyen, H.H.; Yamagishi, J.; Echizen, I. Use of a capsule network to detect fake images and videos. *arXiv* **2019**, arXiv:1910.12467.
96. Chen, Z.; Yang, H. Manipulated face detector: Joint spatial and frequency domain attention network. *arXiv* **2020**, arXiv:2005.02958.
97. Zhu, X.; Wang, H.; Fei, H.; Lei, Z.; Li, S.Z. Face forgery detection by 3d decomposition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2929–2939.
98. Wang, T.; Cheng, H.; Chow, K.P.; Nie, L. Deep convolutional pooling transformer for deepfake detection. *ACM Trans. Multimed. Comput. Commun. Appl.* **2023**, *19*, 1–20. [CrossRef]
99. Masi, I.; Killekar, A.; Mascarenhas, R.M.; Gurudatt, S.P.; AbdAlmageed, W. Two-branch recurrent network for isolating deepfakes in videos. In *Computer Vision—ECCV 2020, Proceedings of the 16th European Conference, Glasgow, UK, 23–28 August 2020*; Proceedings, Part VII 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 667–684.
100. Das, S.; Seferbekov, S.; Datta, A.; Islam, M.S.; Amin, M.R. Towards solving the deepfake problem: An analysis on improving deepfake detection using dynamic face augmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 3776–3785.
101. Yu, X.; Wang, Y.; Chen, Y.; Tao, Z.; Xi, D.; Song, S.; Niu, S. Fake Artificial Intelligence Generated Contents (FAIGC): A Survey of Theories, Detection Methods, and Opportunities. *arXiv* **2024**, arXiv:2405.00711.

102. Li, L.; Bao, J.; Zhang, T.; Yang, H.; Chen, D.; Wen, F.; Guo, B. Face x-ray for more general face forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5001–5010.
103. Yu, P.; Fei, J.; Xia, Z.; Zhou, Z.; Weng, J. Improving generalization by commonality learning in face forgery detection. *IEEE Trans. Inf. Forensics Secur.* **2022**, *17*, 547–558. [CrossRef]
104. Wang, Y.; Hao, Y.; Cong, A.X. Harnessing machine learning for discerning ai-generated synthetic images. *arXiv* **2024**, arXiv:2401.07358.
105. Hussain, S.; Neekhara, P.; Jere, M.; Koushanfar, F.; McAuley, J. Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 10–17 October 2021; pp. 3348–3357.
106. De Lima, O.; Franklin, S.; Basu, S.; Karwoski, B.; George, A. Deepfake detection using spatiotemporal convolutional networks. *arXiv* **2020**, arXiv:2006.14749.
107. Güera, D.; Delp, E.J. Deepfake video detection using recurrent neural networks. In Proceedings of the 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, 27–30 November 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–6.
108. Gallagher, J.; Pugsley, W. Development of a Dual-Input Neural Model for Detecting AI-Generated Imagery. *arXiv* **2024**, arXiv:2406.13688.
109. Mittal, T.; Bhattacharya, U.; Chandra, R.; Bera, A.; Manocha, D. Emotions don't lie: An audio-visual deepfake detection method using affective cues. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 2823–2832.
110. Oorloff, T.; Koppisetti, S.; Bonettini, N.; Solanki, D.; Colman, B.; Yacoob, Y.; Shahriyari, A.; Bharaj, G. AVFF: Audio-Visual Feature Fusion for Video Deepfake Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 27102–27112.
111. Luo, Y.; Zhang, Y.; Yan, J.; Liu, W. Generalizing face forgery detection with high-frequency features. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 16317–16326.
112. Sandotra, N.; Arora, B. A comprehensive evaluation of feature-based AI techniques for deepfake detection. *Neural Comput. Appl.* **2024**, *36*, 3859–3887. [CrossRef]
113. Matern, F.; Riess, C.; Stamminger, M. Exploiting visual artifacts to expose deepfakes and face manipulations. In Proceedings of the 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), Waikoloa Village, HI, USA, 7–11 January 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 83–92.
114. Agarwal, S.; Farid, H.; Fried, O.; Agrawala, M. Detecting deep-fake videos from phoneme-viseme mismatches. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 13–19 June 2020; pp. 660–661.
