Deepfake Attacks on Biometric Systems: Threats, Detection, and Defense Mechanisms – A Systematic Survey

Weibo Ye
Faculty of Information Technology, Monash University, Melbourne, Victoria, VIC 3168, Australia

Abstract: The rapid advancement of deepfake technology poses unprecedented security threats to biometric systems. This paper presents a systematic survey on the latest progress of deepfake attacks in biometric authentication. First, we construct a Deepfake Kill Chain framework that systematically describes the complete attack chain from content generation to authentication decision. Second, we classify and compare defense methods across four dimensions: content-level, behavior-level, environment-level, and generative-end interventions, analyzing their applicability, failure modes, and trade-offs between security and usability in different scenarios. Third, we deeply discuss how Shortcut Learning leads to reduced generalization capability of detectors on unknown variants. Finally, we propose a hierarchical and collaborative defense framework and provide concrete deployment recommendations. This survey aims to provide a unified cognitive framework for both academia and industry.

Keywords: Deepfake; Biometric Recognition; Detection Methods; Defense Strategies; Model Trustworthiness

1. Introduction

1.1 Research Background

Biometric identification technologies have been widely deployed in critical domains such as financial payments, e-government services, and telemedicine. These technologies have become the primary means of modern identity authentication due to their efficiency and convenience. However, with breakthroughs in deep learning techniques such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAE), and diffusion models, deepfake technology has evolved from theoretical research to a practical threat.

In recent years, real security incidents caused by deepfakes have occurred frequently. In the financial sector, attackers have successfully impersonated users using forged faces or voices to conduct remote payments or account transfers, causing substantial economic losses. In government services, forged identity documents have led to unauthorized access to public services, affecting numerous users. On social media platforms, forged videos have been widely spread, triggering serious information security incidents. These events clearly demonstrate that traditional single-modal biometric systems struggle to defend against highly realistic deepfakes.

Although the academic community has produced substantial deepfake detection research, these studies exhibit significant fragmentation. Most research focuses on single modalities (e.g., only facial or only speech recognition), with insufficient cross-modal collaborative defense[1]; existing evaluation frameworks fail to uniformly measure verification performance and spoofing vulnerability[2]; detection models may rely on non-essential features for decision-making, resulting in poor generalization to unknown variants; generative-end prevention and detection systems lack organic integration, forming a fragmented defense ecosystem.

1.2 Core Research Questions

This survey addresses the following core research questions:

What is the complete attack chain for deepfake attacks on biometric systems, and what are the key risk points at each stage?

What categories can existing defense methods be divided into, and what are their respective advantages, disadvantages, and application boundaries?

How can we establish a scientific evaluation framework to uniformly assess defense method effectiveness, especially in balancing security and usability?

How do model trustworthiness issues (particularly Shortcut Learning) affect real-

world deployment of defense systems?
How can we construct a hierarchical and collaborative defense system to counter increasingly complex and diverse deepfake attacks?

2. Threat Modeling of Deepfake Attacks

2.1 Deepfake Kill Chain Framework

To systematically understand the threats posed by deepfakes to biometric systems, we decompose the attack process into five critical stages, forming the Deepfake Kill Chain framework. This framework describes how attacks evolve from the initial stage to final authentication deception from the attacker's perspective.
Forged Content Generation: Attackers use generative models (GANs, diffusion models, etc.) to generate fake biometric features (faces, voice, etc.) from authentic samples or synthetic data. The generation quality at this stage directly affects subsequent attack success rates. High-quality forged content has higher pass rates but requires longer generation times; low-quality forgeries are easier to generate but more easily detected.
Process Injection: Generated forged content must be injected into the identity authentication system workflow. Multiple injection methods are possible, including direct presentation (presenting screen or audio playback of forged content directly to camera or microphone), transmission replacement (replacing authentic data during network transmission), or hybrid attacks (mixing forged content with authentic content).
Feature Impersonation: Deepfakes attempt to mimic authentic user biometric features. For facial recognition systems, this may include impersonating facial texture, lighting conditions, head pose, skin texture, etc.; for voice recognition systems, this may include impersonating voice tone, accent, rhythm, speech rate, etc. The objective at this stage is to minimize differences from authentic user features.
Detection Evasion: Modern deepfakes may employ adversarial techniques to evade known detection methods, including local perturbations (making minor modifications to forged content to evade specific detectors), adaptive attacks (optimizing against known target system characteristics), or cross-modal confusion (exploiting inconsistencies between different modalities to create ambiguity).
Authentication Decision Deception: Forged content passes through feature extraction and similarity matching authentication procedures, ultimately bypassing the decision threshold to cause authentication error. This is the final attack objective stage; successfully reaching this stage means the attack has completely circumvented the defense system.

