Proceedings of CONF-MLA 2025 Symposium: Applied Artificial Intelligence Research
DOI: 10.54254/2755-2721/2025.BJ26546

Deepfake Technology's Dual Nature: A Review of Security Risk Assessment and Defense Strategies

Wufeiyang Chen

Faculty of Science and Engineering (FoSE), University of Nottingham Ningbo China, Ningbo, China
cwfeiyang@gmail.com

Abstract. Deepfake technology, empowered by breakthroughs in deep learning-based image synthesis, is profoundly reshaping identity verification systems, finding extensive application in security, finance, and social media with enhanced convenience. However, its capacity to generate hyper-realistic facial forgeries presents a dual impact: while driving innovation, it simultaneously introduces unprecedented security threats, including privacy violations, identity spoofing, and data poisoning attacks. This paper systematically reviews and assesses current research progress on the security risks and defense strategies associated with Deepfake technology. Through synthesis of existing literature, this paper constructs a multidimensional analytical framework examining three core dimensions: first, the core technological principles underpinning Deepfakes and their evolution; second, the diverse spectrum of security risks arising from their misuse and their underlying mechanisms; and third, the effectiveness and inherent limitations of prevailing defense mechanisms, encompassing detection techniques and legal regulations. This study concludes that although Deepfakes advance facial recognition, mitigating their inherent security threats necessitates a multidimensional synergistic approach. This approach must integrate continuous technological advancements, robust legal oversight, and strengthened public awareness initiatives. Future efforts must prioritize establishing cross-disciplinary collaborative governance mechanisms to achieve a dynamic equilibrium between technological innovation and security assurance.

Keywords: Deepfake, Deepfake Detection, Generative Adversarial Networks, Deep Learning, Fake News

1. Introduction

Revolutionary advances in generative adversarial networks (GANs) and deep learning architectures have enabled Deepfake technology to redefine the technological frontiers of synthetic media. Leveraging frameworks such as StyleGAN and diffusion models, this technology now enables nearly indistinguishable manipulation of facial features, voiceprints, and behavioral biometrics. While demonstrating transformative potential in security, finance, and social media authentication domains, its hyper-realistic synthetic outputs concurrently instigate unprecedented sociotechnical crises, evidenced by its inclusion among 2023’s top five global fraud types. Social media users' tendency to amplify trending material accelerates the spread of conspiracies and misinformation, making these platforms ideal vectors for deploying deepfakes [1]. Furthermore, the growing sophistication of AI-driven deepfake technologies has significantly exacerbated associated risks [2].

Confronting this dual-impact landscape, the research systematically deconstructs Deepfake's evolutionary trajectory while examining three critical threat dimensions—biometric privacy erosion through model inversion attacks reconstructing facial data, cross-modal identity forgery exploiting temporal inconsistencies to bypass liveness detection, and training data poisoning via adversarial sample injection manipulating recognition outcomes. To counter these vectors, research evaluates an integrated defensive framework combining adversarial mechanisms enhancing model robustness, based on deep learning forged face detection, multimodal fusion ensuring audiovisual synchronization, and 3D depth-sensing liveness detection establishing sensor-level security barriers. The findings deliver multidimensional decision support across key stakeholder domains: technology developers fortifying authentication systems, policymakers formulating adaptive compliance frameworks, and the public cultivating deepfake discernment capabilities—collectively advancing secure and governable synthetic media ecosystems.

2. The development and technical foundation of Deepfake

This section systematically discusses the evolutionary trajectory and fundamental technological underpinnings of Deepfake technology, with particular emphasis on the pivotal role of Artificial Intelligence (AI) and deep learning methodologies in driving its rapid advancement.

2.1. Development of Deepfake

The term "Deepfake" first appeared on Reddit in 2017, coined by a user who shared manipulated videos using deep learning techniques. Initially developed and explored within academic and technical communities, Deepfake technology rapidly spread to the public domain with the introduction of user-friendly tools such as FakeApp and FaceSwap. These platforms significantly lowered the barrier to entry, allowing individuals without programming skills to generate manipulated media. Over time, the technology evolved from basic facial swaps to highly realistic videos that could mimic speech, facial expressions, and gestures. As Deepfakes became more sophisticated, their use expanded beyond entertainment into areas such as political misinformation, financial fraud, and identity theft. The proliferation of open-source code and large online datasets further accelerated this development. Today, Deepfakes represent both a technical achievement and a societal challenge, demanding urgent attention from researchers, policymakers, and the public.

