Multimodal Large Models, Deepfake Multimedia, Generative Adversarial Networks, Synthetic Media, and AI Ethics: A Survey

www.surveyx.cn

Abstract

This survey explores the multifaceted domain of multimodal large models, deepfakes, generative adversarial networks (GANs), synthetic media, and AI ethics, emphasizing their profound impact on media integrity and societal trust. It highlights the significant threats posed by deepfakes, necessitating the development of robust detection methods capable of adapting to rapid advancements in deepfake creation. Current detection techniques require continuous refinement to effectively address the complexities of real-world media. A critical gap identified is the scarcity of diverse datasets and the need for detection techniques that generalize across various large artificial intelligence models (LAIMs). The survey underscores the essential role of robust datasets in training effective detection models and emphasizes the importance of developing models capable of generalizing across multiple modalities and tasks. It also identifies promising advancements in audio deepfake detection, suggesting a promising direction for future research. Additionally, the integration of audio analysis into detection frameworks is highlighted as a significant enhancement for misinformation detection. The survey concludes by stressing the importance of addressing challenges posed by multimodal models, deepfakes, and synthetic media to maintain media integrity and public trust. Key takeaways include the necessity for continuous adaptation of detection models, the importance of multimodal approaches, and the potential for active authentication methods to enhance media integrity. By fostering interdisciplinary collaboration and adhering to ethical guidelines, the global community can navigate the complexities of AI-generated content and safeguard the digital landscape.

1 Introduction

1.1 Motivations for the Survey

The rapid advancement of generative artificial intelligence and the pervasive use of deepfake media have sparked significant ethical and moral dilemmas, necessitating a detailed analysis to distinguish between authentic, deepfake, and synthetic images [1]. This survey responds to the urgent need to tackle the challenges and innovations in detecting and creating deepfake content, which increasingly generates realistic but manipulated visuals [2]. The escalating threat posed by such media demands a fundamental reevaluation of media authentication practices to ensure the integrity of information [3].

The issue of deepfake attribution, particularly as generative models like GANs produce synthetic media capable of influencing social discourse, underscores the necessity for enhanced detection and attribution methodologies [4]. The implications of deepfake technology extend into the metaverse, where their potential applications are accompanied by notable security and privacy concerns [5]. Moreover, the transformative potential of Multimodal Large Language Models (MLLMs) in enriching educational experiences highlights the importance of comprehending the capabilities and ethical considerations associated with these technologies [6].

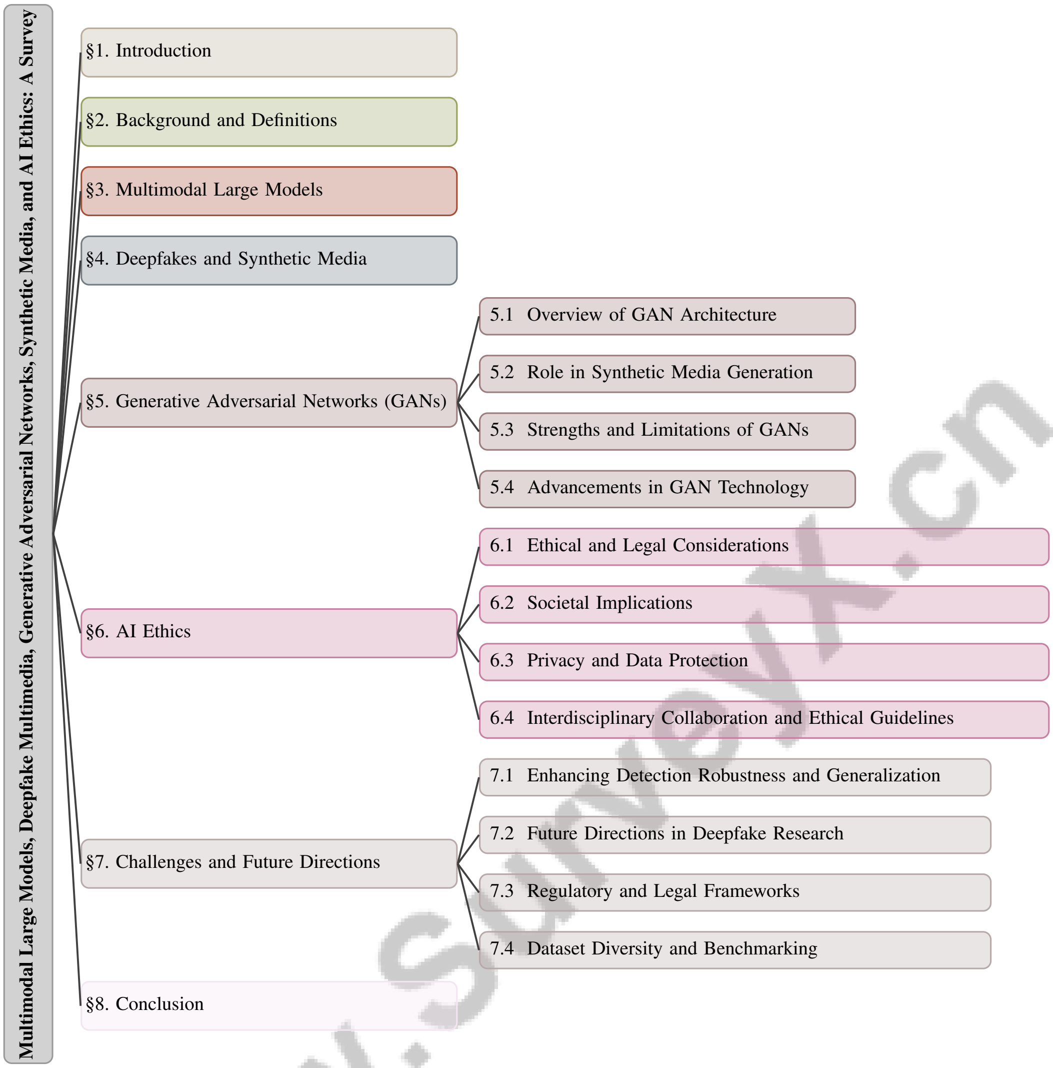

Multimodal Large Models, Deepfake Multimedia, Generative Adversarial Networks, Synthetic Media, and AI Ethics: A Survey
§1. Introduction
§2. Background and Definitions
§3. Multimodal Large Models
§4. Deepfakes and Synthetic Media
§5. Generative Adversarial Networks (GANs)
5.1 Overview of GAN Architecture
5.2 Role in Synthetic Media Generation
5.3 Strengths and Limitations of GANs
5.4 Advancements in GAN Technology
§6. AI Ethics
6.1 Ethical and Legal Considerations
6.2 Societal Implications
6.3 Privacy and Data Protection
6.4 Interdisciplinary Collaboration and Ethical Guidelines
§7. Challenges and Future Directions
7.1 Enhancing Detection Robustness and Generalization
7.2 Future Directions in Deepfake Research
7.3 Regulatory and Legal Frameworks
7.4 Dataset Diversity and Benchmarking
§8. Conclusion

Figure 1: chapter structure

This survey aims to offer a thorough overview of facial manipulation techniques and their detection methods, addressing the societal implications of realistic fake content generated through deep learning [7]. The evolution of deepfake generation and detection presents critical issues, particularly given the serious social and criminal concerns stemming from the realistic nature of deepfake media [8]. The proliferation of social media and multimedia content has intensified the spread of misinformation, necessitating effective multimodal misinformation detection strategies [9].

By synthesizing and analyzing pertinent literature on these technologies, this survey seeks to establish a foundational framework for future research, policy development, and ethical guidelines in AI ethics and technology governance. Investigating effective detection and attribution methods for AI-generated synthetic media is vital for addressing the ethical and security challenges these technologies present, as they can be exploited to manipulate information and erode trust in digital content. This research aims to enhance understanding of human detection capabilities and foster the development of robust detection strategies, ultimately contributing to a safer and more reliable digital environment, thereby bolstering public confidence in multimedia authenticity [10, 11, 12, 4, 13].

1.2 Structure of the Survey

This survey is organized to provide a comprehensive examination of the complex landscape encompassing multimodal large models, deepfake multimedia, generative adversarial networks (GANs),

AI-generated, for reference only.

synthetic media, and AI ethics. It begins with an introduction that outlines the motivations for exploring these topics, emphasizing the importance of addressing the ethical and technical challenges associated with these technologies [4].

The second section, "Background and Definitions," delves into the foundational concepts and definitions essential for understanding the subsequent discussions. It includes an in-depth exploration of deepfake definitions, performance metrics, and standards, alongside relevant datasets, challenges, competitions, and benchmarks in deepfake technology [14]. Additionally, this section addresses the intersection of multimedia and AI, focusing on multimedia intelligence and machine learning applications [15].

Following this, the survey examines "Multimodal Large Models," highlighting their significance and the challenges encountered in creating unified models. This section explores the impact of these models on synthetic media generation and their applications in education and media content creation [16].