2.2 Necessary Conditions for Attack Success

Deepfake attack success is not isolated but depends on an organic combination of multiple factors. Generation quality determines the realism of forged content, directly affecting whether it can pass human inspection; physical constraints involve the attack presentation method, directly affecting whether actual system contact is possible; detection blind spots are weak points of defense systems, and all defense methods may have uncovered scenarios; cross-modal coordination means coordinating multiple modality impersonation in multi-modal authentication systems, with difficulty increasing exponentially with modality count; adaptive capability reflects the attacker's understanding of target system characteristics, with adaptive attacks typically being more effective than generic attacks[3].

3. Classification Survey of Defense Methods

3.1 Content-Level Defense

3.1.1 Image domain feature analysis
Traditional detection methods primarily focus on forensic traces left by deepfakes in images or videos. Gragnaniello et al. systematically compared local descriptors such as LBP (Local Binary Pattern) and HOG (Histogram of Oriented Gradients), finding that these features are somewhat effective for detecting static presentation attacks (Presentation Attacks) but have limited support for dynamic deepfakes[4]. This is because generative models better preserve the coherence of high-frequency information.
Recent research indicates that deepfakes typically exhibit specific forensic traces in the frequency domain. Different generative models leave different fingerprints in the frequency space, causing abnormalities in frequency distribution and spectral gaps. Therefore, hybrid methods combining spatial and frequency

domains demonstrate superior detection performance. However, these methods typically depend on specific forgery models, with limited adaptability to new generation methods.

3.1.2 Temporal domain feature analysis

For video deepfakes, temporal consistency is a critical clue. Due to computational constraints of generative models, forged videos typically exhibit incoherent features between frames, such as facial feature flickering, unnatural head motion, or unnatural facial expression transitions. Through optical flow analysis, we can capture inter-frame motion differences; 3D Convolutional Neural Networks (3D-CNN) can simultaneously model spatial and temporal information; Recurrent Neural Networks (RNN) can learn temporal sequence dependencies. These methods are typically more effective than single-frame analysis for detecting temporal inconsistencies.

3.1.3 Deep learning methods: evolution and limitations

From pre-trained Convolutional Neural Networks (CNN) to attention mechanisms, and then to Transformers, the expressiveness of deepfake detection models continues to increase. However, a critical issue is weak cross-domain generalization. Detection models trained on one dataset typically suffer significant performance degradation on other datasets, a phenomenon known as the "cross-dataset generalization problem"[4]. The primary reasons include different data acquisition conditions across datasets (camera models, lighting conditions, backgrounds, etc.) and the diversity of deepfake generation methods causing varied forensic traces. Consequently, many recent studies have shifted toward using data augmentation and transfer learning techniques to improve generalization performance.

3.2 Behavior-Level Defense

3.2.1 Liveness detection

Liveness detection[5] distinguishes authentic users from forged content by requiring users to perform specific genuine behaviors. Primary methods include reflection analysis (authentic facial surface reflection patterns differ fundamentally from 2D forged content), eye movement detection (requiring users to track moving objects), blink detection (analyzing eyelid motion frequency and patterns), head motion analysis (requiring users to execute specific head poses), etc.[5]

Yang et al. introduced temporal sequence behavior consistency through dynamic lip movement analysis in speaker authentication, enhancing recognition capability against synchronized audio-visual deepfakes[6]. This work's key insight is that static appearance features alone can be easily bypassed by highly realistic deepfakes, while dynamic behavioral features provide an additional defense layer. Voice and lip movement coordination involves physiological constraints that deepfakes struggle to perfectly impersonate without detection.

3.2.2 Cross-modal behavioral coordination

Authentic user biometric features have intrinsic coordination relationships with each other. During speech, voice, lip movement, facial expression, and head pose should be highly coordinated. When deepfakes attempt to impersonate individual modalities, they struggle to simultaneously impersonate these coordination relationships across multiple dimensions. Therefore, detecting cross-modal inconsistencies becomes an effective defense strategy.

