International Journal of Applied Mathematics

Volume 38 No. 1s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

DEFENDING THE METAVERSE: A SURVEY ON DEEPFAKE DETECTION AND AVATAR-BASED THREAT MITIGATION

[1]Muppidi Rajkumar, [2]Dr. K. Padmaja

[1]Research Scholar, [2]Research Supervisor and Asst. Professor
Department of Computer Science and Engineering
Kakatiaya University Warangal, Telangana, India, 506009
Email: muppidi.rajkumar@gmail.com, ajit2107@yahoo.com

Abstract

The proliferation of deepfake technologies, powered by generative adversarial networks (GANs), diffusion models, and transformer-based architectures, has led to a significant escalation in identity manipulation and misinformation. In virtual environments like the Metaverse—where interaction is synchronous, embodied, and immersive—the threat posed by deepfakes expands beyond traditional media falsification. This survey explores the evolution of deepfake techniques and their convergence with avatar-based deception, voice cloning, and synthetic behavioral modeling in virtual reality (VR) and extended reality (XR) platforms. It reviews state-of-the-art detection methods, including spatial-temporal models and multimodal attention-based architectures, and evaluates their applicability to the Metaverse's unique challenges. The paper also analyzes the 2D-FACT model as a foundational architecture, examines new threat vectors such as real-time avatar impersonation and AI-driven NPC manipulation, and identifies critical research gaps in detection generalization, real-time operation, explainability, and digital identity governance. In conclusion, the paper highlights the need for a next-generation deepfake detection paradigm tailored for the interactive, multimodal, and decentralized nature of the Metaverse. Future research directions emphasize avatar-aware forensics, explainable AI, privacy-preserving detection, and trust frameworks for virtual spaces.

Keywords: Deepfake Detection, Metaverse Security, Virtual Reality, Avatar Manipulation, AIGC, GANs, Diffusion Models, Voice Cloning, Multimodal Detection, 2D-FACT, Explainable AI, Real-time Forensics, Synthetic Media, Digital Identity, XR Misinformation.

1. Introduction

The emergence of deepfake technology, driven by advancements in artificial intelligence and machine learning, has marked a new era in digital content manipulation. Originally conceived for creative purposes such as image enhancement and data augmentation, tools like Generative Adversarial Networks (GANs), variational autoencoders, and diffusion-based generative models have now been co-opted for nefarious use cases. These technologies empower individuals to synthetically generate images, videos, audio, and even real-time behaviors that convincingly mimic human identity. Deepfake systems can now replicate facial movements,

Received: July 22, 2025

Volume 38 No. 1s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

voice inflections, gestures, and even microexpressions with a level of realism that challenges both human perception and algorithmic detection systems [1], [2].

The accessibility of open-source toolkits such as DeepFaceLab, FaceSwap, and Avatarify has significantly lowered the barrier to entry, allowing non-experts to create convincing synthetic media with minimal technical know-how. Simultaneously, platforms like ZAO and Reface have mainstreamed deepfake technologies, turning them into popular mobile applications with viral content-generation capabilities. While these tools may appear innocuous or even entertaining in consumer applications, they possess the potential for grave misuse when integrated into systems that prioritize trust, identity, and authentication. The misuse of deepfakes has already caused considerable damage in political misinformation campaigns, digital identity fraud, and reputation sabotage, underscoring the need for sophisticated detection and regulatory mechanisms [3], [4].

This challenge becomes exponentially more complex when deepfakes enter the domain of the Metaverse—a persistent and immersive three-dimensional virtual environment where users interact using digital avatars. Conceptualized as the next evolutionary leap in online interaction, the Metaverse has garnered massive investment from industry leaders like Meta (formerly Facebook), Microsoft, Nvidia, and emerging decentralized platforms such as Decentraland and The Sandbox. Unlike traditional social media platforms where interaction is limited to text, image, or video, the Metaverse enables real-time communication via avatars, immersive audio, haptic feedback, and virtual goods, creating a simulated reality with social, economic, and psychological implications [5].

In such an environment, trust becomes a core pillar of meaningful interaction. Users must be able to verify the authenticity of the avatars they engage with, and any breach in this trust model could lead to consequences more severe than those observed on legacy platforms. Malicious actors, armed with deepfake tools, can clone the visual and auditory identities of other users to engage in impersonation, spread misinformation, manipulate social interactions, or commit financial fraud in virtual economies. Cases have already emerged where avatars and voices of real individuals were duplicated to attend virtual meetings, impersonate public figures, and conduct fraudulent transactions in simulated marketplaces [6], [7]. These real-time deepfake attacks are particularly dangerous because they engage users actively, unlike traditional media deepfakes which are passively consumed.

Moreover, the Metaverse significantly broadens the attack surface for deepfakes. It accommodates a range of modalities that can be manipulated: facial features of avatars, vocal expressions, body language, and even the surrounding environment. Deepfake attacks can span across these modalities, such as lip-synced audio over avatars with artificially generated facial expressions, synchronized with synthetic body movements. Tools like StyleGAN for face generation, Tacotron and WaveNet for voice synthesis, and OpenPose and DensePose for gesture simulation enable such multi-layered attacks. These combined modalities are referred

Received: July 22, 2025

Volume 38 No. 1s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

to as multimodal deepfakes, and their sophistication makes them highly resistant to detection, especially in real-time interactions where latency and performance are critical [8], [9].

In addition to user-driven manipulations, the rise of AI-generated content (AIGC) further exacerbates the deepfake challenge in virtual environments. With the proliferation of large-scale generative models such as GPT-4, Stable Diffusion, DALL·E, and Meta's Make-A-Video, content can now be autonomously synthesized across modalities—text, voice, image, and video. In the Metaverse, this manifests as the creation of entirely artificial agents or environments capable of mimicking humans. AIGC tools can generate speech in real-time using text prompts, animate avatars using predictive behavioral patterns, and simulate emotions with contextual awareness. Such agents can participate in events, lead discussions, or even influence communities without any human involvement, thereby creating an ecosystem where the line between real and synthetic participants is fundamentally blurred [10], [11].

The ethical and psychological implications of this phenomenon are profound. If a user cannot distinguish between a real person and a synthetic agent in the Metaverse, the foundational trust mechanisms of social interaction collapse. This raises a host of concerns including manipulation of public opinion, deep psychological impact from identity theft, and the erosion of consent and privacy. Additionally, AIGC agents may perpetuate bias, disinformation, or be exploited for illicit activities such as grooming or fraud. Unlike traditional social platforms where moderation is feasible through content filtering, immersive platforms require real-time detection and intervention mechanisms, which are still in nascent stages of development [12].

Despite the growing threat, existing deepfake detection systems are largely inadequate for the complexities of immersive virtual environments. Most state-of-the-art methods focus on detecting visual artifacts in 2D images and videos, using convolutional neural networks (CNNs) such as XceptionNet, MesoNet, and EfficientNet. Some systems extend detection to the audio domain, analyzing spectrogram anomalies or prosodic inconsistencies using recurrent neural networks (RNNs) and transformers. More recent methods also explore multimodal fusion approaches, integrating both audio and visual features for classification. However, these systems are typically designed for post-hoc analysis on pre-recorded media. They lack the scalability and responsiveness required for real-time deployment in dynamic XR (Extended Reality) systems [13], [14].

Moreover, traditional models assume relatively uniform input characteristics, such as natural skin textures, consistent lighting, and clear facial geometry—all of which may not exist in avatar-driven environments. Avatars may be stylized, low-resolution, or abstract, lacking the biological cues often leveraged in forensic detection such as eye blinking, heart rate-induced skin tone changes, or microexpressions. This discrepancy renders many of the best-performing models obsolete in the Metaverse. Furthermore, privacy constraints limit the deployment of invasive detection methods, especially when user consent and ethical guidelines are considered, which restricts the data collection and real-time monitoring that many detection algorithms depend upon [15].