The fourth section, "Deepfakes and Synthetic Media," investigates the creation and implications of deepfakes, concentrating on the technologies employed and the challenges they present regarding authenticity and misinformation. It also discusses the societal impact of deepfakes, particularly in the dissemination of misinformation and the difficulties associated with their detection [17].

The fifth section, "Generative Adversarial Networks (GANs)," provides an overview of GAN architecture and its role in generating synthetic media. It analyzes the strengths and limitations of GANs, while also highlighting recent advancements in the field.

The sixth section, "AI Ethics," explores the ethical considerations and responsibilities tied to AI technologies. It examines the ethical and legal issues surrounding AI-generated media, societal implications, privacy concerns, and the necessity for interdisciplinary collaboration and ethical guidelines [18].

Finally, the survey concludes with a discussion on "Challenges and Future Directions," identifying current challenges in AI-generated media and proposing future research trajectories. This section emphasizes the need for enhanced detection robustness, exploration of new avenues in deepfake research, and the implementation of regulatory measures alongside diverse datasets for benchmarking [19].

By employing a comprehensive and structured methodology, this survey aims to deliver an in-depth analysis of the current landscape and future trajectories of transformative technologies, particularly in the realms of AI and synthetic media. It seeks to equip researchers, policymakers, and practitioners with essential insights and strategies for effectively navigating the multifaceted challenges posed by AI-generated content, including detection issues, ethical implications, and societal impacts, as highlighted by recent studies on human perception and detection methodologies across various media types [20, 12, 13, 21].The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Definitions and Core Concepts

The advancement of AI technologies demands a clear understanding of core concepts that define AI-generated media and its ethical implications. This section clarifies these foundational ideas, including multimodal large models, deepfakes, multimedia, generative adversarial networks (GANs), synthetic media, and AI ethics.

Multimodal large models are sophisticated AI systems that integrate and process diverse data types such as text, audio, and images to produce contextually enriched content. They utilize techniques like Multimodal Chain of Thought, Multimodal Instruction Tuning, and Multimodal In-Context Learning [22], enabling applications like visual narratives and multi-view image generation from single captions [23]. The realism of these outputs poses challenges in verifying multimedia authenticity [3].

Deepfakes represent a significant leap in AI content manipulation, using deep learning and GANs to create highly realistic yet deceptive media [2]. The MFMC benchmark emphasizes generating quantitatively dissimilar faces while maintaining essential attributes [24]. This technology enhances

AI-generated, for reference only.

user experiences in the metaverse but also presents risks like identity theft [5]. Benchmarks are crucial for distinguishing real, deepfake, and synthetic images [1].

Multimedia integrates various content forms cohesively, involving text, audio, and visuals. Multimedia Intelligence focuses on the interplay between AI and multimedia, requiring robust methods to detect forgeries in image and text modalities [15]. Benchmarks for classifying (Image, Caption) pairs as truthful or falsified address misinformation challenges [9]. Advances like JPEG AI complicate image forensics, posing challenges for authenticity verification [25].

Generative adversarial networks consist of a generator creating new data instances and a discriminator evaluating them, advancing synthetic media realism and complicating the distinction between genuine and AI-generated content [4]. This duality necessitates examining societal and privacy risks [26].

Synthetic media, including deepfakes and GAN-generated content, mimics real-world media. It can reduce workloads in areas like 3D model editing [27], yet its dual nature raises ethical concerns [26].

AI ethics provides a framework addressing moral implications and responsibilities of AI technologies. AI's integration into products alters privacy risks, highlighting the need for ethical guidelines and interdisciplinary collaboration to navigate AI-generated media complexities [17]. The ethical discourse focuses on transparency, accountability, and safeguarding human rights amid evolving technologies.

2.2 Evolution and Advancements

The evolution of multimodal models and synthetic media is marked by significant advancements in capabilities and applications. Multimodal machine learning has progressed towards systems capable of understanding interactions across diverse data types, enabling sophisticated AI models to synthesize realistic synthetic media, including deepfakes. These developments facilitate the manipulation of various modalities to produce convincing yet deceptive content. Challenges include the absence of a unified model architecture, multimodal reasoning difficulties, and the need for extensive training data [22].

Deepfake technology has transitioned from handcrafted feature-based methods to advanced deep learning techniques, enhancing media realism and complexity. This evolution has increased fabricated content, raising concerns about misinformation [28]. Despite advancements in detection methods, the rapid pace of deepfake creation often surpasses effective countermeasure development [13]. Comprehensive detection frameworks are essential to address these challenges, focusing on deepfake diversity and origin tracing [29].

AI-synthesized human voices in audio media introduce impersonation and disinformation challenges. Developing robust detection methods for synthetic voices is crucial [30]. Evolving datasets with genuine and deepfake audio enhance detection capabilities by providing balanced training and evaluation representations [30]. However, existing benchmarks often focus on audio or video separately, limiting robust multimodal detection system development [31].

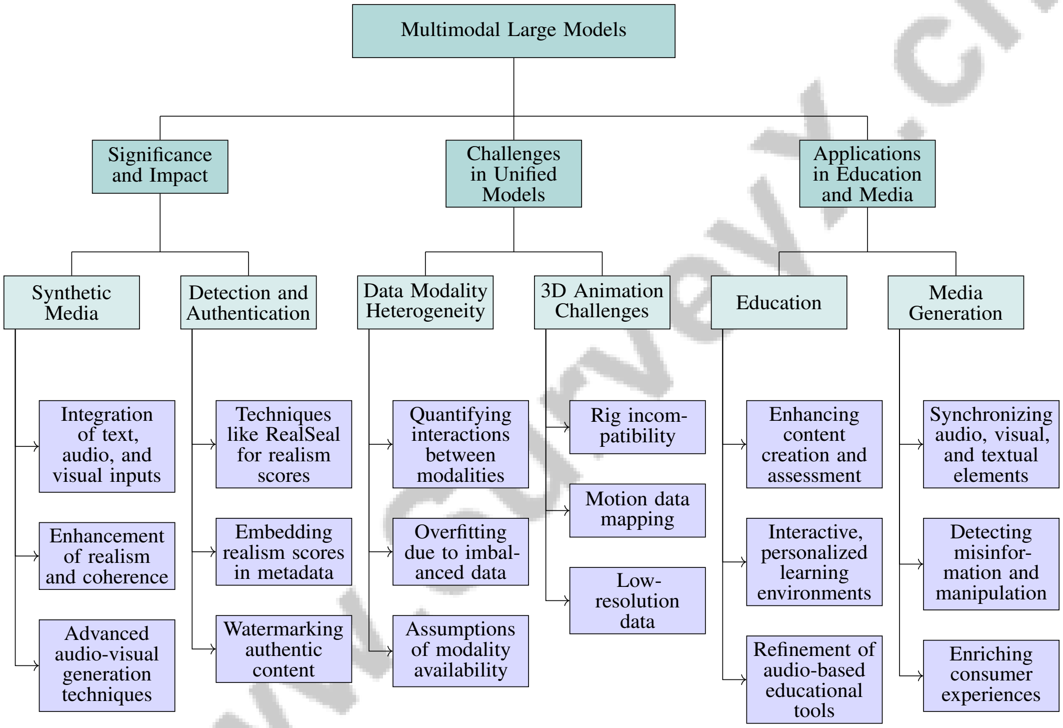
The field grapples with synthetic media technologies' rapid evolution, facing challenges like high-quality training datasets, AI hallucination risks, and AI model integration into workflows without overwhelming users [16]. The gap in audio deepfake detection research compared to image deepfakes highlights the need for advancements [32]. Despite these challenges, watermarking techniques for large language models enhance copyright protection and content traceability [33].

Advancements in multimodal models and synthetic media necessitate continuous research and innovation to harness their potential while mitigating risks. This includes addressing limitations in joint audio and video generation methods, which often fail to consider inherent modality correlations, leading to unrealistic outputs when generated independently [34]. The survey categorizes research into face swapping, reenactment, talking face generation, and facial attribute editing, along with forgery detection, highlighting deepfake technology's complexity [35]. Additionally, benchmarks by Papadopoulos et al. provide comparative studies of synthetic misinformers against real-world misinformation, addressing research gaps [9].

3 Multimodal Large Models

Multimodal large models (MLMs) have significantly influenced fields like synthetic media by proficiently processing and integrating diverse data modalities. Their role in both generating and combating misinformation is crucial, particularly in the realm of deepfake technology, which threatens societal trust and democratic values [36, 9, 37]. This examination highlights MLMs' foundational impact on enhancing media content's realism and coherence, driving transformative applications.

As illustrated in Figure 2, the hierarchical structure of Multimodal Large Models underscores their significance in synthetic media, elucidating the challenges faced in developing unified models and their applications in education and media generation. This figure categorizes the integration of diverse modalities, techniques for media authentication, and the transformative potential in creating dynamic content. Such a comprehensive overview reinforces the critical role of MLMs in advancing the field and addressing the complexities of modern media landscapes.