2.2. Technical foundation

Deepfake technology is built upon the foundations of artificial intelligence (AI), particularly deep learning (DL) and machine learning (ML). At its core, Deepfake creation relies on neural networks, especially a class of DL models known as Generative Adversarial Networks (GANs). GANs consist of two competing networks—a generator that creates synthetic media and a discriminator that evaluates its authenticity [3]. Through iterative training on large datasets containing images and videos of target individuals, these models learn to produce highly realistic facial expressions, movements, and speech patterns. The success of GANs in synthesizing audio, video, and images makes them the cornerstone of Deepfake generation.

Furthermore, the availability of graphical user interface (GUI)-based tools such as FakeApp has democratized Deepfake creation, enabling users with minimal technical expertise to produce manipulated content given sufficient time and computational resources [4]. Moreover, the technology requires substantial visual data to train the models, enabling the replication of photorealistic human features. Convolutional neural networks (CNNs) are often used in video and image synthesis due to their ability to extract and process spatial features. Also, originally prominent in natural language processing (NLP), transformer models have recently shown significant progress in the generation of deepfakes. Built on an encoder–decoder backbone with self-attention, these models can generate synthetic images or videos through fine-tuning of pre-trained weights on targeted datasets [5].

3. Deepfake-related security risks

The rapid evolution of Deepfake technology has raised critical security concerns, particularly in the domains of privacy protection, identity authentication, and model integrity. This section outlines three major security risks associated with Deepfake systems.

3.1. Privacy invasion and data leakage

One of the most urgent issues is the massive amount of facial data required to train Deepfake models. These data are typically collected from social media, public datasets, or surveillance systems, and sometimes even obtained through hacking attempts to access private photo albums. Individuals are often unaware that their facial data is being harvested, whether through legitimate or illegitimate means, and this data contains sensitive biometric information that is inherently private and unique. Even more, the biometric traits of facial data are distinctive, permanent, and irreplaceable, which makes it especially sensitive. Unlike passwords or credit card numbers, once facial data is compromised, it cannot be altered, leading to irreversible damage to an individual's privacy.

3.2. Identity forgery and face swapping attacks

Deepfake technology also significantly elevates the risk of identity forgery. By extracting facial features from a target and manipulating them onto another individual’s body, malicious actors can easily create fake identities. This is particularly concerning in contexts where facial recognition is used for authentication, such as in financial transactions, security systems, or access control. For instance, attackers can use Deepfake videos to impersonate CEOs or public figures, thereby gaining unauthorized access to sensitive information or causing reputational damage. In 2023, several unscrupulous advertisers exploited celebrity deepfakes for commercial purposes: a counterfeit Tom Hanks was used to promote dental plans without his permission [6].

Beyond personal threats, Deepfakes also pose political risks. They have been used to spread disinformation, manipulate public opinion, and undermine trust in democratic processes. The weaponization of Deepfakes in political campaigns, where fabricated videos can discredit candidates or sway voters, exemplifies the far-reaching implications of this technology. For example, Governor Ron DeSantis aired a damaging deepfake audio of Trump [7]; and a Bellingcat researcher fabricated images of Trump being arrested using deepfake technology [8]. The Republican Party published a campaign video featuring a deepfake of President Biden rejoicing over his 2024 victory, accompanied by fictional outcomes such as the invasion of Taiwan, bank closures, and San

Francisco being overrun by crime [9]. In summary, the widespread misuse of Deepfake technology not only threatens personal security and privacy but also poses significant political and societal risks.

3.3. Data poisoning and vulnerabilities in model training

In addition to the risks posed by the deployment of Deepfake technology, there are also vulnerabilities in the model training phase, especially through data poisoning attacks. In such cases, malicious actors inject misleading or incorrect data into the training dataset, altering the model's learning process. For instance, by uploading a set of images with specific features—such as individuals wearing a particular type of glasses—the trained model may mistakenly recognize other people wearing similar glasses as the target person. In this way, attackers can manipulate the trained model to misidentify or fail to recognize certain targets under specific conditions. Furthermore, when the Deepfake generation model is affected by poisoned data, it may produce high-quality images or videos that are highly realistic but do not align with real-world patterns (e.g., facial movements, blinking, lighting, etc.). Under certain attack conditions, the line between "fake" and "real" becomes blurred, and subtle differences in Deepfakes become harder to detect. This not only impacts the accuracy of Deepfake detection systems but also allows malicious actors to bypass biometric security systems. Therefore, data poisoning attacks make Deepfake generation models more susceptible to manipulation, with certain facial features (such as glasses or masks) being used as tools to deceive the model.