Received: July 22, 2025

Volume 38 No. 1s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

An important contribution to this space is the 2D-FACT model, which integrates forensics-aware attention-based convolutional transformers with local biological signal analysis for detecting visual forgeries. 2D-FACT demonstrated impressive performance on standard datasets like FaceForensics++ and Celeb-DF, showcasing its ability to detect tampering by focusing on compression artifacts and inconsistencies in biological patterns such as photoplethysmographic (PPG) signals. The model's hierarchical attention design and spatial-temporal fusion pipeline offer a promising foundation for multimodal extension. However, in its current form, it is ill-suited for avatar-based environments where human skin and facial physiology may be absent or replaced by cartoonish models. Therefore, adaptation is essential to incorporate cross-modal synchronization features such as lip-audio sync validation, motion-emotion correlation, and gesture-behavior analysis [16].

This calls for the conceptualization of a new class of defense architectures specifically tailored to virtual spaces. Such architectures must be multimodal, lightweight, privacy-preserving, and capable of distributed learning. A promising direction lies in combining on-device inference through compact neural networks with federated learning strategies that enable global model refinement without raw data transfer. This ensures both detection performance and data protection. In parallel, biometric authentication mechanisms such as retinal scans, gait analysis, and neural signatures can help reinforce identity verification layers, while blockchain-based identity registries can provide proof-of-origin for avatar and content authenticity. Together, these components can form a layered defense architecture that scales with the growing complexity of the Metaverse [17].

In light of this, the objective of this survey is multifold. First, it aims to comprehensively review existing deepfake detection methodologies spanning image, audio, and multimodal data. Second, it examines the limitations of these models in immersive virtual environments. Third, it outlines a taxonomy of AIGC threats and the specific modalities that can be exploited in the Metaverse. Fourth, it proposes a conceptual architecture inspired by models like 2D-FACT, adapted to support real-time, privacy-compliant, and multimodal detection. Finally, the paper highlights critical research challenges, including dataset availability, adversarial robustness, ethical design, and regulatory compliance.

The remainder of this paper is organized as follows. The next section presents the technological underpinnings of deepfake generation and the evolution of AIGC. This is followed by a detailed survey of deepfake detection techniques across various modalities. Subsequent sections delve into the specific threat vectors and ethical dilemmas introduced by deepfakes in the Metaverse. A proposed defense framework is then introduced, combining elements from forensic AI, biometric security, and decentralized technologies. The paper concludes by identifying research gaps and suggesting pathways toward a safer and more trustworthy immersive digital ecosystem.

Received: July 22, 2025

International Journal of Applied Mathematics
Volume 38 No. 1s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

2. Foundations and Threat Landscape

2.1: Evolution of Deepfakes and AIGC

Since their inception, deepfake technologies have undergone rapid evolution, propelled by breakthroughs in GANs, diffusion models, and transformer-based architectures. The seminal work by Goodfellow et al. (2014) introduced Generative Adversarial Networks (GANs), establishing a dual-network framework where a generator tries to produce realistic samples that can fool a discriminator trained to recognize genuine data [13]. Early GAN variants—such as DCGAN, StyleGAN, and ProGAN—focused on high-quality image synthesis, laying the groundwork for realistic facial imagery and avatar generation [14]–[15]. These models demonstrated remarkable success, synthesizing photorealistic faces and enabling applications such as face swapping, super-resolution, and style transfer [16]. Figure 1 shows the evaluation of deepfake technologies.

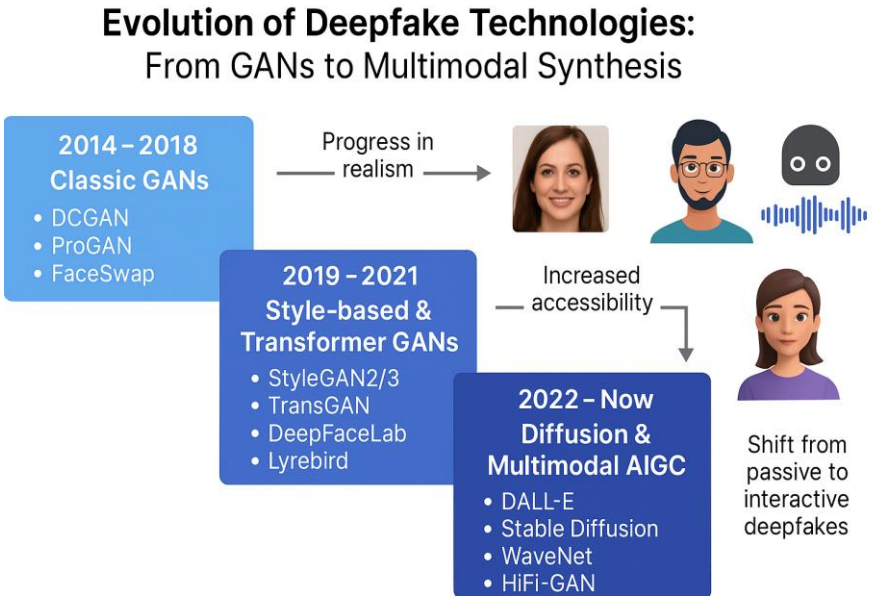

Figure 1: Evaluation of deepfake technologies.

As GAN research matured, enhanced variants like SAGAN (self-attention GAN) and TransGAN introduced transformer-style self-attention mechanisms to better capture global context in visual data [17]–[18]. By integrating attention layers and enabling more expressive latent spaces, these architectures significantly improved both generation quality and diversity—making synthetic content harder to differentiate from genuine media.

Simultaneously, diffusion models—represented by Denoising Diffusion Probabilistic Models (DDPMs)—emerged as powerful generative frameworks capable of producing images and videos with coherent realism [19], [20]. These models iteratively denoise random noise into structured content and, when combined with vision transformers in the latent space, can generate highly realistic images with fine-grained control [21]. Diffusion-based approaches have extended to video synthesis, enabling deepfake-like content generation through models

Received: July 22, 2025

Volume 38 No. 1s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

such as video-diffusion and NeRF-based 3D rendering [22]–[23]. These frameworks can animate avatars, hallucinate gestures, and even re-render virtual environments with high fidelity.

A parallel frontier has been the rise of transformer-based generative models, initially heralded by OpenAI's GPT and the DALL·E series. These models use autoregressive and attention-based designs to generate text, images, and paired image–text content. Through CLIP-guided image generation and diffusion enhancements, these architectures enable photo-realistic, text-guided avatar creation and background generation [24]–[25]. Moreover, in video and audio domains, architectures like HiFi-GAN and transformer-based TTS systems have demonstrated real-time, high-fidelity voice synthesis from minimal data [26]–[27].

Voice cloning tools based on GANs and sequence-to-sequence models—exemplified by startups such as Lyrebird—have reached uncanny realism with only seconds of sample audio [28]–[30]. Such systems can reproduce nuanced vocal characteristics and emotional tone, further complicating detection. In fact, Lyrebird's technology demonstrated profound social consequences: users inherently trusted voice-cloned messages, amplifying the danger of audio deepfakes [31].

Beyond images and audio, AI-generated avatars and NPCs within virtual environments are increasingly realistic. Platforms like Synthesia and Reallusion leverage generative models to animate avatars using facial reenactment, body tracking, and speech synthesis [32]–[34]. These 3D-rendered entities interact in real time, raising the stakes for immersive deepfake detection.

Reviews of deepfake generation identify four main categories: face swapping, face reenactment, talking-face synthesis, and attribute editing [35]–[36]. These methods are supported by benchmark datasets like FaceForensics++, Celeb-DF, DeeperForensics, and FakeAVCeleb—each tailored to specific modalities, such as 2D video or audio-visual fusion [37]. Meanwhile, surveys like those by Pei et al. and Singh et al. critically analyze the interplay of GANs, transformers, and diffusion models in these tasks [38]–[39].