Multimodal Large Models
Significance and Impact
Challenges in Unified Models
Applications in Education and Media
Synthetic Media
Detection and Authentication
Data Modality Heterogeneity
3D Animation Challenges
Education
Media Generation
Integration of text, audio, and visual inputs
Enhancement of realism and coherence
Advanced audio-visual generation techniques
Techniques like RealSeal for realism scores
Embedding realism scores in metadata
Watermarking authentic content
Quantifying interactions between modalities
Overfitting due to imbalanced data
Assumptions of modality availability
Rig incompatibility
Motion data mapping
Low-resolution data
Enhancing content creation and assessment
Interactive, personalized learning environments
Refinement of audio-based educational tools
Synchronizing audio, visual, and textual elements
Detecting misinformation and manipulation
Enriching consumer experiences

Figure 2: This figure illustrates the hierarchical structure of Multimodal Large Models, highlighting their significance in synthetic media, challenges in unified models, and applications in education and media generation. It categorizes the integration of diverse modalities, techniques for media authentication, and the transformative potential in creating dynamic content.

3.1 Significance of Multimodal Large Models and Synthetic Media

MLMs advance synthetic media by integrating multiple modalities—text, audio, and visual inputs—essential for producing realistic and coherent media. This integration enriches multimedia storytelling and educational tools, creating engaging experiences [6]. In synthetic media, particularly deepfake videos, advanced audio-visual generation techniques enhance realism through synchronized elements [22]. Datasets like COSMOS, with 850 image-caption pairs from credible sources, are vital for evaluating and improving misinformation detection [9].

Beyond generation, MLMs are pivotal in detecting and authenticating synthetic media. Techniques like RealSeal use multisensory inputs and machine learning to generate realism scores, crucial for combating deepfakes. By embedding realism scores in metadata and watermarking authentic content, RealSeal revolutionizes trust verification in complex media landscapes [38, 3, 31]. This approach addresses challenges like misinformation and trust erosion in digital content.

3.2 Challenges in Unified Multimodal Models

Unified multimodal models face challenges due to data modality heterogeneity. Quantifying and learning interactions between text, audio, and visual inputs complicate creating cohesive models [39]. Neural networks often overfit certain modalities, especially when data is imbalanced, undermining realistic output production, such as synchronized lip motion with speech, essential for photorealistic virtual humans [40]. Traditional deepfake methods yield 3D rendering inconsistencies, resulting in blurred outputs due to complex 3D facial mapping [41].

Another obstacle is assuming all modalities are available during training and inference, which is rarely practical. This assumption can lead to overfitting, reducing robustness and generalization capabilities in scenarios with missing modalities [42]. Current multimodal deepfake detection methods requiring both audio and visual data are impractical in many applications [43]. Existing benchmarks focus on single-modality binary classification tasks, lacking the sophistication to analyze cross-modality manipulations, hindering comprehensive model development [28].

In 3D animation, challenges like rig incompatibility, motion data mapping, and low-resolution data complicate realistic multimodal model creation. Overcoming these hurdles is essential for advancing AI's capability in generating high-fidelity synthetic media [27].

Figure 3 illustrates the primary challenges faced by unified multimodal models, categorized into data modality heterogeneity, assumptions regarding modality availability, and specific challenges in 3D animation. Each category highlights key issues, such as interaction quantification, overfitting risks, and 3D rendering complexities, which hinder the development of cohesive multimodal models. Addressing these challenges is crucial for the progress of unified multimodal models, poised to revolutionize complex media content synthesis and analysis.

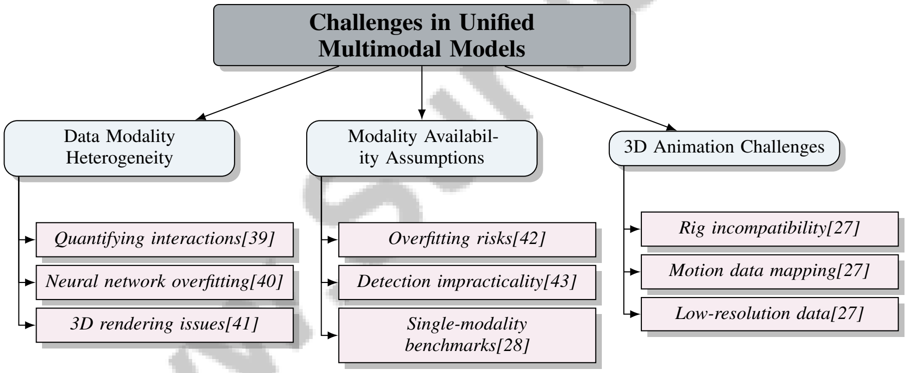

Challenges in Unified Multimodal Models
Data Modality Heterogeneity
Modality Availability Assumptions
3D Animation Challenges
Quantifying interactions[39]
Neural network overfitting[40]
3D rendering issues[41]
Overfitting risks[42]
Detection impracticality[43]
Single-modality benchmarks[28]
Rig incompatibility[27]
Motion data mapping[27]
Low-resolution data[27]

Figure 3: This figure illustrates the primary challenges faced by unified multimodal models, categorized into data modality heterogeneity, assumptions regarding modality availability, and specific challenges in 3D animation. Each category highlights key issues, such as interaction quantification, overfitting risks, and 3D rendering complexities, which hinder the development of cohesive multimodal models.

3.3 Applications in Education and Media Generation

MLMs hold potential in transforming education and media generation by integrating diverse data types. In education, they enhance content creation, support, assessment, and feedback, particularly in science education, fostering interactive, personalized learning environments [6]. In audio manipulation, benchmarks evaluating human perception of audio deepfakes refine educational tools reliant on audio, ensuring effectiveness and engagement [44].

In media, multimodal models enable rich experiences by synchronizing audio, visual, and textual elements, enhancing storytelling and multimedia production. This capability allows creators to detect misinformation and manipulation, ensuring authentic storytelling resonating with viewers [45, 28, 38, 46]. Generating realistic, coherent media content enriches consumer experiences and opens creative expression avenues in media industries.

Applications of multimodal models in education and media underscore their transformative potential in creating dynamic, personalized content. Leveraging models like MLLMs and synthetic media, educators and creators transcend traditional methodologies, fostering interactive environments enhancing personalization, accessibility, and engagement. These technologies facilitate innovative content creation and tailored support for diverse learning needs while raising ethical considerations necessitating careful navigation. They represent opportunities to enrich educational experiences across disciplines while fostering critical competencies in students [6, 21].

4 Deepfakes and Synthetic Media

4.1 Deepfake Technology and Challenges

Deepfake technology utilizes advanced generative algorithms, notably Generative Adversarial Networks (GANs) and diffusion models, to create highly realistic synthetic media. This sophistication makes distinguishing AI-generated content from genuine media increasingly difficult, as GANs often produce images that are virtually indistinguishable from real ones [47]. The ease of access to deepfake creation tools raises significant concerns about their potential to deceive audiences, posing societal and political risks [47].

A major challenge is the rapid advancement of generation techniques, which frequently outpace the development of effective detection methods. This gap highlights the urgent need for robust detection systems to counter the threats posed by misleading deepfakes [48]. Current detection efforts focus mainly on binary classifications of real versus AI-generated images, leaving a gap in distinguishing between different AI architectures and their unique characteristics.

The introduction of audio deepfakes adds complexity, necessitating sophisticated detection methods to address security and privacy threats. Integrating audio and visual modalities in detection frameworks is crucial for effectively identifying multimodal deepfakes and maintaining digital media integrity. Benchmarks that assess visual realism, such as those predicting Mean Opinion Scores (MOS), are vital for evaluating face-swapping models and enhancing detection capabilities [49].

The democratization of deepfake technology allows even inexperienced users to create realistic deepfakes swiftly, exacerbating the detection challenge and raising ethical concerns. Addressing these issues requires developing generalized detection methods resilient to adversarial attacks, particularly against videos degraded to simulate lower quality, diverging from training-time data augmentations [48].

In the context of the metaverse, deepfakes pose unique security risks, especially concerning impersonation and user safety. The rise of synthetic multimedia content underscores the need for comprehensive multimodal detection methods that analyze audio, video, text, and images to effectively combat misuse and misinformation. This necessity is driven by increasingly sophisticated misinformation tactics that exploit the nuances of each modality, requiring robust frameworks for accurate detection and clarification [38, 28, 31, 45, 46].

4.2 Impact of Deepfakes and Misinformation

Deepfakes are a powerful tool for spreading misinformation, capable of producing highly realistic synthetic media that can alter public perception and propagate false narratives. The sophistication of generative models, particularly those using diffusion techniques, complicates detection efforts, necessitating the development of robust and adaptable detectors to keep pace with these advancements. The continuous evolution of deepfake generation techniques often outstrips detection advancements, presenting ongoing challenges for effective detection systems [50].