4. Overview of defensive strategies against Deepfake security risks

4.1. Technical countermeasures

4.1.1. Adversarial defense mechanisms

Adversarial defense mechanisms aim to enhance the robustness of deepfake detection models by introducing perturbed samples during the training process. In facial recognition tasks, researchers often use images altered by subtle noise or adversarial patterns—such slight facial distortions as denoising and compressing the images—as part of the training dataset. These manipulated inputs help the model become accustomed to such attacks, thereby improving its ability to withstand adversarial manipulations. This approach strengthens the model's classification boundaries and improves its resilience against malicious input. By training with adversarial data, the system becomes better equipped to distinguish between authentic and fake content, ultimately reducing the risk of deepfake manipulation.

4.1.2. Deep ensemble learning for fake face detection

In Deepfake detection, deep learning techniques play a pivotal role, with significant research focused on two main approaches: Conventional Neural Networks (CNN) and Region-based Convolutional Neural Networks (RCNN) [10].

CNN is a deep learning model designed for image data processing. It automatically extracts features from images through convolutional operations. By stacking convolution and pooling layers, CNN progressively captures simple to complex features and uses fully connected layers for final classification or regression. In Deepfake detection, CNN is primarily used to extract facial features from individual video frames to identify forged content [11]. However, CNN is mostly applied to static images or single frames. RCNN, an extension of CNN, is primarily used for object detection tasks. It first generates multiple candidate regions and applies convolutional operations to each, detecting multiple objects in the image. Unlike CNN, RCNN not only classifies objects but also considers their location (bounding boxes) [11]. In Deepfake detection, RCNN analyzes multiple video frames, leveraging both spatial and temporal features to detect forged content in dynamic videos, offering a more comprehensive and accurate detection result.

4.1.3. Multi-modal fusion verification

As deepfake technology advances, traditional single-modal detection methods struggle with subtle distortions in dynamic videos. As a result, multi-modal detection, combining audio and visual signals, has emerged to improve accuracy by analyzing both modalities for anomalies.

Audio-Visual Deepfake Detection (AV-DFD) [12] exemplifies a multimodal framework: it aligns audio and visual features and feeds them into a cross-attention module for joint temporal scrutiny. This method not only considers facial movements in video frames but also integrates audio information, such as the synchronization between speech and facial movements, effectively enhancing the detection of forged videos. In audio-visual consistency detection, these approaches pinpoint deepfake artifacts—most notably the mismatch between mouth shapes (visemes) and corresponding sounds (phonemes) [13]. Recent Vision-Transformer-based encoders/decoders, exemplified by AVFakeNet [14], have further sharpened multimodal fusion performance. These approaches fuse audio and visual features in the embedding space, improving deepfake detection performance when dealing with dynamic videos. By integrating audio-visual information, they capture more detailed forged features, making them more robust in identifying deepfake content.

Overall, multi-modal fusion methods, by synchronizing cross-perceptual channels, strengthen detection systems, making them crucial for advancing deepfake detection technologies.

4.2. Legal regulation and policy frameworks

Existing legal frameworks are crucial for privacy protection and deepfake regulation, yet many remain fragmented, vague, or outdated, limiting their effectiveness against evolving synthetic media threats. Employing natural-language processing and linguistic analysis, an empirical study examined 96 bills introduced during 2024–2025 and identified 29 that explicitly targeted child protection in synthetic-media contexts. The findings reveal that while legislative attention is increasing, it remains fragmented across states [15]. One of the main challenges lies in the rapid evolution of deepfake technology, which often outpaces regulatory responses. In many cases, legal provisions remain vague, overly general, or fail to address the full spectrum of risks posed by deepfakes. This leaves significant loopholes for malicious actors to exploit. These shortcomings highlight the need for a unified legal framework to better protect minors and prevent deepfake misuse.

4.3. Public awareness and education

Raising public awareness and educating users about Deepfake technology is crucial in combating its negative impacts. Many individuals are unaware of the potential risks associated with Deepfakes, making them more susceptible to manipulation. Through targeted awareness campaigns and educational programs, individuals can become more discerning when interacting with digital content. These initiatives should include information on how to spot Deepfakes, the dangers they pose, and the importance of digital literacy in recognizing synthetic media. Furthermore, educating content creators, influencers, and the public about ethical use and responsible content sharing is critical. By fostering a more informed society, we can mitigate the impact of Deepfakes on personal privacy, security, and societal trust.