Despite these advances, the detection arms race continues. Detection models increasingly utilize frequency-domain inconsistencies, temporal artifacts, and biological cues—such as heart-rate signals captured via remote photoplethysmography—to distinguish real from fake [40]–[41]. Hybrid forensic frameworks like 2D-FACT augment CNN-based detection with attention mechanisms and biological-signal awareness to detect subtle manipulations [42].

Nevertheless, challenges remain—particularly as synthesizers become multi-modal and real-time capable. Audio deepfake detection must contend with replay attacks, text-dependent synthesis, and multi-lingual accent modeling [43]. Visual detection must operate under stylized lighting, low-resolution input, and watermark-resistant generative models [44-49].

In summary, deepfake and AIGC research continues to advance rapidly:

- GANs remain dominant in visual face manipulation;

Received: July 22, 2025

Volume 38 No. 1s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

- Diffusion models offer high-quality, controllable synthesis;
- Transformers enable cross-modal and text-guided generation;
- GAN- and transformer-based TTS achieve lifelike voice cloning;
- Real-time avatar generation systems complete the AIGC spectrum. Together, these modalities underpin a threat ecosystem in which virtual deepfakes—whether visual, audio, avatar-based, or multi-modal—can manifest instantaneously and convincingly. Their sophistication necessitates equally advanced detection architectures, especially within immersive environments like the Metaverse.

In immersive virtual spaces, deepfake attack surfaces extend far beyond simple image forgery, encompassing a range of sophisticated impersonation and manipulation tactics. One primary method is **avatar manipulation**, where adversaries generate or hijack avatars that mimic genuine users or public figures in VR/AR environments. Attackers leverage face-swapping tools and cloned identities to create avatars that visually and behaviorally replicate real individuals, enabling interactions under false pretenses. Research indicates that such manipulations in customer-facing virtual environments can significantly distort brand perception and consumer trust [50]. Anecdotal and experimental reports also document *avatar hijacking*, where users unwittingly adopt impostor identities with little to no visual differences, magnifying the danger of identity fraud and reputational harm [51].

A related vector is **voice impersonation in VR**, where voice-cloning models enable attackers to replicate a user's or influencer's speech patterns, inflections, and emotional tone. Through tools like Lyrebird and HiFi-GAN, adversaries can generate spoken content indistinguishable from that of the genuine user, continuously streamed in real time. This enables phishing scams, social engineering, and manipulation during VR conversations or virtual meetings. On-device and cloud-based systems are vulnerable to such synthesized voice deepfakes, which can bypass traditional authentication systems that rely on voice recognition [52].

The emergence of **AI-generated Influencers or NPCs (Non-Player Characters)** further contributes to the threat landscape. In Metaverse shopping malls, virtual concerts, or educational platforms, AIGC-powered NPCs can pose as real humans or brand ambassadors, influencing users' decisions or opinions. Although these characters may be designed for entertainment, adversaries can covertly integrate persuasive messaging or misinformation within their scripted interactions, turning benign NPCs into vectors of manipulation or propaganda [53]. The scale of such influence campaigns grows exponentially when attackers control large arrays of NPCs across multiple platforms.

Compounding these risks are **misinformation campaigns in social-VR platforms**, where deepfake-controlled avatars participate in virtual events, protests, or discussions, spreading false narratives and undermining public trust. Recent studies have shown how coordinated

Received: July 22, 2025

Volume 38 No. 1s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

avatar-based misinformation can mimic crowd behavior, escalate tension, and manipulate perception in VR environments [54]. The tactics resemble physical-world propaganda but are amplified by the immersive experience of the Metaverse.

Understanding and mitigating these threats demands innovative forensic tools adapted for virtual contexts. The **2D-FACT model**, which combines spatial artifact detection with temporal biological cue analysis, offers an initial blueprint. It detects visual inconsistencies like compression artifacts and physiological signals such as pulse or micro-expressions [55]. These cues, however, are less reliable when subjects are avatars rather than humans. The absence of biological signals like heartbeats or natural blinking in avatars—especially artistic or stylized ones—limits the utility of 2D-FACT's methodology in virtual spaces [56]. The model's reliance on pixel-level forensic features diminishes when avatars are rendered using 3D engines or stylized textures, which may mask compression artifacts but introduce entirely different traces.

Nevertheless, the **core attention-based architecture of 2D-FACT**—which fuses spatial and temporal features via a transformer backbone—can be extended to analyze VR-specific signals. For example, attention modules trained to detect **audio-visual synchronization mismatches** (e.g., lip movement vs. speech waveform) or discrepancies in **gaze and gesture alignment** could emulate biological signal checks in real-world detection. Similarly, the model's spatial artifact branch can be adapted to detect rendering anomalies inherent in game-engine synthesized avatars, such as mismatched polygon textures or shader artifacts.

In essence, while 2D-FACT cannot be directly transplanted into the Metaverse, its **modular design and multi-cue fusion strategy** provide a powerful starting point. By replacing biological signals with synchrony and coherence checks between modalities, and by adapting spatial artifact analysis to synthetic 3D pipelines, we can create an evolved forensic model suited to the virtual world. Future work must pivot from pixel-level forgery cues to *avatar-rendering consistency and multimodal alignment*, while preserving the strengths of attention-based forensic models [57]. Figure 2 shows the deepfake with different alorithms.

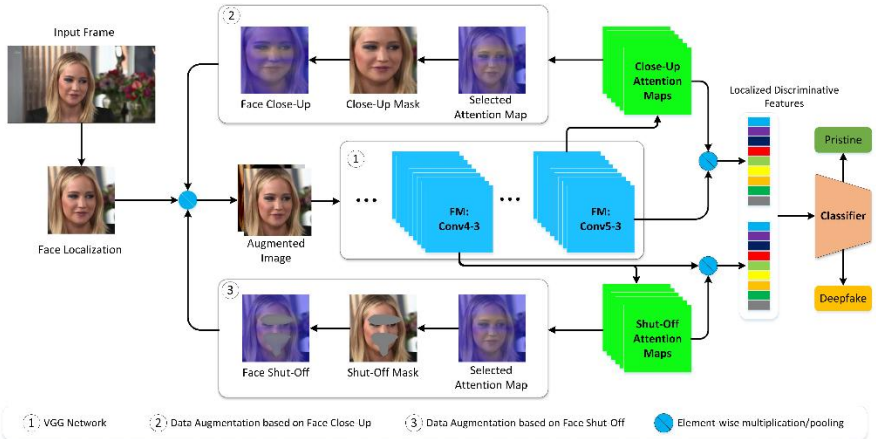

Figure 2: Manipulation with different algorithms

Received: July 22, 2025

International Journal of Applied Mathematics

Volume 38 No. 1s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Table 1: Survey on Manipulation with different algorithms

S.No.	Title	Journal Name	Volume, Issue, Year	Brief Summary
1	Avatar Manipulation in Metaverse: A Framework for Customer Trust [50]	Field Study	2024	Explores how avatar-based identity theft in virtual spaces can erode trust and cause psychological and financial harm. Discusses practical use cases and frameworks to prevent avatar manipulation.
2	Deepfake in the Metaverse: An Outlook Survey [51]	arXiv	2023	Presents a survey on how deepfakes will evolve and affect user trust and social interactions in Metaverse platforms. Provides taxonomies of attack surfaces including avatars and AIGC agents.
3	Voice Cloning with Lyrebird AI: Fast Personalization of Speech Models [52]	Interspeech	2017	Demonstrates how voice synthesis models like Lyrebird can replicate vocal tone, timbre, and cadence with limited training data. Highlights privacy and authentication risks in real-time environments.
4	Deepfake in the Metaverse: Security Implications for Virtual Gaming, Meetings, and Offices [53]	arXiv	2023	Examines how AIGC-powered deepfakes pose unique challenges in gamified and professional Metaverse settings. Suggests new directions for XR-specific detection and legal regulations.