115. Qian, Y.; Yin, G.; Sheng, L.; Chen, Z.; Shao, J. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 86–103.
116. Afchar, D.; Nozick, V.; Yamagishi, J.; Echizen, I. Mesonet: A compact facial video forgery detection network. In Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Kowloon, Hong Kong, 11–13 December 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–7.
117. Laptev, I.; Marszalek, M.; Schmid, C.; Rozenfeld, B. Learning realistic human actions from movies. In Proceedings of the 2008 IEEE Conference on Computer Vision And Pattern Recognition, Anchorage, Alaska, 23–28 June 2008; IEEE: Piscataway, NJ, USA, 2008; pp. 1–8.
118. Nirkin, Y.; Wolf, L.; Keller, Y.; Hassner, T. Deepfake detection based on the discrepancy between the face and its context. *arXiv* **2020**, arXiv:2008.12262.
119. Ismail, A.; Elpeltagy, M.; S. Zaki, M.; Eldahshan, K. A new deep learning-based methodology for video deepfake detection using XGBoost. *Sensors* **2021**, *21*, 5413. [CrossRef]
120. Cozzolino, D.; Rössler, A.; Thies, J.; Nießner, M.; Verdoliva, L. Id-reveal: Identity-aware deepfake video detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 15108–15117.
121. Alnaim, N.M.; Almutairi, Z.M.; Alsuwat, M.S.; Alalawi, H.H.; Alshobaili, A.; Alenezi, F.S. DFFMD: A deepfake face mask dataset for infectious disease era with deepfake detection algorithms. *IEEE Access* **2023**, *11*, 16711–16722. [CrossRef]
122. Nadimpalli, A.V.; Rattani, A. ProActive deepfake detection using gan-based visible watermarking. *ACM Trans. Multimed. Comput. Commun. Appl.* **2023**, *20*, 1–27. [CrossRef]
123. Tang, L.; Ye, Q.; Hu, H.; Xue, Q.; Xiao, Y.; Li, J. DeepMark: A Scalable and Robust Framework for DeepFake Video Detection. *ACM Trans. Priv. Secur.* **2024**, *27*, 1–26. [CrossRef]
124. Jiang, Z.; Guo, M.; Hu, Y.; Gong, N.Z. Watermark-based Detection and Attribution of AI-Generated Content. *arXiv* **2024**, arXiv:2404.04254.
125. Combs, K.; Bihl, T.J.; Ganapathy, S. Utilization of generative AI for the characterization and identification of visual unknowns. *Nat. Lang. Process. J.* **2024**, *7*, 100064. [CrossRef]

126. Cao, J.; Zhang, K.Y.; Yao, T.; Ding, S.; Yang, X.; Ma, C. Towards Unified Defense for Face Forgery and Spoofing Attacks via Dual Space Reconstruction Learning. *Int. J. Comput. Vis.* **2024**, *132*, 5862–5887. [CrossRef]
127. Li, Y.; Wang, Z.; Papatheodorou, T. Staying vigilant in the Age of AI: From content generation to content authentication. *arXiv* **2024**, arXiv:2407.00922.
128. Chakraborty, U.; Gheewala, J.; Degadwala, S.; Vyas, D.; Soni, M. Safeguarding Authenticity in Text with BERT-Powered Detection of AI-Generated Content. In Proceedings of the 2024 International Conference on Inventive Computation Technologies (ICICT), Lalitpur, Nepal, 24–26 April 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 34–37.
129. Bai, J.; Lin, M.; Cao, G. AI-Generated Video Detection via Spatio-Temporal Anomaly Learning. *arXiv* **2024**, arXiv:2403.16638.
130. Sun, K.; Chen, S.; Yao, T.; Liu, H.; Sun, X.; Ding, S.; Ji, R. DiffusionFake: Enhancing Generalization in Deepfake Detection via Guided Stable Diffusion. *arXiv* **2024**, arXiv:2410.04372.
131. Li, Y.; Chang, M.C.; Lyu, S. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, China, 11–13 December 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–7.
132. Wang, R.; Juefei-Xu, F.; Ma, L.; Xie, X.; Huang, Y.; Wang, J.; Liu, Y. Fakespotter: A simple yet robust baseline for spotting ai-synthesized fake faces. *arXiv* **2019**, arXiv:1909.06122.