Specifically, this includes detecting audio-visual synchronization (whether voice-to-lip-movement synchronization conforms to physical laws), facial expression-speech consistency (facial expression changes during speech should relate to speech content), and head pose-eye motion coordination (natural head motion should accompany corresponding eye movements). These multi-dimensional consistency checks substantially increase attack difficulty.

3.3 Environment-Level Defense

3.3.1 Electrical network frequency signal verification

DeFakePro[7] proposed utilizing Electrical Network Frequency (ENF) as an innovative trust signal[7]. ENF is the power system's frequency signal, unconsciously recorded through various physical pathways (such as device electromagnetic interference) during video or audio recording. Because ENF is determined globally by regional power systems, attackers are nearly unable to precisely replicate specific local ENF characteristics without knowing them when generating forged content.

This method's advantages include difficulty of forgery, strong cross-domain robustness (not dependent on specific biometric feature characteristics), and complementarity with other methods (can be combined with content-level

and behavior-level methods). However, ENF method limitations include requiring high-quality audio recording environments, potentially unsuitable for certain scenarios.

3.3.2 Other physical constraints

Research also explores other environmental information for defense. Lighting condition analysis: authentic environment natural lighting differs significantly from studio lighting in light direction, color temperature, shadow patterns, etc. Background consistency analysis: background replacement or editing during forgery typically leaves obvious traces such as unnatural boundaries and color shifts. Noise characteristic analysis: authentic environmental noise (background sound, device noise) differs markedly from artificial noise generated by deep learning models in time-frequency characteristics. These physical-level constraints provide forensic evidence difficult to forge.

3.4 Generative-End Intervention

3.4.1 Adversarial perturbation injection

Dong et al.'s work considers the problem from an attack-defense adversarial perspective[8]. Rather than passively detecting forgeries, we can inject adversarial perturbations during the forgery generation process to weaken quality from the source. The specific method involves injecting minor perturbations during generative model optimization, causing the generated forged content to remain perceptually realistic to human observers while degraded for biometric systems, unable to pass the system similarity threshold. This "attacker's attacker" perspective provides a new defense paradigm.

3.4.2 Generation quality control

Another direction involves identifying key parameters in the generation process to reduce forgery success probability at the source. This includes identifying generation parameter configurations prone to forgery failure, setting minimum quality thresholds to reject sub-threshold forgeries, constraining generation models themselves to limit forgery diversity, etc. While these measures cannot completely prevent forgery generation, they substantially increase attack costs.

4. Evaluation Framework and Model Trustworthiness

4.1 EPS Framework

Chingovska et al.'s EPS (Expected Performance and Spoofability) framework[2] represents an important attempt at unified defense method evaluation. The framework's core insight is that defense systems must simultaneously consider verification performance (ability to accept authentic users) and spoofing vulnerability (ability to reject attacks). The framework defines key indicators including APCER (Attack Presentation Classification Error Rate - probability of incorrectly accepting attacks as authentic users, lower is better), BPCER (Bona Fide Presentation Classification Error Rate - probability of rejecting authentic users, lower is better), and ACER (average of both), etc. Its value lies in emphasizing the trade-off between security and usability: a system focusing only on security while ignoring user experience (rejecting many authentic users) is as unacceptable as a system prioritizing usability while vulnerable to attacks (accepting forged content).

4.2 Shortcut Learning Issue

4.2.1 Definition and typical manifestations

Shortcut Learning refers to models achieving superficially high performance by learning non-essential, spurious correlations. In deepfake detection, this means models may learn to depend on dataset-specific forensic traces rather than essential deepfake characteristics. According to Sahidullah[9] et al.'s research, Shortcut Learning in deepfake detection manifests in four primary forms[9]:

Generative Model-Specific Features: Detectors learn to identify specific generative model (e.g., particular GAN architecture) distinctive traces, causing substantial performance degradation when new generative models appear.

Dataset-Specific Features: Models over-adapt to training dataset-specific characteristics (color bias, resolution, compression artifacts), with severe performance degradation when applied to deepfakes from other sources.