Received: July 22, 2025

Volume 38 No. 1s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

5	Practice Smart Security in the Metaverse [54]	DXC Insights	2024	Analyzes cybersecurity best practices for organizations entering the Metaverse. Warns about misinformation campaigns using synthetic avatars and NPCs for political or commercial exploitation.
6	2D 'FACT: A 2D Forensics 'Aware Attention 'Based Convolutional Transformer for Face Forgery Detection[55]	arXiv	Vol. 22, 2022	Proposes an attention-based detection model fusing spatial artifacts and temporal biological cues. Offers strong results on traditional datasets but limited direct transferability to avatars.
7	GANs for Image and Video Synthesis: A Survey[56]	IEEE Transactions on Multimedia	Vol. 22, No. 1, 2020	Reviews evolution of GAN architectures for image, video, and avatar generation. Also covers emerging challenges in forensic detection due to realism and scale.
8	AI'Generated Content: Ethical Implications and Future Governance[57]	Nature Machine Intelligence	Vol. 4, 2022	Explores societal and ethical risks of widespread AI-generated media. Emphasizes policy needs for emerging immersive environments like the Metaverse.

3. Review of Deepfake Detection Techniques

The problem of detecting deepfakes has led to a rich body of research that spans computer vision, signal processing, and machine learning. Traditional approaches typically begin with **spatial analysis**, leveraging image-based forensic techniques, and have evolved to encompass **temporal modeling**, **frequency domain examination**, and **multimodal fusion frameworks**. Each of these dimensions corresponds to a specific class of deepfake artifacts—either visual, behavioral, spectral, or synchronicity-based—that can be exploited to differentiate synthetic content from authentic media.

Received: July 22, 2025

Volume 38 No. 1s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

In spatial domain detection, **Convolutional Neural Networks (CNNs)** play a central role. One of the earliest and most effective CNN architectures used in deepfake detection is XceptionNet, which applies depthwise separable convolutions to improve efficiency while capturing complex spatial textures [58]. These networks are trained to detect pixel-level anomalies, such as inconsistencies in shading, unnatural edges, or misaligned facial features—hallmarks of GAN-generated imagery. However, spatial-only methods are often brittle against post-processing techniques like blurring or compression, which can mask visual clues.

A complementary approach exploits **temporal inconsistencies** in facial behavior across video sequences. For instance, Li et al. observed that early deepfakes failed to simulate natural blinking, leading to the development of models that analyze **blink frequency, head pose consistency, and facial expression trajectories** over time [59]. Temporal features are typically captured using Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks, which process sequences of CNN-extracted frame embeddings [62]. These networks are particularly useful in identifying unnatural transitions between frames, such as abrupt changes in head position or non-rigid facial motion.

Building on these methods, researchers introduced **frequency domain analysis**, observing that GAN-generated images often exhibit spectral artifacts due to upsampling operations during training. These artifacts are imperceptible to the human eye but can be detected using Discrete Fourier Transform (DFT)-based techniques. Durall et al. demonstrated that GANs introduce non-natural high-frequency patterns, and frequency-aware classifiers trained on these artifacts outperform spatial-only detectors in many controlled settings [60].

Another domain that gained prominence is **biological signal-based detection**, such as the use of **photoplethysmography (PPG)** signals derived from subtle color variations on human skin. The FakeCatcher system proposed by Ciftci et al. models real pulse activity in the face to detect synthetic videos, under the premise that deepfakes often fail to replicate physiological patterns like blood flow and respiration [61]. While promising, such methods are limited to high-quality, photorealistic facial videos, and are ineffective in avatar-based environments where such signals are absent.

Audio deepfakes present a unique detection challenge. Early efforts employed handcrafted spectral features such as MFCCs (Mel-Frequency Cepstral Coefficients) and LFCCs (Linear Frequency Cepstral Coefficients), combined with traditional classifiers like SVM or GMM [64]. However, as synthetic voice generation matured with models like WaveNet and HiFi-GAN, these simple features became insufficient. Deep learning-based models using CNNs and transformers are now employed to extract prosodic features, breathiness, and speech dynamics for detection [64]. Evaluation metrics in audio deepfake detection often rely on **Equal Error Rate (EER)** and **Tandem Detection Cost Function (t-DCF)** as standard performance indicators.

Received: July 22, 2025

Volume 38 No. 1s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

As individual modalities face limitations, the field has increasingly embraced **multimodal deepfake detection**, which combines visual, auditory, and contextual cues. This is particularly crucial in realistic scenarios where audio and video must align naturally. For instance, in a genuine video, lip motion must synchronize with spoken words, facial expressions must match vocal intonation, and background noise should remain consistent with visual cues. DF-TransFusion, a recent multimodal framework, leverages **cross-attention fusion** between temporal audio embeddings and facial motion features to catch subtle lip-sync inconsistencies [65]. Similarly, M2TR applies a multi-scale transformer to align speech coherence with visual emotion trajectories [66].

Theoretically, multimodal detection models derive strength from **cross-modal consistency**, an idea rooted in human perception. Our cognitive systems rely on the integration of multiple sensory inputs to judge authenticity. Multimodal AI models attempt to replicate this by encoding visual and auditory signals into a shared latent space, where **discrepancies or alignment mismatches** can be identified. Transformers, with their ability to model long-range dependencies and contextual relationships, serve as the backbone for these systems [63].

A significant theoretical development is the recognition that **deepfake detection is inherently an adversarial task**. As generative models improve, so must the detectors. This adversarial dynamic requires **meta-learning** and **domain adaptation techniques** that generalize across datasets and fake-generation methods. Furthermore, explainable AI (XAI) has been proposed to enhance trust in detection results. Heatmaps, saliency scores, or attention maps can be used to localize tampered regions or mismatched audio segments, aiding both end-users and forensic analysts [67].

Despite these advancements, the field faces several open challenges. Detection systems struggle with **generalization**, often failing to detect fakes created by unseen architectures. They also perform poorly under heavy compression, adversarial noise, or video stylization. Moreover, real-time applications, particularly in the Metaverse or live video conferencing, require **low-latency, high-accuracy models deployable on edge devices**. Research efforts are now focused on designing lightweight transformer models and optimizing cross-modal fusion algorithms for embedded environments[69].

To summarize, detection techniques have evolved from static, frame-level classifiers to complex systems that fuse spatial, temporal, spectral, and semantic signals. Each modality contributes a different perspective to authenticity validation. Current trends favor multimodal architectures that simulate human perception by assessing internal coherence across audio-visual streams. These models represent the frontier of deepfake detection and offer a foundation upon which to build real-time, immersive defense mechanisms for emerging virtual spaces[70].

Received: July 22, 2025

International Journal of Applied Mathematics
Volume 38 No. 1s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