133. Rafique, R.; Nawaz, M.; Kibriya, H.; Masood, M. Deepfake detection using error level analysis and deep learning. In Proceedings of the 2021 4th International Conference on Computing & Information Sciences (ICCIS), Karachi, Pakistan, 29–30 November 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–4.
134. Zhou, T.; Wang, W.; Liang, Z.; Shen, J. Face forensics in the wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 5778–5788.
135. Khalil, S.S.; Youssef, S.M.; Saleh, S.N. iCaps-Dfake: An integrated capsule-based model for deepfake image and video detection. *Future Internet* **2021**, *13*, 93. [CrossRef]
136. Groh, M.; Epstein, Z.; Firestone, C.; Picard, R. Deepfake detection by human crowds, machines, and machine-informed crowds. *Proc. Natl. Acad. Sci. USA* **2022**, *119*, e2110013119. [CrossRef] [PubMed]
137. Guan, L.; Liu, F.; Zhang, R.; Liu, J.; Tang, Y. MCW: A Generalizable Deepfake Detection Method for Few-Shot Learning. *Sensors* **2023**, *23*, 8763. [CrossRef] [PubMed]
138. Guo, Z.; Wang, S. Content-Insensitive Dynamic Lip Feature Extraction for Visual Speaker Authentication Against Deepfake Attacks. In Proceedings of the ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–5.
139. Vora, V.; Savla, J.; Mehta, D.; Gawade, A. A Multimodal Approach for Detecting AI Generated Content using BERT and CNN. *Int. J. Recent Innov. Trends Comput. Commun.* **2023**, *11*, 691–701. [CrossRef]
140. Huang, L.; Zhang, Z.; Zhang, Y.; Zhou, X.; Wang, S. RU-AI: A Large Multimodal Dataset for Machine Generated Content Detection. *arXiv* **2024**, arXiv:2406.04906.
141. Mone, G. Outsmarting Deepfake Video. *Commun. ACM* **2023**, *66*, 18–19.
142. Khaleel, Y.L.; Habeeb, M.A.; Alnabulsi, H. Adversarial Attacks in Machine Learning: Key Insights and Defense Approaches. *Appl. Data Sci. Anal.* **2024**, *2024*, 121–147.
143. Ren, K.; Zheng, T.; Qin, Z.; Liu, X. Adversarial attacks and defenses in deep learning. *Engineering* **2020**, *6*, 346–360. [CrossRef]
144. Zhang, T. Deepfake generation and detection, a survey. *Multimed. Tools Appl.* **2022**, *81*, 6259–6276. [CrossRef]
145. Ling, X.; Ji, S.; Zou, J.; Wang, J.; Wu, C.; Li, B.; Wang, T. Deepsec: A uniform platform for security analysis of deep learning model. In Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 20–22 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 673–690.
146. Carlini, N.; Farid, H. Evading deepfake-image detectors with white-and black-box attacks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 658–659.
147. Aneja, S.; Markhasin, L.; Nießner, M. TAFIM: Targeted adversarial attacks against facial image manipulations. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 58–75.
148. Panariello, M.; Ge, W.; Tak, H.; Todisco, M.; Evans, N. Malafide: A novel adversarial convolutive noise attack against deepfake and spoofing detection systems. *arXiv* **2023**, arXiv:2306.07655.
149. Zhong, H.; Chang, J.; Yang, Z.; Wu, T.; Mahawaga Arachchige, P.C.; Pathmabandu, C.; Xue, M. Copyright protection and accountability of generative ai: Attack, watermarking and attribution. In Proceedings of the Companion Proceedings of the ACM Web Conference 2023, Austin, TX, USA, 30 April–4 May 2023; pp. 94–98.
150. Gong, L.Y.; Li, X.J. A contemporary survey on deepfake detection: Datasets, algorithms, and challenges. *Electronics* **2024**, *13*, 585. [CrossRef]
151. Firc, A.; Malinka, K.; Hanáček, P. Diffuse or Confuse: A Diffusion Deepfake Speech Dataset. In Proceedings of the 2024 International Conference of the Biometrics Special Interest Group (BIOSIG), Darmstadt, Germany, 25–27 September 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 1–7.

152. Yang, X.; Li, Y.; Lyu, S. Exposing deep fakes using inconsistent head poses. In Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 8261–8265.