5. Conclusion

Deepfake technology stands at the intersection of technological advancement and security vulnerability, embodying the paradoxical nature of innovation in the era of synthetic media. On one hand, it pushes the boundaries of digital creativity and human–machine interaction; on the other, it exposes critical flaws in our ability to ensure authenticity and trust in visual information. This paper has systematically examined the technological evolution of Deepfakes and the foundational AI architectures that enable their development, as well as the multifaceted security risks they introduce —including privacy infringement and biology data leakage risks, cross-modal identity and face forgery, and data poisoning during model training. In response, technical defense strategies such as adversarial training to enhance model robustness, deep ensemble learning for fake face detection, and multimodal audio-visual fusion have demonstrated significant, albeit still incomplete, potential in mitigating synthetic media threats.

Beyond technological measures, the analysis revealed substantial gaps in existing legal frameworks, particularly regarding the protection of minors. These frameworks remain fragmented and frequently lag the rapid pace of Deepfake innovation. Public awareness, though a critical line of defense, remains insufficient in both scope and effectiveness. Collectively, these findings underscore the need for a multidimensional, adaptive, and ethically grounded approach that integrates ongoing technological advancement, agile policymaking, and proactive public education.

Looking ahead, addressing the challenges posed by Deepfake technology will require sustained interdisciplinary collaboration across computer science, legal studies, ethics, and media research. Only through such integrative efforts can we establish a robust and flexible governance ecosystem—one that preserves public trust, safeguards individual privacy, and ensures the responsible development of synthetic media in an increasingly digitized world.

References

[1] Masood M., Nawaz M., Malik K.M., et al. (2023) Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. Appl Intell 53(4): 3974–4026

[2] Liu M.Y., Huang X., Yu J., et al. (2021b) Generative adversarial networks for image and video synthesis: algorithms and applications. Proc IEEE, 109(5): 839–862.

[3] Khan, R., Sohail, M., Usman, I., Sandhu, M., Raza, M., Yaqub, M. A., & Liotta, A. (2024). Comparative study of deep learning techniques for DeepFake video detection. ICT Express, 10(6), 1226-1239.

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

[5] Mubarak R., Alsboui T., Alshaikh O., et al. (2023) A survey on the detection and impacts of deepfakes in visual, audio, and textual formats. IEEE Access 11: 144497–14452. https: //doi.org/10.1109/ ACCESS.2023.3344653

[6] Guardian. (2023) Tom Hanks says AI version of him used in dental plan ad. https: //www.theguardian.com/film/2023/oct/02/tom-hanks-dental-ad-ai-version-fake.

[7] Isenstadt A. (2023) DeSantis PAC uses AI-generated Trump. Politico. https: //www.politico.com/news/2023/07/17/desantis-pac-ai-generated-trump-in-ad-00106695.

[8] Stanley-Becker I., and Nix N. (2023) Fake images of Trump arrest...Washington Post (March 22). https: //www.washingtonpost.com/politics/2023/03/22/trump-arrest-deepfakes/.

[9] Binder M. (2023) Republicans release AI-generated attack ad. Mashable, April 25 https: //mashable.com/video/republican-attack-ad-biden-reelection-ai.

[10] Zhou, X., Wang, Y., & Wu, P. (2020). Detecting deepfake videos via frame serialization learning. Paper presented at the 2020 IEEE 3rd International Conference of Safe Production and Informatization (IICSPI).

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

[12] Zhou, Y., & Lim, S. N. (2021). Joint audio-visual deepfake detection. In Proceedings of the IEEE/CVF international conference on computer vision, pp. 14800-14809.

[13] Oorloff, T., Koppisetti, S., Bonettini, N., Solanki, D., Colman, B., Yacoob, Y., ... & Bharaj, G. (2024). Avff: Audio-visual feature fusion for video deepfake detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 27102-27112.

[14] Ilyas, H., Javed, A., & Malik, K. M. (2023). AVFakeNet: A unified end-to-end Dense Swin Transformer deep learning model for audio–visual deepfakes detection. Applied Soft Computing, 136, 110124.

[15] Roundtree, A. K. (2025). Deepfake Laws Protecting Childhood Safety and Future: An NLP Analysis. In Proceedings of the 24th Interaction Design and Children, pp. 1086-1090.