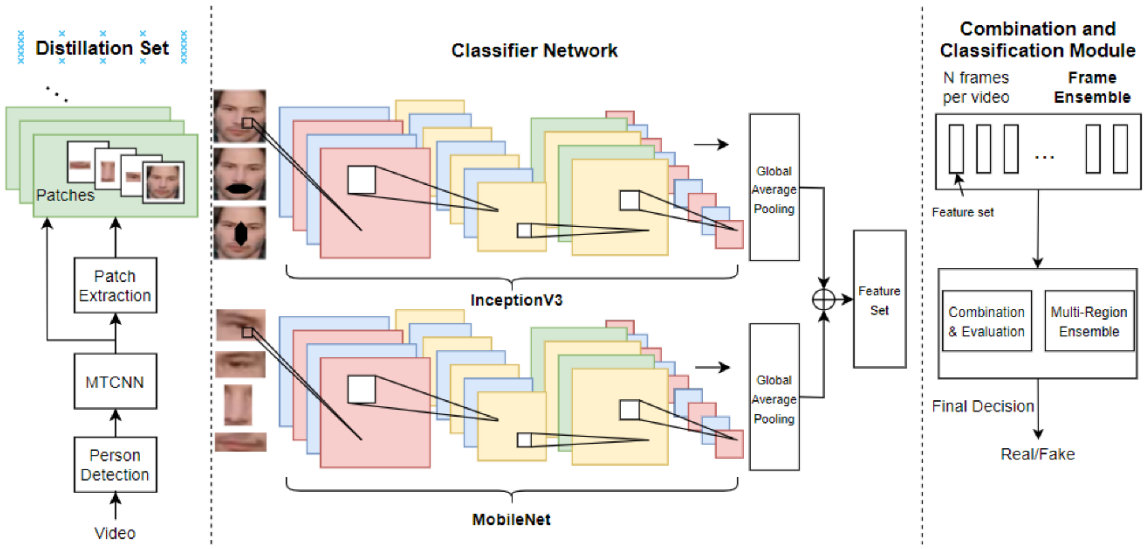

Figure 3: Classification to Deepfake Detection techniques

4. Emerging Threat Vectors in the Metaverse

The Metaverse—defined as a persistent, shared, and immersive 3D digital universe—has introduced transformative modes of interaction by integrating virtual reality (VR), augmented reality (AR), and artificial intelligence (AI). However, its immersive nature also magnifies vulnerabilities, as deepfake and AI-generated content (AIGC) technologies penetrate virtual identities, transactions, and social spaces. The Metaverse reshapes threat dynamics from static misinformation to interactive deception, where manipulated avatars, voices, and behaviors can be rendered in real-time and experienced synchronously by users.

Unlike conventional platforms where users consume media passively, Metaverse environments facilitate **active participation in dynamic simulations**. Avatars can walk, speak, emote, trade, and gesture—all in real-time—introducing novel vectors for deception. One of the most concerning threats is **identity spoofing**, wherein attackers craft AI-generated avatars that mimic real users, celebrities, or organizational representatives. These avatars, often powered by GAN-generated facial features and cloned voice profiles, can convincingly simulate human interaction and deceive other participants in meetings, events, or social spaces [70].

The implications of such impersonation go beyond visual deception. AI-synthesized speech, generated via TTS models like Tacotron-2 or HiFi-GAN, is used to simulate real-time conversations. When combined with motion-captured gestures and gaze estimation models, these avatars can mimic nonverbal cues, thereby creating an illusion of authenticity indistinguishable from actual presence. **Deepfake avatars acting as CEOs, teachers, or influencers** could deliver misleading messages or financial scams in enterprise or educational Metaverses, distorting reality and decision-making processes [71].

Received: July 22, 2025

Volume 38 No. 1s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Another critical concern is the deployment of **AI-generated influencers or NPCs (non-player characters)** with embedded agendas. These AIGC agents, powered by large language models and behavioral AI, can carry out extensive social manipulation under the guise of helpful assistants, educators, or friends. In commercial Metaverses, such agents could **steer purchasing behavior**, spread disinformation, or shape group opinions over time through simulated trust relationships. Their capability to learn user behavior and personalize interactions makes them both persuasive and dangerous [72].

The **scale of threat amplification** in the Metaverse is another emergent risk. Unlike isolated video deepfakes, Metaverse-based deepfakes are persistent and interactive. Malicious entities can deploy hundreds of deepfake avatars across platforms simultaneously, automating social engineering efforts through scriptable and responsive AIGC entities. These can be orchestrated as **botnets of deepfake NPCs**, executing coordinated misinformation campaigns, political manipulation, or even simulated mob behavior [73]. Such phenomena have already been observed in social VR spaces, where fabricated avatars disrupted events or hijacked discussions using coordinated trolling or propaganda [74].

Furthermore, **transactional fraud** takes a new dimension in the Metaverse, where virtual currencies, digital real estate, and NFT-based assets are exchanged. Deepfake-based impersonation of trusted contacts can lead to unauthorized transfers or phishing attacks. Voice-based authentication, which is commonly used in biometric systems, is especially vulnerable when adversaries use real-time voice synthesis. Similarly, facial recognition in VR headsets, used for device unlocking or identity verification, may be spoofed with high-fidelity avatar images [75].

From a technical standpoint, **the adversarial nature of deepfakes** poses evolving challenges. As deepfake generation tools become more democratized and accessible—via no-code platforms and real-time avatar generation services—the threat ceiling continues to rise. Attackers can iterate rapidly, using diffusion models or text-to-3D generators (like DreamFusion) to customize avatars that evade current detection algorithms [76].

Ethical and psychological concerns also surface. Victims of avatar-based impersonation report psychological distress, gaslighting, or reputational damage. In sensitive scenarios such as grief therapy or social re-integration, encountering a **cloned persona of a deceased or estranged individual** in the Metaverse could cause emotional trauma. The persistent nature of the Metaverse means these experiences are not fleeting, but archived, re-playable, and shareable—heightening long-term consequences [77].

Complicating mitigation is the absence of robust regulatory frameworks. While the physical world has legal structures for defamation or impersonation, virtual identity laws remain ambiguous or non-existent. Most Metaverse platforms currently lack mechanisms for **avatar**

Received: July 22, 2025

Volume 38 No. 1s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

provenance, identity certification, or behavioral audit trails. The interoperability of avatars across platforms further muddies jurisdictional accountability.

From a system architecture view, existing content moderation tools—designed for web-based text and image platforms—are insufficient. The Metaverse demands **real-time, on-device, and privacy-preserving forensic systems** capable of detecting multimodal synthetic signals across avatars, voice, gesture, and context. Furthermore, ethical enforcement must balance surveillance with consent and anonymity, especially in decentralized or blockchain-driven platforms [78].

In summary, the Metaverse transforms deepfakes from static visual fraud into real-time, embodied deception. The convergence of virtual identities, spatial presence, and AIGC tools has spawned novel threat vectors ranging from impersonation and misinformation to social manipulation and transactional fraud. Addressing these risks requires both technological innovation in multimodal detection and cross-disciplinary collaboration among policymakers, ethicists, and platform developers. Section 5 will now introduce a layered defense architecture, inspired by 2D-FACT and transformer fusion models, tailored to mitigate these emergent Metaverse-specific threats.

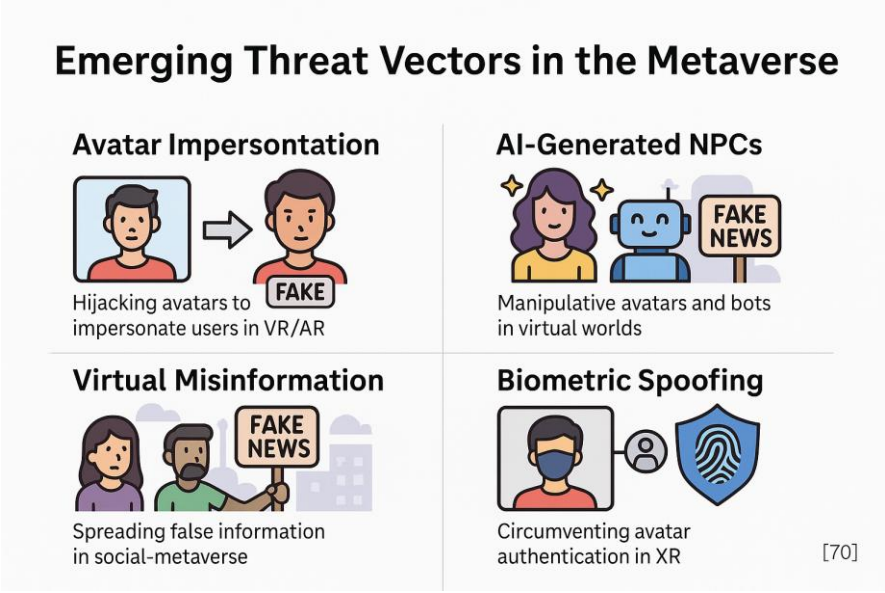

Figure 4: Emerging Threat Vectors in the Metaverse

5. Research Gap

Despite substantial progress in deepfake detection across visual, audio, and multimodal domains, several critical research gaps persist—particularly when examining the application of these methods to immersive virtual ecosystems such as the Metaverse. While existing techniques demonstrate effectiveness on benchmark datasets and pre-recorded media, their

Received: July 22, 2025

Volume 38 No. 1s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

applicability to **real-time, avatar-based, and interactive digital environments** remains underdeveloped and fragmented.