The societal implications of deepfakes extend beyond misinformation, encompassing significant security and privacy concerns. Their ability to manipulate identities and create convincing forgeries poses substantial risks in various contexts, from political arenas to personal privacy breaches [49]. Recent benchmarks introducing multi-level hierarchical classification systems enhance the understanding of deepfake generation and recognition, improving detection and categorization capabilities [51].

Despite advancements in detection methods, challenges persist, particularly in generalizing across different generative models and addressing the limitations of existing multimodal detection approaches [52]. Traditional detection methods primarily focus on facial manipulations, underscoring the need

AI-generated, for reference only.

for comprehensive solutions that can identify videos with background manipulations or those entirely AI-generated [53].

Moreover, human detection abilities are significantly influenced by real-world conditions, such as distractions and video quality, which can lower detection rates [54]. This underscores the necessity for detection systems that function effectively in diverse and challenging environments. Performance degradation under certain conditions, such as high constant rate factors and datamoshing, further emphasizes the need for resilient detection methods [48].

4.3 Challenges in Deepfake Detection

Detecting deepfakes involves numerous challenges, primarily due to sophisticated generation techniques that often render them indistinguishable from authentic media. A significant obstacle is the context-specific and computationally intensive nature of current detection methods, which limits their effectiveness in real-world scenarios [55]. This limitation is exacerbated by rapid advancements in deepfake technology, which continually outpace detection system development, allowing attackers to exploit existing weaknesses [56].

A primary difficulty in deepfake detection is the poor generalization of current systems to unseen data and newer manipulation techniques, particularly those that do not focus on facial features. Existing detectors often struggle with non-face-centric manipulations, limiting their applicability in diverse real-world scenarios where various forms of synthetic media are prevalent [53]. This challenge is compounded by the reliance on pixel-level perturbations in current benchmarks, which overlook the potential for attribute-based attacks that can bypass detection algorithms without significantly altering pixel distributions [57].

Integrating audio-visual features in detection frameworks presents a core challenge due to distributional modality gaps that hinder the effective fusion of audio and visual data [47]. This gap complicates the development of robust multimodal detection systems capable of addressing the complexities of audiovisual deepfakes. Moreover, the significant increase in false alarms, particularly with pristine JPEG AI compressed images misclassified as deepfakes, underscores the need for improved detection methodologies [25].

Developing robust detection methods is further challenged by current systems' inadequacies in low-quality environments. Realistic evaluation frameworks that emphasize these limitations are crucial for advancing detection technologies [48]. Additionally, adaptive training strategies and data augmentations are essential for enhancing detection system performance across diverse datasets and manipulation techniques [58].

5 Generative Adversarial Networks (GANs)

5.1 Overview of GAN Architecture

Generative Adversarial Networks (GANs) represent a pivotal advancement in AI through their dual-network architecture consisting of a generator and a discriminator. This adversarial setup allows the generator to create data closely resembling real samples, while the discriminator differentiates between authentic and synthetic data [59]. The interaction of these networks facilitates highly realistic image generation, making GANs essential for applications like image synthesis and deepfake production [60]. Enhancements such as StyleGAN2 improve the generation process, producing high-resolution images by learning intricate patterns from datasets, crucial for applications requiring high fidelity [60, 61].

Despite their effectiveness, GANs pose challenges for deepfake detection, as detectors trained on one model often fail to generalize to others, revealing vulnerabilities in detection frameworks [62]. Solutions like GDA-Net have been developed to accurately attribute fake images to their originating GAN architectures, showing resilience to variations in model training and fine-tuning [63]. Additionally, integrating transformer-based frameworks, such as AVA, enhances GAN architectures by merging audio and visual features for more effective deception detection, particularly in scenarios with missing modalities [42].

5.2 Role in Synthetic Media Generation

GANs have revolutionized synthetic media generation by employing their adversarial structure to create realistic content across images, audio, and video. The generator-discriminator interaction enables media production often indistinguishable from authentic content, presenting both opportunities and challenges [4]. This is particularly evident in deepfake creation, where GANs manipulate facial features to generate synthetic videos, as demonstrated in various datasets [50]. Advanced techniques further enhance GANs' role; for instance, the M2M model, a conditional diffusion model, generates multiple interrelated images from a single input caption, illustrating complex image synthesis potential [23]. Frameworks like MIS-AVoiDD utilize modality-invariant and specific representations to improve audio-visual deepfake detection, showcasing GANs' versatility in multimodal contexts [47].

Beyond video and image synthesis, GANs are crucial in audio generation, contributing to the development of comprehensive multimodal datasets. This integration supports more realistic and contextually coherent synthetic media creation, facilitating sophisticated detection methods for multimodal deepfakes [50]. Innovative detection approaches, such as utilizing GAN-specific frequencies (GSF) in the DCT domain, further illustrate GANs' dual role in both generating and detecting synthetic media [55]. The advancement of GAN-based systems is bolstered by strategies that integrate supervised and reinforcement learning, optimizing data augmentations tailored to individual test samples, thus improving cross-dataset generalization and classification accuracy [64, 61, 65].

5.3 Strengths and Limitations of GANs

GANs are instrumental in synthetic media generation, offering strengths that enhance their utility in creating realistic content. A key advantage is their ability to generate lifelike images and videos through an adversarial framework, resulting in synthetic media often indistinguishable from authentic content [66]. Additionally, GANs can leave identifiable artificial fingerprints, aiding forensic analysis by identifying the specific GAN architecture used [59]. GANs' adaptability is another strength, particularly in capturing manipulation cues across modalities. Frameworks like MIS-AVoiDD have demonstrated superior performance by leveraging modality-invariant and specific representations, enhancing audio-visual deepfake detection [47]. Methods such as CTF-DCT are computationally efficient and provide explainable results, making them suitable for real-world applications requiring GAN-generated content detection [55].

However, GANs also exhibit notable limitations. Their vulnerability to adversarial attacks can significantly degrade deepfake detection models' performance. For instance, certain models, like SpecRNet, show substantial performance drops under adversarial conditions, while others, such as RawNet3, demonstrate greater robustness [67]. Moreover, GAN-based systems may overfit to specific faces or scenarios, limiting their generalizability and necessitating ongoing innovation to enhance robustness [68]. The integration of techniques like attention-diversity (AD) loss in UNITE represents a promising advancement in improving detection performance across varied manipulation contexts by encouraging models to distribute attention across different spatial regions in video frames [53]. Nevertheless, the need for high-quality datasets and the potential biases in AI outputs remain significant obstacles, as these factors can limit GANs' effectiveness and raise concerns regarding data privacy and security [16].

5.4 Advancements in GAN Technology

Recent advancements in GAN technology have significantly enhanced synthetic media generation capabilities and detection methods. Artificial fingerprinting techniques, which achieve perfect detection and attribution accuracy across various generative models, surpass existing methods and offer a sustainable solution for deepfake detection [69]. The integration of advanced machine learning frameworks, such as the Multimodal Graph Learning (MGL) framework, has improved the robustness and generalization capabilities of deepfake detection systems, significantly outperforming existing state-of-the-art methods [70]. Novel benchmarks focusing on synthetic videos generated by state-of-the-art video generators address previous datasets' limitations, providing a comprehensive platform for evaluating GAN-generated content [71].

In audio generation, the emergence of the Codecfake dataset marks a shift from traditional vocoder-generated benchmarks, focusing on audio from neural codecs, thus providing a more accurate representation of current audio synthesis technologies [72]. The integration of both audio and visual modalities in detection frameworks, exemplified by approaches like AVTENet, significantly enhances detection accuracy by addressing previous single-modality approaches' limitations [73]. Exploring adversarial training techniques, such as pixel-wise Gaussian blurring, presents a promising avenue for improving deepfake detection, proving more effective than traditional adversarial training methods [74]. Furthermore, hybrid models combining convolutional neural networks (CNNs) with other techniques have shown considerable effectiveness in detecting deepfakes, although further improvements are necessary for practical implementation [75].

These advancements underscore the dynamic nature of GAN technology and its implications for synthetic media creation and detection. As GANs evolve, developing innovative techniques and comprehensive benchmarks will be crucial in addressing the challenges posed by deepfakes, ensuring digital content integrity and mitigating risks to societal and electoral integrity [76].

6 AI Ethics

6.1 Ethical and Legal Considerations

The proliferation of AI-generated media, particularly deepfakes, raises complex ethical and legal challenges that necessitate a multifaceted approach. The advancement of deepfake technology increases risks related to misinformation and privacy violations, underscoring the need for effective detection methods and heightened public awareness to mitigate potential harms [77, 52]. The dual-use nature of deepfakes, serving both beneficial and malicious purposes, highlights the urgent requirement for robust detection strategies and comprehensive legal frameworks that address regulatory and privacy challenges, particularly concerning the CIA triad—confidentiality, integrity, and availability [78, 79].