Non-Semantic Feature Dependence: Detectors depend on pixel-level forensic traces rather than semantic information; adversarial modification of these low-level features may cause detection failure.

Modality Imbalance: In multi-modal detection, models over-depend on dominant modalities; when the dominant modality is specifically optimized by attackers, the entire system fails.

4.2.2 Impact and mitigation strategies

This problem's severity lies in the fact that

laboratory high performance may not translate to effective real-world deployment. A detector achieving 98% accuracy on standard datasets might only achieve 60% accuracy when encountering new deepfake generation methods. This creates a critical trust crisis: we cannot ensure real deployment defense system effectiveness.

Existing methods for mitigating Shortcut Learning include data augmentation (expanding training data diversity), transfer learning and domain adaptation (leveraging pre-trained models and domain adaptation techniques to improve generalization), adversarial training (using adversarial examples to enhance model robustness), multi-task learning (simultaneously learning related tasks to force learning more universal features), and explainability analysis (using feature visualization and gradient analysis to identify whether models depend on non-essential features).

5. Multi-Modal Fusion Detection

5.1 Cross-Modal Alignment

Du et al.'s cross-modal alignment and distillation framework's core idea is that authentic audio-visual content should be highly aligned in semantic space while deepfakes often exhibit misalignment[10]. By separately extracting visual and audio features, projecting them into the same semantic space, computing feature consistency measures, and marking inconsistent feature pairs as deepfake evidence. This method's advantage lies in capturing subtle cross-modal inconsistencies, which are critical for deepfake recognition.

5.2 Lightweight Multi-Modal Models

Considering computational resource constraints and real-time requirements in practical deployment, lightweight joint audio-visual models embed audio-visual coordination into single-stream architectures. Compared to dual-stream architectures, this approach achieves significant computational reduction, stronger real-time performance, and more effective multi-modal interaction. These models can also be deployed on mobile and embedded devices.

5.3 Hierarchical Fusion and Inconsistency Detection

Recent research explores hierarchical multi-modal fusion. These methods perform feature fusion at different network depths, considering both local pixel/spectral features and global semantic information, explicitly modeling contradictions and conflicts between modalities. These contradictions are often key evidence of deepfakes. Multi-modal inconsistency detection includes synchronization detection (audio-visual synchronization degree), semantic consistency (emotional consistency, content consistency), and feature conflict recognition.

6. Application Scenario Analysis

6.1 Financial Authentication

In bank transfers and payment authentication scenarios[11], deepfake threats are most direct[11]. These scenarios feature high economic value, multiple biometric channels, and rapid authentication requirements. Existing defenses face multiple challenges: attackers can combine forgeries across multiple modalities; user expectations for rapid authentication limit defense method complexity; attackers may leverage historical data to generate personalized deepfakes for specific users. Defense strategies can employ risk-adaptive tiered verification: low-risk transactions use single features, high-risk transactions employ multi-factor authentication.

6.2 E-Government Services

In identity documents and driver's license verification scenarios[12], identity forgery may cause crimes with lasting legal consequences[12]. Defense requires establishing historical baseline and anomaly detection systems, periodically updating user biometric baselines, detecting significant deviations from historical baselines, and initiating human review when suspicious.

6.3 Telemedicine

Telemedicine authentication must ensure patient privacy, medical safety, and complete audit logs. Multi-dimensional verification combinations can be employed: biometric recognition plus knowledge authentication, real-time monitoring plus post-event audit, and anomalous behavior detection.

7. Hierarchical Collaborative Defense Framework

Based on the preceding analysis, we propose a hierarchical collaborative attack-defense framework integrating generative-end prevention,

content detection, behavioral authentication, environment verification, risk fusion decision-making, and feedback learning into six levels of organic integration.

7.1 Generative-End Prevention Layer
Reducing forged material quality and usability at the generation stage

7.2 Content Detection Layer
Detecting forgeries by analyzing forensic traces

7.3 Behavioral Authentication Layer
Authenticating through detecting genuine user behavior and biological constraints

7.4 Environment Verification Layer
Utilizing difficult-to-forge environmental and physical information

7.5 Risk Fusion and Decision Layer
Integrating all defense layer information for final authentication decisions

7.6 Feedback and Learning Layer
Learning from false alarms and attacks for continuous defense improvement

8. Domestic Research Status and Innovation

8.1 Major Achievements
Domestic academia has made important contributions to deepfake defense. Li Chunxian[13] et al. examined deepfake threats from social-institutional perspectives, proposing governance strategies based on socio-technical coupling. Ran Lian and Zhang Wei et al. constructed AIGC deepfake analysis frameworks using actor-network theory, proposing targeted governance strategies including virtue governance, law governance, technology governance, and public governance.