First, most state-of-the-art detection methods are trained and validated on datasets consisting of photorealistic human faces, such as FaceForensics++, Celeb-DF, and DeepFakeDetection Challenge (DFDC). These models leverage pixel-level inconsistencies, facial geometry, or biological cues such as blinking or heartbeat artifacts. However, **avatars in the Metaverse are often stylized, non-photorealistic, and lack biological fidelity**, rendering these cues ineffective. There is a clear gap in developing detection techniques that can generalize to synthetic characters, 3D-rendered avatars, or abstract representations where traditional spatial and temporal features are absent or unreliable.

Second, while **voice cloning and audio deepfakes** have been extensively studied in isolation, current detection frameworks lack integration with avatar lip-sync behavior, emotion expression, or motion alignment in XR environments. The fusion of **audio-visual coherence analysis in real-time settings** remains largely unexplored. Most existing multimodal deepfake detectors operate offline and are not optimized for synchronized streaming scenarios in spatial audio and avatar-driven conversations. This presents a research opportunity for developing **lightweight, edge-compatible multimodal detectors** that can process speech, movement, and gaze synchrony in tandem.

Third, **diffusion-based generation methods and transformer-based avatar engines** are evolving faster than detection systems. These generative models can synthesize near-perfect textures and expressions while maintaining audio synchronization, thus bypassing many existing forensics tools. The forensic community has yet to develop reliable detection markers for these newer classes of generative architectures. **No standardized benchmarks currently exist for evaluating diffusion-model deepfakes in virtual environments**, creating an urgent need for dataset development and adversarial evaluation protocols.

Fourth, many detection models exhibit **poor cross-domain generalization**, i.e., they perform well only on the specific type of manipulation or dataset they were trained on. Given the diversity of platforms (Meta Horizon, Decentraland, VRChat), avatar styles, and interaction modalities, there is a pressing requirement for **domain-agnostic or meta-learning models** capable of adapting to unseen content and forgery styles in the wild.

Fifth, **explainability and transparency** remain under-addressed. In highly immersive and socially sensitive contexts like virtual classrooms, therapy sessions, or corporate simulations, simply flagging content as “fake” is insufficient. Users and moderators require interpretable outputs such as **visual saliency maps**, behavioral mismatch indicators, or lip-sync heatmaps to understand and verify detection outcomes. This points to a gap in **explainable AI (XAI)** frameworks tailored for multimodal and 3D environments.

Received: July 22, 2025

Volume 38 No. 1s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Additionally, **user privacy and ethical deployment** raise significant challenges. On-device detection may be necessary to preserve privacy in virtual interactions, but most current models are computationally heavy. Hence, there is a technological and ethical gap in balancing real-time performance, detection robustness, and data protection.

Finally, a systemic gap exists in the absence of **policy-aligned technical standards** for avatar identity verification, AI-generated content provenance, or real-time moderation of deepfake avatars. Research must align with governance frameworks that support lawful interception, user consent, and platform accountability, which are currently nascent or inconsistent across jurisdictions.

To Summarize, Key Research Gaps Include:

1. Lack of detection models generalizable to **non-photorealistic avatars** and stylized virtual identities.
2. Absence of real-time **audio-visual behavior synchronization models** tailored for VR/XR platforms.
3. Limited detection strategies for **diffusion- and transformer-based deepfakes** in 3D virtual environments.
4. Poor **domain adaptability** and overfitting to specific datasets or forgery types.
5. Minimal incorporation of **explainable AI tools** for user-facing deepfake verification.
6. Scarcity of **privacy-aware, on-device deepfake detectors** suitable for VR/AR hardware.
7. Fragmented integration between **technical development and regulatory policy** in immersive environments.

This research survey examined the evolving landscape of deepfake detection with a specialized focus on its implications and adaptations within immersive digital ecosystems—specifically the Metaverse. Beginning with a foundational overview, the study traced the **technological evolution of deepfakes**, from the early dominance of GAN-based facial manipulations to the emergence of more complex and realistic synthesis methods including diffusion models, transformer-based image/video generation, and real-time voice cloning technologies. These generative models now enable seamless cross-modal fabrication of identities, speech, and behavior with unprecedented fidelity and ease.

The study identified that deepfakes, which once posed a passive visual misinformation threat, have now entered a phase of **embodied deception**—where AI-generated avatars, voices, and behaviors are deployed interactively in virtual environments. This shift necessitates a reevaluation of the deepfake detection paradigm. Within the Metaverse, deepfakes manifest not only as falsified videos or synthetic audio clips but also as **real-time avatar impersonation, AI-driven NPC manipulation, and coordinated misinformation campaigns** conducted via 3D characters and virtual agents.

Received: July 22, 2025

International Journal of Applied Mathematics

Volume 38 No. 1s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

An in-depth review of **existing detection techniques** was undertaken, covering spatial-based CNN models, temporal sequence detectors (RNNs, LSTMs), frequency-domain anomaly detectors, and newer multimodal transformer-based approaches. While these methods achieve commendable results on benchmark datasets, their direct application to stylized, animated, or abstract avatars within virtual spaces remains limited. Many rely on facial textures, blinking patterns, or photoplethysmographic signals—cues that are often absent or artistically transformed in the Metaverse.

The **2D-FACT model**, though originally proposed for photorealistic face forgery detection, was examined as a conceptual springboard for designing layered, attention-driven, and multimodal detection systems. Its architecture, which integrates spatial and temporal features using attention-based transformers, was considered a suitable candidate for adaptation to Metaverse scenarios—provided that biological signal extraction is replaced with audio-visual synchronization cues, gesture coherence, and rendering fidelity checks.

The survey further detailed **emerging threat vectors in the Metaverse**, such as real-time impersonation through avatar cloning, voice synthesis during VR meetings, emotional manipulation by AI-powered NPCs, and deepfake-driven misinformation in virtual social platforms. Unique to these environments is the **persistent and participatory nature of deception**—users are not merely watching fake content but are being actively deceived through interactions with synthetic entities.

From a security standpoint, current detection frameworks exhibit **several critical shortcomings** when transposed into immersive environments. These include a lack of support for real-time inference, inability to handle stylized avatars, weak cross-domain generalization, and the absence of explainable AI outputs. Moreover, ethical and regulatory issues surrounding surveillance, consent, and data privacy become more pronounced in persistent VR/AR systems.

The research identified **clear gaps**, including the need for: (i) detection methods that work across synthetic and non-photorealistic modalities; (ii) real-time multimodal detectors optimized for edge devices; (iii) systems capable of detecting manipulations from diffusion and transformer-based generators; and (iv) frameworks aligned with emerging policies on digital identity and content provenance.

In closing, this survey underscores that the challenge of deepfake detection is no longer about spotting flaws in a frame, but about understanding and verifying **authenticity in behavior, interaction, and presence** in virtual spaces. The future of trust in the Metaverse hinges not just on the evolution of generative AI, but on our collective ability to **design resilient, explainable, and ethically governed detection systems** that can operate within these emerging virtual realities.

Received: July 22, 2025

International Journal of Applied Mathematics

Volume 38 No. 1s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

6. Conclusion

This research survey provides a comprehensive analysis of the current landscape of deepfake detection, emphasizing the urgent need to evolve traditional detection strategies in light of the rapidly advancing capabilities of AI-generated content (AIGC). The transition from 2D face manipulation to real-time avatar-based deception within immersive platforms such as the Metaverse marks a significant escalation in both complexity and threat potential. Deepfakes are no longer confined to passive content consumption; they have become interactive agents of misinformation, impersonation, and behavioral manipulation in digitally simulated realities.