In educational settings, the integration of Multimodal Large Language Models (MLLMs) presents significant ethical challenges, necessitating frameworks that ensure data protection and responsible use [6]. The advocacy for Green AI principles aligns with sustainability considerations by emphasizing eco-friendly and computationally efficient systems [80]. Resilient detection systems are crucial to counter vulnerabilities, as adversarial attacks pose serious security and privacy risks [57, 67]. The MFMC benchmark emphasizes privacy-preserving technologies in face recognition systems to protect individual privacy [24].

Interdisciplinary collaboration is vital in addressing deepfakes' ethical challenges, with contributions from linguists, psychologists, and technologists advancing detection methods and understanding synthetic media's behavioral impacts [49]. Ethical committee-approved experiments stress the importance of ethical considerations in research [56]. A comprehensive strategy emphasizing transparency, accountability, and privacy protections is essential to navigate AI ethics and legal frameworks, as AI-generated content introduces new privacy vulnerabilities and exacerbates existing risks [17, 12, 13]. Developing robust legal frameworks and ethical guidelines is crucial to prevent synthetic media exploitation and ensure responsible deployment of these technologies.

6.2 Societal Implications

The rise of synthetic media and AI technologies, including deepfakes, has profound societal implications across various sectors, from education to public discourse. Deepfakes' ability to manipulate content poses significant threats to information integrity, potentially influencing public opinion and disrupting democratic processes [81]. This necessitates robust detection methods and benchmarks to mitigate adverse societal effects [82]. In education, AI technologies offer opportunities and challenges, necessitating the addressing of bias, privacy, and educational equity to ensure beneficial impacts [6]. Educating students about deepfakes is crucial for fostering awareness and discernment in a digitally manipulated content landscape [83].

The music industry faces challenges from AI-generated media, where detection technologies are vital for protecting copyright and artistic integrity [30]. The development of eco-friendly AI approaches aligns with societal sustainability goals by reducing computational costs and carbon footprints [80]. Privacy technologies like D-CAPTCHA enhance security against deepfake threats [56], while privacy-enhancing deepfakes empower individuals to control their digital identities, contributing to privacy protection [24].

AI-generated, for reference only.

The societal implications are pronounced in regions with limited technological infrastructure, where detection technologies' vulnerabilities can lead to significant consequences [48]. Effective detection and moderation tools are essential to prevent synthetic media misuse and preserve information integrity. Addressing AI's societal implications requires interdisciplinary collaboration, robust detection methods, and ethical guidelines to navigate these transformative technologies' complexities. This involves developing frameworks for content verification and policies to combat AI misuse in misinformation and cyber threats [76, 17, 84, 85, 26].

6.3 Privacy and Data Protection

AI technologies, particularly in synthetic media and deepfakes generated by frameworks like GANs and Diffusion Models, present significant privacy and data protection challenges. These technologies enable the creation of realistic yet fabricated content, introducing privacy risks such as exposure from deepfake pornography and increased surveillance due to data collection. The rise of deepfakes necessitates urgent research into detection methods, highlighting ethical and security implications and prompting the development of a comprehensive taxonomy of AI privacy risks [17, 11]. Safeguarding personal information and ensuring data integrity are paramount as AI-generated content becomes more prevalent. The unauthorized use of personal data to create deepfakes can lead to identity theft and reputational damage, emphasizing the need for privacy-enhancing technologies [24].

AI integration into various sectors raises concerns about data collection and usage, necessitating stringent data protection measures, particularly in educational contexts [6]. Emphasizing privacy and data protection is crucial for maintaining trust and ensuring ethical AI use. Technological solutions like AI-driven image-based automated fact verification and advanced detection frameworks for deepfakes, alongside robust regulatory frameworks, are essential for addressing privacy and data protection challenges [36, 17, 85]. Comprehensive legal guidelines governing AI-generated content are vital for protecting individual rights and ensuring accountability. These frameworks should address ethical implications, promote transparency, and enforce strict data protection standards to safeguard personal information.

Addressing privacy and data protection challenges in AI and synthetic media requires a multifaceted approach combining technological innovation with robust regulatory measures. By emphasizing privacy and data protection, stakeholders can effectively address risks associated with AI-generated content, such as deepfake misuse and heightened surveillance concerns, fostering responsible and ethical technology use. Implementing privacy-preserving measures like federated learning and differential privacy is essential to mitigate these risks and ensure information integrity in the digital realm [17, 12, 85, 13, 26].

6.4 Interdisciplinary Collaboration and Ethical Guidelines

The complexity and global reach of deepfake technologies necessitate interdisciplinary collaboration and ethical guidelines to address the challenges they present effectively. The international nature of deepfakes underscores the need for collaborative efforts to develop regulatory measures that adapt to rapid technological advancements [76]. Such collaboration is essential for creating comprehensive strategies encompassing legal, technological, and societal dimensions, ensuring a coordinated response to synthetic media threats [86].

Interdisciplinary collaboration integrates insights from diverse fields, including computer science, law, ethics, and social sciences, to develop holistic solutions for combating deepfake threats. This approach enhances detection systems' robustness and facilitates adaptive frameworks that generalize across different deepfakes and datasets, addressing current limitations in generalization and rapid evolution of generative techniques. Incorporating multi-modal data integration and exploring larger model architectures are promising directions for improving detection systems' performance and adaptability [87].

Open-source availability of datasets and models plays a pivotal role in fostering further research and development in deepfake detection and prevention. Public access to these resources enables researchers to collaboratively refine detection methodologies, contributing to a more resilient understanding of synthetic media. Developing standardized evaluation metrics is crucial for assessing detection techniques' effectiveness, ensuring consistency and comparability across studies [88].

Ethical guidelines are crucial in navigating deepfakes' challenges, addressing synthetic media's moral implications, and emphasizing transparency, accountability, and individual rights protection. The relationship between traditional digital watermarking and LLM watermarking illustrates evolving ethical considerations in AI-generated content, highlighting the need for innovative approaches to safeguard digital integrity [33].

Effective strategies against synthetic media require a nuanced understanding of detection technologies and stakeholder collaboration. By fostering interdisciplinary partnerships and adhering to ethical principles, the global community can better navigate deepfakes' complexities, ensuring responsible AI deployment and societal values protection [89].

7 Challenges and Future Directions

Addressing the challenges posed by deepfake technologies requires innovative strategies aimed at enhancing detection robustness and generalization. The ever-evolving nature of deepfake generation necessitates a proactive approach to detection methodologies, ensuring systems remain effective against advancing adversarial tactics. This section explores critical advancements needed in detection systems, emphasizing the integration of diverse modalities and the refinement of existing models to improve performance across various contexts. By prioritizing these enhancements, the research community can better tackle the complexities introduced by synthetic media and develop more resilient detection frameworks.

7.1 Enhancing Detection Robustness and Generalization

The ongoing evolution of deepfake technologies necessitates the development of detection systems that are robust and capable of generalizing across diverse datasets and generative models. Future research should focus on bolstering the robustness of detection models against emerging deepfake generation techniques and exploring additional modalities for detection [47]. This can be achieved through the integration of advanced AI techniques, such as ensemble and fusion methods, which utilize multiple data sources to enhance model robustness, particularly in audio deepfake detection [56].

Expanding datasets to include a wider variety of video types and refining models for improved prediction accuracy are essential steps in enhancing detection systems [49]. Additionally, investigating cross-modality and within-modality regularization techniques can significantly boost detection accuracy and robustness against independent modality manipulations, addressing current challenges in AI-generated media [57].

Developing robust deepfake detection algorithms that perform reliably across a broader range of video qualities is crucial. Future research should also explore various corruption methods to ensure detection systems remain effective in real-time scenarios [48]. Furthermore, establishing ethical guidelines and investigating emerging technologies such as large language models in conjunction with deepfakes are vital areas for future exploration [9].

Metrics that assess the effectiveness of generated deepfakes in misleading face recognition systems and their dissimilarity from original faces underscore the necessity for improved detection robustness [24]. By prioritizing these areas, the research community can develop innovative solutions to combat the challenges posed by synthetic media, ultimately enhancing detection model performance in diverse contexts and improving dataset quality.

7.2 Future Directions in Deepfake Research

The trajectory of deepfake research is set for significant advancements, focusing on enhancing detection capabilities and addressing the multifaceted challenges these technologies present. A promising direction involves integrating multimodal data to improve detection accuracy and provide a comprehensive understanding of deepfake content [35]. This includes developing two-stream networks that combine handcrafted and deep features, incorporating temporal data to enhance detection processes [90].

Future research should prioritize expanding datasets to encompass a broader range of forgery techniques and audio sources, thereby enhancing the robustness and applicability of detection methods across diverse scenarios. Exploring cross-modality and within-modality regularization techniques will further augment detection capabilities, addressing the complexities of multimodal deepfakes [91].

Integrating advanced deep learning techniques with large language models (LLMs) presents another promising research avenue, potentially enhancing detection performance and interpretability [79]. Additionally, developing adaptive defenses that can respond to attribute-level attacks will be crucial for maintaining the integrity of detection systems against evolving adversarial strategies [57].