153. Korshunov, P.; Marcel, S. Deepfakes: A new threat to face recognition? assessment and detection. *arXiv* **2018**, arXiv:1812.08685.
154. Dang, H.; Liu, F.; Stehouwer, J.; Liu, X.; Jain, A.K. On the detection of digital face manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5781–5790.
155. Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M. Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1–11.
156. Agarwal, S.; Farid, H.; Gu, Y.; He, M.; Nagano, K.; Li, H. Protecting World Leaders Against Deep Fakes. In Proceedings of the CVPR Workshops, Long Beach, CA, USA, 15–20 June 2019; Volume 1, p. 38.
157. Jiang, L.; Li, R.; Wu, W.; Qian, C.; Loy, C.C. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2889–2898.
158. Zi, B.; Chang, M.; Chen, J.; Ma, X.; Jiang, Y.G. Wilddeepfake: A challenging real-world dataset for deepfake detection. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 2382–2390.
159. He, Y.; Gan, B.; Chen, S.; Zhou, Y.; Yin, G.; Song, L.; Sheng, L.; Shao, J.; Liu, Z. Forgerynet: A versatile benchmark for comprehensive forgery analysis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual Conference, Nashville, TN, USA, 20–25 June 2021; pp. 4360–4369.
160. Khalid, H.; Tariq, S.; Kim, M.; Woo, S.S. FakeAVCeleb: A novel audio-video multimodal deepfake dataset. *arXiv* **2021**, arXiv:2108.05080.
161. Barrington, S.; Bohacek, M.; Farid, H. DeepSpeak Dataset v1. 0. *arXiv* **2024**, arXiv:2408.05366.
162. Neves, J.C.; Tolosana, R.; Vera-Rodriguez, R.; Lopes, V.; Proença, H.; Fierrez, J. Ganprintr: Improved fakes and evaluation of the state of the art in face manipulation detection. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 1038–1048. [CrossRef]
163. Peng, B.; Fan, H.; Wang, W.; Dong, J.; Li, Y.; Lyu, S.; Li, Q.; Sun, Z.; Chen, H.; Chen, B.; et al. Dfgc 2021: A deepfake game competition. In Proceedings of the 2021 IEEE International Joint Conference on Biometrics (IJCB), Shenzhen, China, 4–7 August 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–8.
164. Deepfake Detection Challenge. 2019. Available online: https://www.kaggle.com/c/deepfake-detection-challenge (accessed on 25 January 2025).
165. DeepForensics Challenge. 2020. Available online: https://competitions.codalab.org/competitions/25228 (accessed on 25 January 2025).
166. Deepfake Game Competition. 2021. Available online: https://competitions.codalab.org/competitions/29583 (accessed on 25 January 2025).
167. Face Forgery Analysis Challenge. 2021. Available online: https://competitions.codalab.org/competitions/33386 (accessed on 25 January 2025).
168. Shim, K.; Sung, W. A comparison of transformer, convolutional, and recurrent neural networks on phoneme recognition. *arXiv* **2022**, arXiv:2210.00367.
169. Lu, Z.; Wang, F.; Xu, Z.; Yang, F.; Li, T. On the performance and memory footprint of distributed training: An empirical study on transformers. *arXiv* **2024**, arXiv:2407.02081.
170. Panopoulos, I.; Nikolaidis, S.; Venieris, S.I.; Venieris, I.S. Exploring the Performance and Efficiency of Transformer Models for NLP on Mobile Devices. In Proceedings of the 2023 IEEE Symposium on Computers and Communications (ISCC), Tunis, Tunisia, 9–12 July 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–4.
171. Heidari, A.; Jafari Navimipour, N.; Dag, H.; Unal, M. Deepfake detection using deep learning methods: A systematic and comprehensive review. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2024**, *14*, e1520. [CrossRef]
172. Akhtar, Z. Deepfakes generation and detection: A short survey. *J. Imaging* **2023**, *9*, 18. [CrossRef] [PubMed]
173. Lee, H.; Lee, C.; Farhat, K.; Qiu, L.; Geluso, S.; Kim, A.; Etzioni, O. The Tug-of-War Between Deepfake Generation and Detection. *arXiv* **2024**, arXiv:2407.06174.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.