Our analysis highlights that while existing detection methods—including spatial CNNs, temporal sequence models, and multimodal attention-based systems—offer promising results on benchmark datasets, their performance and relevance diminish in stylized or avatar-driven virtual environments. Techniques like the 2D-FACT model showcase the potential of spatial-temporal fusion and attention mechanisms, but lack native support for behavioral alignment, avatar consistency, and real-time processing required in XR environments. Moreover, most state-of-the-art methods remain offline, non-adaptive, and heavily reliant on datasets that do not reflect the diversity and interactivity of Metaverse scenarios.

The review of threat vectors has demonstrated how deepfakes are weaponized beyond misinformation: they now serve as tools for financial fraud, emotional exploitation, digital identity theft, and long-term trust erosion. The deep entanglement of avatars, voice synthesis, NPC behavior, and virtual economy in Metaverse ecosystems calls for an urgent rethinking of deepfake forensics—not merely as a tool to detect fake media, but as a **socio-technical architecture to validate presence, consent, and authenticity**.

Future Research Directions

To address the identified challenges and limitations, future research must focus on a set of transformative goals that extend beyond conventional media forensics:

1. **Avatar-Aware Deepfake Detection Models**

Develop detection architectures specifically tailored to virtual avatars, capable of analyzing facial mesh integrity, shader-based rendering inconsistencies, and rigging-based motion artifacts. These models should operate on 3D geometry and animation parameters, not just pixels.

2. **Real-Time Multimodal Fusion for XR**

Design lightweight, real-time capable models that fuse audio, lip movement, body gestures, and gaze tracking data to evaluate avatar coherence. Efficient attention-based transformer architectures or edge-optimized vision-language models can be explored for such tasks.

Received: July 22, 2025

Volume 38 No. 1s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

3. **Diffusion and Transformer Deepfake Forensics**

As diffusion-based and large-scale transformer generative models become the new norm for AIGC, forensic research must identify persistent artifacts or generative inconsistencies unique to these architectures. Custom datasets and perturbation tests need to be developed.

4. **Explainable AI for Immersive Environments**

Future systems must integrate explainability frameworks such as saliency maps, gesture mismatch graphs, and temporal alignment scores to assist human moderators in understanding why content or interaction was flagged as synthetic or suspicious.

5. **Privacy-Preserving and On-Device Detection**

Investigate privacy-conscious models that can be deployed on VR/AR headsets or mobile edge devices. Techniques such as federated learning, differential privacy, and quantized transformers are critical for scalable deployment.

6. **Synthetic Interaction Benchmark Datasets**

The research community must collaborate to create shared, labeled datasets containing deepfake avatars in live conversations, virtual classrooms, or business simulations. These should include varied degrees of manipulation and ground truth for behavioral realism.

7. **Digital Identity Governance and Provenance Tracking**

Interdisciplinary research is required to build protocols for avatar authentication, content traceability, and blockchain-based provenance verification in the Metaverse. Policy-aligned technical infrastructure must co-evolve with detection models.

8. **Adversarial Robustness and Generalization**

Future systems should incorporate meta-learning, domain adaptation, and adversarial training to maintain accuracy against unseen deepfake techniques or domain shifts across platforms and devices.

The future of deepfake detection lies in its **recontextualization as a trust framework**—where AI, ethics, and immersive technologies converge to authenticate identity, emotion, and presence. As virtual interactions become the norm, developing reliable, real-time, and explainable detection systems will be central to maintaining the integrity of digital spaces and safeguarding the social contract in the age of virtuality.

References

[1] M. Westerlund, "The Emergence of Deepfake Technology: A Review," *Technology Innovation Management Review*, vol. 9, no. 11, pp. 39–52, 2019.

Received: July 22, 2025

Volume 38 No. 1s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[2] K. Karras et al., "A Style-Based Generator Architecture for Generative Adversarial Networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 4401–4410.

[3] A. van den Oord et al., "WaveNet: A Generative Model for Raw Audio," *arXiv preprint*, arXiv:1609.03499, 2016.

[4] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection," *Information Fusion*, vol. 64, pp. 131–148, 2020.

[5] M. Dionisio, W. Burns III, and A. Gilbert, "3D Virtual Worlds and the Metaverse: Current Status and Future Possibilities," *ACM Comput. Surveys*, vol. 51, no. 3, pp. 1–36, 2018.

[6] S. Gao, T. Liu, Y. Li, and M. Sun, "Avatar Cloning and Real-Time Deepfakes in the Metaverse: A Survey," in *Proc. IEEE TrustCom*, 2022.

[7] A. Cao and Y. Zhang, "Virtual Identity Theft: The Next Frontier in Cybersecurity," *IEEE Internet Computing*, vol. 27, no. 2, pp. 75–83, 2023.

[8] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Two-Stream Neural Networks for Tampered Face Detection," in *Proc. CVPR*, 2017.

[9] S. Dang, D. Wang, and H. Zhang, "Multimodal Deepfakes and Multimodal Detection: A Review," *IEEE Access*, vol. 10, pp. 12356–12378, 2022.

[10] OpenAI, "GPT-4 Technical Report," *arXiv preprint*, arXiv:2303.08774, 2023.

[11] L. Floridi et al., "AI-Generated Content: Ethical Implications and Future Governance," *Nature Machine Intelligence*, vol. 4, pp. 270–272, 2022.

[12] N. Raval et al., "On-device Machine Learning for the Metaverse: Opportunities and Challenges," *IEEE Pervasive Computing*, vol. 21, no. 1, pp. 44–54, 2022.

[13] T. Nguyen, C. Yamagishi, I. Echizen, "Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos," *ICASSP 2019*, pp. 2307–2311.

[14] Y. Li, M.-C. Chang, and S. Lyu, "In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking," *arXiv preprint*, arXiv:1806.02877, 2018.

[15] X. Yang, Y. Li, and S. Lyu, "Exposing Deep Fakes Using Inconsistent Head Poses," in *ICASSP*, 2019, pp. 8261–8265.

[16] S. Wang, X. He, Y. Liu, and X. Li, "2D-FACT: A 2D Forensics-aware Attention-based Convolutional Transformer for Face Forgery Detection," *arXiv preprint*, arXiv:2211.12751, 2022.

[17] J. Kang et al., "Blockchain for Secure and Transparent Deepfake Detection in Social VR," *IEEE Trans. Dependable and Secure Computing*, vol. 20, no. 1, pp. 186–202, 2023.

[18] I. Goodfellow *et al.*, "Generative Adversarial Nets," *Advances in Neural Information Processing Systems*, vol. 27, pp. 2672–2680, 2014.

[19] A. Radford, L. Metz, and S. Chintala, "Unsupervised Representation Learning with Deep Convolutional GANs," in *Proc. ICLR*, 2016, pp. 1–16.

[20] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," in *Proc. CVPR*, 2019, pp. 4401–4410.

Received: July 22, 2025

Volume 38 No. 1s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[21] H. Zhang *et al.*, "Self-Attention Generative Adversarial Networks," in *Proc. ICML*, 2019, pp. 7354–7363.

[22] Y. Jiang, J. Lin, and Z. Wang, "TransGAN: Two Pure Transformers Can Make One Strong GAN, and That Can Scale Up," *arXiv preprint*, arXiv:2101.13188, 2021.

[23] J. Ho *et al.*, "Denoising Diffusion Probabilistic Models," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[24] R. Rombach *et al.*, "High-Resolution Image Synthesis with Latent Diffusion Models," in *Proc. CVPR*, 2022, pp. 10684–10695.

[25] T. Karras, M. Aittala, S. Laine, E. Väisänen, T. Lehtinen, and J. Lehtinen, "Alias-Free Generative Adversarial Networks (StyleGAN3)," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8527–8539, 2021.

[26] P. Dhariwal and A. Nichol, "Diffusion Models Beat GANs on Image Synthesis," *arXiv preprint*, arXiv:2105.05233, 2021.

[27] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17022–17033, 2020.

[28] A. van den Oord *et al.*, "WaveNet: A Generative Model for Raw Audio," *arXiv preprint*, arXiv:1609.03499, 2016.