Innovative frameworks like D-CAPTCHA could be expanded to address video-based deepfakes, enhancing their utility across various communication contexts [56]. Furthermore, applying emerging loss functions and improving training datasets, as demonstrated in models like UNITE, will be essential for enhancing real-time detection capabilities [53].

Lastly, exploring new adversarial attack strategies and developing provably robust deepfake detectors will be critical for continuously challenging and refining existing defenses, ensuring detection systems remain resilient against the rapid evolution of deepfake technologies [92].

By pursuing these research avenues, the academic and technological communities can create advanced and adaptable detection methods that improve the identification of AI-generated media across various formats, including audio, images, and text. This effort is vital for preserving media integrity and fostering public trust in digital content, particularly in light of findings indicating the increasing sophistication of forgeries, which are often indistinguishable from authentic media. Additionally, understanding human detection capabilities, influenced by generalized trust and familiarity with deepfakes, will inform the development of more effective strategies to combat misinformation and media manipulation [28, 13].

7.3 Regulatory and Legal Frameworks

The proliferation of deepfake technologies necessitates comprehensive regulatory and legal frameworks to govern AI-generated media, addressing the multifaceted challenges they pose to privacy, security, and societal integrity. Significant gaps exist in understanding the implications of deepfakes within legal contexts, underscoring the urgent need for standardized detection protocols [93]. The complexity of these technologies requires a structured approach, as illustrated by research architectures that categorize challenges into layers such as DeepFake Architectural Robustness, DeepFake Detection, and DeepFake Forensics, each focusing on specific aspects of detection and analysis [94].

Robust regulatory frameworks should hold all parties in the deepfake supply chain accountable, from creators to distributors, to ensure responsible use and mitigate potential harms [95]. This includes establishing legal standards governing the creation, distribution, and use of synthetic media, ensuring that these technologies do not infringe on individual rights or societal norms.

Open-source benchmarks, such as those provided by the Deepfake Detection Challenge Dataset, are critical for facilitating contributions from the research community and enabling the continuous refinement of detection methods. These benchmarks are essential for evaluating the effectiveness of various detection strategies, highlighting the importance of considering both semantic and perceptual features in the detection process [96].

The necessity for regulatory measures is further emphasized by benchmarks designed to evaluate the effectiveness of models in generating privacy-enhancing deepfakes. These benchmarks allow users to control their digital identities, indicating a critical requirement for regulatory frameworks that protect individual privacy while fostering technological innovation [24]. However, the need for substantial training data and reference videos in some detection methods may limit their applicability, highlighting the importance of scalable solutions that can be implemented across diverse scenarios [97].

7.4 Dataset Diversity and Benchmarking

The advancement of AI-generated media research, particularly in deepfake detection, critically depends on the availability and diversity of datasets. Diverse datasets form the foundation for training robust detection models and facilitate comprehensive evaluation across varying scenarios and modalities [101]. The Celeb-DF dataset, for example, serves as a large-scale collection of high-

Benchmark	Size	Domain	Task Format	Metric
DDDB[29]	62,154	Art Detection	Image Classification	AA, mAP
DREAM[51]	62,500	Multimedia Forensics	Image Classification	Accuracy
HADDI[98]	100	Image Deepfake Detection	Binary Classification	Accuracy, Confidence
ADDB[44]	12,274	Audio Processing	Audio Classification	Accuracy
Codecfake[72]	1,058,216	Audio Deepfake Detection	Detection	EER
LookupForensics[85]	40,000	Image Forgery Detection	Fact Verification	µAP
SONAR[99]	2,274	Audio Deepfake Detection	Audio Classification	EER, Accuracy
SID-Set[100]	300,000	Image Manipulation	Deepfake Detection	Accuracy, F1-score

Table 1: Table summarizing key benchmarks for AI-generated media detection, highlighting dataset size, domain focus, task format, and evaluation metrics. These benchmarks represent diverse modalities and tasks, providing a comprehensive overview of current datasets used in deepfake and multimedia forensics research.

quality deepfake videos, playing a pivotal role in assessing the effectiveness of detection methods [102]. Similarly, a dataset comprising 40,000 images generated by each model, divided into training, validation, and testing sets, is specifically designed to evaluate detector performance on synthetic human faces [103]. Table 1 presents a detailed comparison of various benchmarks employed in the study of AI-generated media detection, illustrating the diversity and scope of datasets available for advancing deepfake detection research.

However, notable gaps persist in the availability of large-scale datasets for audio and video detection, alongside a lack of standardized evaluation benchmarks, complicating reliable comparisons of detection methods [12]. This underscores the necessity for future research to focus on collecting more diverse datasets and exploring additional methods for detecting AI-generated art and media [29]. Creating datasets using a combination of real images from established datasets like CelebA, FFHQ, and ImageNet, alongside synthetic images generated by various GAN architectures and diffusion models, exemplifies the approach needed to enhance dataset diversity and benchmarking efforts [51].

Furthermore, comprehensive datasets such as FakeMusicCaps and SONICS are instrumental in advancing research on Audio-Generated Media (AIGM) detection, providing rich resources for evaluating detection model performance in this domain [30]. Emphasizing cross-dataset performance evaluation is crucial, ensuring that detection methods are not only effective within a single dataset's context but also generalize well across different datasets, thereby improving their robustness and applicability in real-world scenarios [58].

8 Conclusion

This survey explores the complex landscape of multimodal large models, deepfakes, generative adversarial networks (GANs), synthetic media, and AI ethics, highlighting their significant implications for media integrity and societal trust. The findings indicate that deepfakes pose substantial threats, necessitating the development of robust detection methods that can keep pace with the rapid evolution of deepfake creation [104]. Current detection techniques must undergo continual refinement to effectively address the challenges presented by real-world media [105].

A notable gap in the literature is the lack of diverse datasets and the requirement for detection techniques that generalize across various large artificial intelligence models (LAIMs) [12]. This survey underscores the critical importance of robust datasets in training effective detection models, drawing comparisons between tampering and deepfake detection [19]. Furthermore, advancing our understanding of modality interactions and developing models capable of generalizing across multiple modalities and tasks is essential [39].

Current research is limited by its failure to account for contextual factors that affect detection capabilities in real-world scenarios, emphasizing the need for context-aware detection frameworks [13]. Promising advancements in audio deepfake detection have emerged, with codec-trained countermeasures achieving near-zero error rates, indicating a fruitful direction for future research [106]. The proposed datasets and benchmarks provide a valuable foundation for advancing audio deepfake detection, highlighting the necessity for effective evaluation methods [32].

Moreover, audio features significantly improve the accuracy of misinformation detection in short videos, reinforcing the importance of integrating audio analysis into detection frameworks [38]. The increasing demand for enhanced detection methods for audio deepfakes is clear, warranting further investigation in this domain [107].

AI-generated, for reference only.

References

[1] Shahzeb Naeem, Ramzi Al-Sharawi, Muhammad Riyyan Khan, Usman Tariq, Abhinav Dhall, and Hasan Al-Nashash. Real, fake and synthetic faces – does the coin have three sides?, 2024.

[2] Sm Zobaed, Md Fazle Rabby, Md Istiaq Hossain, Ekram Hossain, Sazib Hasan, Asif Karim, and Khan Md. Hasib. Deepfakes: Detecting forged and synthetic media content using machine learning, 2021.

[3] Bhaktipriya Radharapu and Harish Krishna. Realseal: Revolutionizing media authentication with real-time realism scoring, 2024.

[4] Baiwu Zhang, Jin Peng Zhou, Ilia Shumailov, and Nicolas Papernot. On attribution of deepfakes, 2021.

[5] Haojie Wu, Pan Hui, and Pengyuan Zhou. Deepfake in the metaverse: An outlook survey, 2023.

[6] Arne Bewersdorff, Christian Hartmann, Marie Hornberger, Kathrin Seßler, Maria Bannert, Enkelejda Kasneci, Gjergji Kasneci, Xiaoming Zhai, and Claudia Nerdel. Taking the next step with generative artificial intelligence: The transformative role of multimodal large language models in science education, 2024.

[7] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection, 2020.

[8] Trung-Nghia Le, Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Robust deepfake on unrestricted media: Generation and detection, 2022.

[9] Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C. Petrantonakis. Synthetic misinformers: Generating and combating multimodal misinformation, 2023.

[10] Mirko Casu, Luca Guarnera, Pasquale Caponnetto, and Sebastiano Battiato. Genai mirage: The impostor bias and the deepfake detection challenge in the era of artificial illusions, 2024.

[11] Irene Amerini, Mauro Barni, Sebastiano Battiato, Paolo Bestagini, Giulia Boato, Tania Sari Bonaventura, Vittoria Bruni, Roberto Caldelli, Francesco De Natale, Rocco De Nicola, Luca Guarnera, Sara Mandelli, Gian Luca Marcialis, Marco Micheletto, Andrea Montibeller, Giulia Orru', Alessandro Ortis, Pericle Perazzo, Giovanni Puglisi, Davide Salvi, Stefano Tubaro, Claudia Melis Tonti, Massimo Villari, and Domenico Vitulano. Deepfake media forensics: State of the art and challenges ahead, 2024.

[12] Li Lin, Neeraj Gupta, Yue Zhang, Hainan Ren, Chun-Hao Liu, Feng Ding, Xin Wang, Xin Li, Luisa Verdoliva, and Shu Hu. Detecting multimedia generated by large ai models: A survey. *arXiv preprint arXiv:2402.00045*, 2024.

[13] Joel Frank, Franziska Herbert, Jonas Ricker, Lea Schönherr, Thorsten Eisenhofer, Asja Fischer, Markus Dürmuth, and Thorsten Holz. A representative study on human detection of artificially generated media across countries, 2023.

[14] Enes Altuncu, Virginia N. L. Franqueira, and Shujun Li. Deepfake: Definitions, performance metrics and standards, datasets and benchmarks, and a meta-review, 2022.

[15] Wenwu Zhu, Xin Wang, and Wen Gao. Multimedia intelligence: When multimedia meets artificial intelligence. *IEEE Transactions on Multimedia*, 22(7):1823–1835, 2020.

[16] Liangrui Pan, Zhenyu Zhao, Ying Lu, Kewei Tang, Liyong Fu, Qingchun Liang, and Shaoliang Peng. Opportunities and challenges in the application of large artificial intelligence models in radiology, 2024.

[17] Hao-Ping Lee, Yu-Ju Yang, Thomas Serban von Davier, Jodi Forlizzi, and Sauvik Das. Deepfakes, phrenology, surveillance, and more! a taxonomy of ai privacy risks, 2024.

[18] Deep insights of deepfake techno.

AI-generated, for reference only.

[19] Junke Wang, Zhenxin Li, Chao Zhang, Jingjing Chen, Zuxuan Wu, Larry S. Davis, and Yu-Gang Jiang. Fighting malicious media data: A survey on tampering detection and deepfake detection, 2022.

[20] Zhikan Wang, Zhongyao Cheng, Jiajie Xiong, Xun Xu, Tianrui Li, Bharadwaj Veeravalli, and Xulei Yang. A timely survey on vision transformer for deepfake detection, 2024.

[21] Jasper Roe and Mike Perkins. Deepfakes and higher education: A research agenda and scoping review of synthetic media, 2024.

[22] Xinji Mai, Zeng Tao, Junxiong Lin, Haoran Wang, Yang Chang, Yanlan Kang, Yan Wang, and Wenqiang Zhang. From efficient multimodal models to world models: A survey, 2024.

[23] Ying Shen, Yizhe Zhang, Shuangfei Zhai, Lifu Huang, Joshua M. Susskind, and Jiatao Gu. Many-to-many image generation with auto-regressive diffusion models, 2024.

[24] Umur A. Ciftci, Gokturk Yuksek, and Ilke Demir. My face my choice: Privacy enhancing deepfakes for social media anonymization, 2022.

[25] Edoardo Daniele Cannas, Sara Mandelli, Natasa Popovic, Ayman Alkhateeb, Alessandro Gnutti, Paolo Bestagini, and Stefano Tubaro. Is jpeg ai going to change image forensics?, 2024.

[26] Subash Neupane, Ivan A. Fernandez, Sudip Mittal, and Shahram Rahimi. Impacts and risk of generative ai technology on cyber defense, 2023.

[27] Julius Girbig, Changkun Ou, and Sylvia Rothe. Generative 3d animation pipelines: Automating facial retargeting workflows, 2022.

[28] Rui Shao, Tianxing Wu, and Ziwei Liu. Detecting and grounding multi-modal media manipulation, 2023.

[29] Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. Benchmarking deepart detection, 2023.

[30] Yupei Li, Manuel Milling, Lucia Specia, and Björn W. Schuller. From audio deepfake detection to ai-generated music detection – a pathway and overview, 2024.

[31] Davide Salvi, Brian Hosler, Paolo Bestagini, Matthew C. Stamm, and Stefano Tubaro. Timit-tts: a text-to-speech dataset for multimodal synthetic media detection, 2022.

[32] Joel Frank and Lea Schönherr. Wavefake: A data set to facilitate audio deepfake detection, 2021.

[33] Yuqing Liang, Jiancheng Xiao, Wensheng Gan, and Philip S. Yu. Watermarking techniques for large language models: A survey, 2024.

[34] Vinod K Kurmi, Vipul Bajaj, Badri N Patro, K S Venkatesh, Vinay P Namboodiri, and Preethi Jyothi. Collaborative learning to generate audio-video jointly, 2021.

[35] Gan Pei, Jiangning Zhang, Menghan Hu, Zhenyu Zhang, Chengjie Wang, Yunsheng Wu, Guangtao Zhai, Jian Yang, Chunhua Shen, and Dacheng Tao. Deepfake generation and detection: A benchmark and survey, 2024.

[36] Claudio S. Pinhanez, German H. Flores, Marisa A. Vasconcelos, Mu Qiao, Nick Linck, Rogério de Paula, and Yuya J. Ong. Towards a new science of disinformation, 2022.

[37] Nikolaos Misirlis and Harris Bin Munawar. From deepfake to deep useful: risks and opportunities through a systematic literature review, 2023.

[38] Moyang Liu, Yukun Liu, Ruibo Fu, Zhengqi Wen, Jianhua Tao, Xuefei Liu, and Guanjun Li. Exploring the role of audio in multimodal misinformation detection, 2024.

[39] Paul Pu Liang. Foundations of multisensory artificial intelligence, 2024.

[40] Siddarth Ravichandran, Ondřej Texler, Dimitar Dinev, and Hyun Jae Kang. Synthesizing photorealistic virtual humans through cross-modal disentanglement, 2023.

[41] Georgii Stanishevskii, Jakub Steczkiewicz, Tomasz Szczepanik, Sławomir Tadeja, Jacek Tabor, and Przemysław Spurek. Deepfake for the good: Generating avatars through face-swapping with implicit deepfake generation, 2024.

[42] Zhaoxu Li, Zitong Yu, Nithish Muthuchamy Selvaraj, Xiaobao Guo, Bingquan Shen, Adams Wai-Kin Kong, and Alex Kot. Flexible-modal deception detection with audio-visual adapter, 2023.

[43] Cai Yu, Peng Chen, Jiahe Tian, Jin Liu, Jiao Dai, Xi Wang, Yesheng Chai, Shan Jia, Siwei Lyu, and Jizhong Han. A unified framework for modality-agnostic deepfakes detection, 2023.

[44] Nicolas M. Müller, Karla Pizzi, and Jennifer Williams. Human perception of audio deepfakes, 2024.

[45] Ekraam Sabir, Ayush Jaiswal, Wael AbdAlmageed, and Prem Natarajan. Meg: Multi-evidence gnn for multimodal semantic forensics, 2020.

[46] Yizhou Zhang, Loc Trinh, Defu Cao, Zijun Cui, and Yan Liu. Interpretable detection of out-of-context misinformation with neural-symbolic-enhanced large multimodal model, 2024.

[47] Vinaya Sree Katamneni and Ajita Rattani. Mis-avoidd: Modality invariant and specific representation for audio-visual deepfake detection, 2023.

[48] Yang A. Chuming, Daniel J. Wu, and Ken Hong. Practical deepfake detection: Vulnerabilities in global contexts, 2022.

[49] Luka Dragar, Peter Peer, Vitomir Štruc, and Borut Batagelj. Beyond detection: Visual realism assessment of deepfakes, 2023.

[50] Junhao Xu, Jingjing Chen, Xue Song, Feng Han, Haijun Shan, and Yugang Jiang. Identity-driven multimedia forgery detection via reference assistance, 2024.

[51] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. Level up the deepfake detection: a method to effectively discriminate images generated by gan architectures and diffusion models, 2023.

[52] Hong-Hanh Nguyen-Le, Van-Tuan Tran, Dinh-Thuc Nguyen, and Nhien-An Le-Khac. Passive deepfake detection across multi-modalities: A comprehensive survey, 2024.

[53] Rohit Kundu, Hao Xiong, Vishal Mohanty, Athula Balachandran, and Amit K. Roy-Chowdhury. Towards a universal synthetic video detector: From face or background manipulations to fully ai-generated content, 2024.

[54] Emilie Josephs, Camilo Fosco, and Aude Oliva. Artifact magnification on deepfake videos increases human detection and subjective confidence, 2023.

[55] Oliver Giudice, Luca Guarnera, and Sebastiano Battiato. Fighting deepfakes by detecting gan dct anomalies, 2021.

[56] Lior Yasur, Guy Frankovits, Fred M. Grabovski, and Yisroel Mirsky. Deepfake captcha: A method for preventing fake calls, 2023.

[57] Xiangtao Meng, Li Wang, Shanqing Guo, Lei Ju, and Qingchuan Zhao. Ava: Inconspicuous attribute variation-based adversarial attack bypassing deepfake detection, 2023.