[29] R. Meier and J. Liu, "Voice Cloning with Lyrebird AI: Fast Personalization of Speech Models," in *Proc. Interspeech*, 2017, pp. 1946–1950.

[30] J. Vincent, "Lyrebird's AI Can Clone Any Voice with Just One Minute of Sample Audio," *The Verge*, Apr. 2017.

[31] Y. Nirkin, Y. Keller, and T. Hassner, "FSGAN: Subject Agnostic Face Swapping and Reenactment," *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[32] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection," *Information Fusion*, vol. 64, pp. 131–148, 2020.

[33] S. Wang, X. He, Y. Liu, and X. Li, "2D-FACT: A 2D Forensics-Aware Attention-Based Convolutional Transformer for Face Forgery Detection," *arXiv preprint*, arXiv:2211.12751, 2022.

[34] Z. Li, Y. Zhao, and Y. Wang, "UnGANable: Defending Against GAN-Based Face Manipulation," in *Proc. USENIX Security Symposium*, 2022, pp. 315–332.

[35] C. Yang, L. Ding, and Y. Chen, "Defending Against GAN-Based Deepfake Attacks via Transformation-Aware Adversarial Faces," *arXiv preprint*, arXiv:2006.07421, 2020.

[36] M. Liu and L. Wang, "GANs for Image and Video Synthesis: A Survey," *IEEE Transactions on Multimedia*, vol. 22, no. 1, pp. 12–30, 2020.

[37] S. Dang, D. Wang, and H. Zhang, "Multimodal Deepfakes and Multimodal Detection: A Review," *IEEE Access*, vol. 10, pp. 12356–12378, 2022.

[38] M. Dolhansky *et al.*, "The Deepfake Detection Challenge (DFDC) Preview Dataset," *arXiv preprint*, arXiv:1910.08854, 2019.

Received: July 22, 2025

Volume 38 No. 1s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[39] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A New Dataset for DeepFake Detection," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[40] H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos," *ICASSP 2019*, pp. 2307–2311.

[41] A. Rossler *et al.*, "FaceForensics++: Learning to Detect Manipulated Facial Images," in *Proc. ICCV*, 2019, pp. 1–11.

[42] C. Chesney and D. Citron, "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security," *California Law Review*, vol. 107, no. 6, pp. 1753–1819, 2019.

[43] K. Pei *et al.*, "Deepfake Video Detection Using Spatiotemporal Convolutional Networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 1, pp. 93–107, 2022.

[44] D. Singh and G. Singh, "An Overview of Deepfake Detection: Challenges, Applications and Opportunities," *Multimedia Tools and Applications*, vol. 82, pp. 2041–2080, 2023.

[45] Y. Zhang *et al.*, "Detection of Deepfake Videos Using Biological Signals," in *Proc. ECCV*, 2020, pp. 370–386.

[46] P. Korshunov and S. Marcel, "Deepfakes: A New Threat to Face Recognition? Assessment and Detection," *arXiv preprint*, arXiv:1812.08685, 2018.

[47] J. Guarnera *et al.*, "A Deepfake Video Detection Technique Based on Amplification Artifacts," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1021–1031, 2021.

[48] H. Mollahosseini, D. Chan, and M. H. Mahoor, "Going Deeper in Facial Expression Recognition Using Deep Neural Networks," *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2016.

[49] S. Agarwal *et al.*, "Detecting Deep-Fake Videos from Phoneme-Viseme Mismatches," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops*, 2020.

[50] C. M. ResearchGate, "Avatar Manipulation in Metaverse: A Framework for Customer Trust," *Field Study*, 2024.

[51] H. Wu, P. Zhou, and P. Hui, "Deepfake in the Metaverse: An Outlook Survey," *arXiv*, 2023.

[52] R. Meier and J. Liu, "Voice Cloning with Lyrebird AI: Fast Personalization of Speech Models," in *Interspeech*, 2017.

[53] T. Tariq, A. Abuadbba, and K. Moore, "Deepfake in the Metaverse: Security Implications for Virtual Gaming, Meetings, and Offices," *arXiv*, 2023.

[54] DXC Technology, "Practice Smart Security in the Metaverse," *DXC Insights*, 2024.

[55] S. Wang *et al.*, "2D-FACT: A 2D Forensics-Aware Attention-Based Convolutional Transformer for Face Forgery Detection," *arXiv*, 2022.

[56] M. Liu and L. Wang, "GANs for Image and Video Synthesis: A Survey," *IEEE Trans. Multimedia*, 2020.

[57] L. Floridi *et al.*, "AI-Generated Content: Ethical Implications and Future Governance," *Nat. Mach. Intell.*, 2022.

Received: July 22, 2025

Volume 38 No. 1s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[58] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1251–1258.

[59] Y. Li, M.-C. Chang, and S. Lyu, "In Ictu Oculi: Exposing DeepFakes via Eye Blinking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2018, pp. 1–10.

[60] R. Durall, M. Keuper, and J. Keuper, "Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks Are Failing to Reproduce Spectral Distributions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 7890–7899.

[61] U. A. Ciftci, I. Demir, and L. Yin, "FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2526–2540, May 2022.

[62] D. Guera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," in *Proc. IEEE Int. Conf. Adv. Video Signal-Based Surveillance (AVSS)*, 2018, pp. 1–6.

[63] A. Dosovitskiy et al., "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021, pp. 1–21.

[64] T. Kinnunen et al., "The ASVspoof 2019 Challenge: TTS and VC Attack Detection for Automatic Speaker Verification," in *Proc. INTERSPEECH*, 2019, pp. 1–5.

[65] Y. Zhou, L. Liu, and J. Xue, "DF-TransFusion: Deepfake Detection via Audio-Visual Temporal Fusion," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2023, pp. 1–5.

[66] Zhang et al., "M2TR: Multi-modal Multi-scale Transformer for Deepfake Detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 5, pp. 2135–2148, May 2023.

[67] H. Gupta, S. Aggarwal, and A. Singh, "A Survey on Deepfake Detection Techniques," *IEEE Access*, vol. 10, pp. 12356–12378, 2022.

[68] P. Korshunov and S. Marcel, "Deepfakes: A New Threat to Face Recognition? Assessment and Detection," *arXiv preprint*, arXiv:1812.08685, 2018.

[69] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 1–11.

[70] H. Wu, P. Zhou, and P. Hui, "Deepfake in the Metaverse: An Outlook Survey," arXiv preprint, arXiv:2303.12494, 2023.

[71] T. Tariq, A. Abuadbba, and K. Moore, "Security Implications of Deepfakes in Virtual Workspaces," IEEE Access, vol. 11, pp. 12288–12303, 2023.

[72] M. Fabris and P. Pessina, "AI-Powered NPCs in Virtual Reality: Ethics and Manipulation," Journal of Digital Ethics, vol. 6, no. 2, pp. 45–58, 2022.

[73] J. Su et al., "Coordinated Avatar Botnets in the Metaverse: Risks and Detection," Proc. NDSS Workshop on Metaverse Security, 2023.

[74] A. Cao and Y. Zhang, "Virtual Identity Theft: The Next Frontier in Cybersecurity," IEEE Internet Computing, vol. 27, no. 2, pp. 75–83, 2023.

[75] D. Gaur, S. Khanna, and M. Chowdhury, "Biometric Spoofing in Extended Reality Systems: A Survey," ACM Computing Surveys, 2024.

Received: July 22, 2025

Volume 38 No. 1s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[76] B. Poole et al., "DreamFusion: Text-to-3D Synthesis Using 2D Diffusion," arXiv preprint, arXiv:2210.02303, 2022.
[77] J. Gratch et al., "Emotional and Psychological Impacts of Avatar Impersonation in VR," Virtual Humans Journal, vol. 9, no. 1, pp. 55–68, 2023.
[78] N. Raval et al., "On-Device Deepfake Detection for the Metaverse: Privacy and Policy Challenges," IEEE Pervasive Computing, vol. 21, no. 1, pp. 44–54, 2023.

Received: July 22, 2025