[58] Luca Bondi, Edoardo Daniele Cannas, Paolo Bestagini, and Stefano Tubaro. Training strategies and data augmentations in cnn-based deepfake video detection, 2020.

[59] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do gans leave artificial fingerprints?, 2018.

[60] Simranjeet Singh, Rajneesh Sharma, and Alan F. Smeaton. Using gans to synthesise minimum training data for deepfake generation, 2020.

AI-generated, for reference only.

[61] Mohammad Lataifeh, Xavier Carrasco, Ashraf Elnagar, and Naveed Ahmed. Augmenting character designers creativity using generative adversarial networks, 2023.

[62] Florinel-Alin Croitoru, Andrei-Iulian Hiji, Vlad Hondru, Nicolae Catalin Ristea, Paul Irofti, Marius Popescu, Cristian Rusu, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. Deepfake media generation and detection in the generative ai era: A survey and outlook, 2024.

[63] Sowdagar Mahammad Shahid, Sudev Kumar Padhi, Umesh Kashyap, and Sk. Subidh Ali. Generalized deepfake attribution, 2024.

[64] Sohail Ahmed Khan and Duc-Tien Dang-Nguyen. Clipping the deception: Adapting vision-language models for universal deepfake detection, 2024.

[65] Aakash Varma Nadimpalli and Ajita Rattani. On improving cross-dataset generalization of deepfake detectors, 2022.

[66] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. Deepfake detection by analyzing convolutional traces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 666–667, 2020.

[67] Piotr Kawa, Marcin Plata, and Piotr Syga. Defense against adversarial attacks on audio deepfake detection, 2023.

[68] Sowmen Das, Selim Seferbekov, Arup Datta, Md Saiful Islam, and Md Ruhul Amin. 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*, pages 3776–3785, 2021.

[69] Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data, 2022.

[70] Zhiyuan Yan, Peng Sun, Yubo Lang, Shuo Du, Shanzhuo Zhang, Wei Wang, and Lei Liu. Multimodal graph learning for deepfake detection, 2023.

[71] Danial Samadi Vahdati, Tai D. Nguyen, Aref Azizpour, and Matthew C. Stamm. Beyond deepfake images: Detecting ai-generated videos, 2024.

[72] Yi Lu, Yuankun Xie, Ruibo Fu, Zhengqi Wen, Jianhua Tao, Zhiyong Wang, Xin Qi, Xuefei Liu, Yongwei Li, Yukun Liu, Xiaopeng Wang, and Shuchen Shi. Codecfake: An initial dataset for detecting llm-based deepfake audio, 2024.

[73] Ammarah Hashmi, Sahibzada Adil Shahzad, Chia-Wen Lin, Yu Tsao, and Hsin-Min Wang. Avtenet: Audio-visual transformer-based ensemble network exploiting multiple experts for video deepfake detection, 2023.

[74] Zhi Wang, Yiwen Guo, and Wangmeng Zuo. Deepfake forensics via an adversarial game, 2022.

[75] Jacob mallet, Laura Pryor, Rushit Dave, and Mounika Vanamala. Deepfake detection analyzing hybrid dataset utilizing cnn and svm, 2023.

[76] Hriday Ranka, Mokshit Surana, Neel Kothari, Veer Pariawala, Pratyay Banerjee, Aditya Surve, Sainath Reddy Sankepally, Raghav Jain, Jhagrut Lalwani, and Swapneel Mehta. Examining the implications of deepfakes for election integrity, 2024.

[77] Aniruddha Tiwari, Rushit Dave, and Mounika Vanamala. Leveraging deep learning approaches for deepfake detection: A review, 2023.

[78] Shahroz Tariq, Alsharif Abuadbba, and Kristen Moore. Deepfake in the metaverse: Security implications for virtual gaming, meetings, and offices, 2023.

[79] Sahibzada Adil Shahzad, Ammarah Hashmi, Yan-Tsung Peng, Yu Tsao, and Hsin-Min Wang. How good is chatgpt at audiovisual deepfake detection: A comparative study of chatgpt, ai models and human perception, 2024.

[80] Subhajit Saha, Md Sahidullah, and Swagatam Das. Exploring green ai for audio deepfake detection, 2024.

[81] Kundan Patil, Shrushti Kale, Jaivanti Dhokey, and Abhishek Gulhane. Deepfake detection using biological features: A survey, 2023.

[82] Sanjay Saha, Rashindrie Perera, Sachith Seneviratne, Tamasha Malepathirana, Sanka Rasnayaka, Deshani Geethika, Terence Sim, and Saman Halgamuge. Undercover deepfakes: Detecting fake segments in videos, 2023.

[83] Gabrielle Watson, Zahra Khanjani, and Vandana P. Janeja. Audio deepfake perceptions in college going populations, 2021.

[84] Christina P. Walker, Daniel S. Schiff, and Kaylyn Jackson Schiff. Merging ai incidents research with political misinformation research: Introducing the political deepfakes incidents database, 2024.

[85] Shuhan Cui, Huy H. Nguyen, Trung-Nghia Le, Chun-Shien Lu, and Isao Echizen. Lookupforensics: A large-scale multi-task dataset for multi-phase image-based fact verification, 2024.

[86] Achhardeep Kaur, Azadeh Noori Hoshyar, Vidya Saikrishna, Selena Firmin, and Feng Xia. Deepfake video detection: challenges and opportunities. *Artificial Intelligence Review*, 57(6):159, 2024.

[87] Dinesh Srivasthav P and Badri Narayan Subudhi. Adaptive meta-learning for robust deepfake detection: A multi-agent framework to data drift and model generalization, 2024.

[88] Momina Masood, Marriam Nawaz, Khalid Mahmood Malik, Ali Javed, and Aun Irtaza. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward, 2021.

[89] Claire Leibowicz, Sean McGregor, and Aviv Ovadya. The deepfake detection dilemma: A multistakeholder exploration of adversarial dynamics in synthetic media, 2021.

[90] Ying Xu and Sule Yildirim Yayilgan. When handcrafted features and deep features meet mismatched training and test sets for deepfake detection, 2022.

[91] Heqing Zou, Meng Shen, Yuchen Hu, Chen Chen, Eng Siong Chng, and Deepu Rajan. Cross-modality and within-modality regularization for audio-visual deepfake detection, 2024.

[92] Shehzeen Hussain, Paarth Neekhara, Malhar Jere, Farinaz Koushanfar, and Julian McAuley. Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples, 2020.

[93] Naciye Celebi, Qingzhong Liu, and Muhammed Karatoprak. A survey of deep fake detection for trial courts, 2022.

[94] Amin Azmoodeh and Ali Dehghantanha. Deep fake detection, deterrence and response: Challenges and opportunities, 2022.

[95] Andrea Miotti and Akash Wasil. Combatting deepfakes: Policies to address national security threats and rights violations, 2024.

[96] Roberto Amoroso, Davide Morelli, Marcella Cornia, Lorenzo Baraldi, Alberto Del Bimbo, and Rita Cucchiara. Parents and children: Distinguishing multimodal deepfakes from natural images, 2024.

[97] Davide Cozzolino, Alessandro Pianese, Matthias Nießner, and Luisa Verdoliva. Audio-visual person-of-interest deepfake detection, 2023.

[98] Sergi D. Bray, Shane D. Johnson, and Bennett Kleinberg. Testing human ability to detect deepfake images of human faces, 2023.

[99] Xiang Li, Pin-Yu Chen, and Wenqi Wei. Sonar: A synthetic ai-audio detection framework and benchmark, 2024.

[100] Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. Sida: Social media image deepfake detection, localization and explanation with large multimodal model, 2024.

[101] Luisa Verdoliva. Media forensics and deepfakes: an overview, 2020.

[102] Siwei Lyu. Deepfake detection: Current challenges and next steps. In *2020 IEEE international conference on multimedia & expo workshops (ICMEW)*, pages 1–6. IEEE, 2020.

[103] Yuhang Lu and Touradj Ebrahimi. Towards the detection of ai-synthesized human face images, 2024.

[104] Thanh Thi Nguyen, Quoc Viet Hung Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Thanh Tam Nguyen, Quoc-Viet Pham, and Cuong M. Nguyen. Deep learning for deepfakes creation and detection: A survey, 2022.

[105] Siwei Lyu. Deepfake detection: Current challenges and next steps, 2020.

[106] Yuankun Xie, Chenxu Xiong, Xiaopeng Wang, Zhiyong Wang, Yi Lu, Xin Qi, Ruibo Fu, Yukun Liu, Zhengqi Wen, Jianhua Tao, Guanjun Li, and Long Ye. Does current deepfake audio detection model effectively detect alm-based deepfake audio?, 2024.

[107] Zahra Khanjani, Gabrielle Watson, and Vandana P. Janeja. How deep are the fakes? focusing on audio deepfake: A survey, 2021.

AI-generated, for reference only.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.