In detection methods, Yao Wenda[14] et al. systematically organized facial deepfake detection feature space division and evaluation metrics. Li Junjie et al. focused on video deepfake detection generalization issues[15]. Zeng Zhiping et al. reviewed deepfake audio generation and authentication from audio perspectives. In attack-defense techniques[16], Wang Li et al. constructed facial deepfake liveness verification attack frameworks and unified evaluation systems[17], revealing vulnerabilities in commercial liveness verification APIs. Wu Hanyu et al. focused on anti-forensic adversarial attacks against deepfake video detection[18]. Mo Yonghua et al. proposed face anti-spoofing verification systems based on big data analysis[19].

8.2 Main Innovation Points of This Survey
First systematically describing deepfake threats from the complete attack chain perspective; (2) establishing unified classification framework across content, behavior, environment, and generative-end dimensions enabling method comparability; (3) systematically expounding model trustworthiness issues like Shortcut Learning with mitigation strategies; (4) proposing complete defense system from generative-end prevention to risk fusion; (5) analyzing defense strategies combining financial, government, and healthcare scenarios.

9. Open Issues and Future Directions

9.1 Currently Unresolved Issues
Real-time Detection vs. Usability Trade-off: Current multi-modal fusion methods have high computational complexity with insufficient real-time performance on mobile devices; Generalization to Unknown Attacks: Continuously emerging generative models and attack methods make existing detector effectiveness uncertain on unseen attacks; Cross-Organization Collaboration Feasibility: Threat intelligence sharing faces privacy, trade secret, and legal barriers; Legal and Ethical Framework: Biometric feature collection and use face privacy-security balance issues.

9.2 Future Research Directions
Generative model explainability, active defense strategies, physics-constraint-based defense, industry standards establishment, and zero-trust verification paradigms will be important future research directions.

10. Conclusion
Deepfake technology development poses unprecedented security challenges to biometric systems. This survey systematized attack processes through the Deepfake Kill Chain framework, classified and compared defense methods across multiple dimensions, deeply discussed model trustworthiness issues like Shortcut Learning, and proposed a hierarchical collaborative defense architecture. Deepfake

defense is not merely a technical issue but a comprehensive problem involving legal, ethical, and social dimensions. Future research requires multi-disciplinary and multi-stakeholder collaborative frameworks.

References

[1] Li Junjie, Wang Jianzong, Zhang Xulong, Qian Xiaoyang. Generalization problem in video deepfake detection: methods, challenges and technical progress. Big Data Research, 2025.

[2] Chingovska I, dos Anjos A R, Marcel S. Biometrics Evaluation Under Spoofing Attacks. IEEE Transactions on Information Forensics and Security, 2014, 9(12): 2264-2277.

[3] Gragnaniello D, Poggi G, Sansone C, et al. An Investigation of Local Descriptors for Biometric Spoofing Detection. IEEE Transactions on Information Forensics and Security, 2015, 10(4): 849-863.

[4] Li Junjie, Wang Jianzong, Zhang Xulong, Qian Xiaoyang. Generalization problem in video deepfake detection: methods, challenges and technical progress. Big Data Research, 2025.

[5] Wang H, Su L, Zeng H, et al. Anti-spoofing study on palm biometric features. Expert Systems with Applications, 2023, 218: 119546.

[6] Yang C Z, Ma J, Wang S, et al. Preventing DeepFake Attacks on Speaker Authentication by Dynamic Lip Movement Analysis. IEEE Transactions on Information Forensics and Security, 2021, 16.

[7] Nagothu D, Xu R, Chen Y, et al. DeFakePro: Decentralized Deepfake Attacks Detection Using ENF Authentication. arXiv preprint arXiv:2207.13070, 2022.

[8] Dong J, Wang Y, Lai J, et al. Restricted Black-Box Adversarial Attack Against DeepFake Face Swapping. IEEE Transactions on Information Forensics and Security, 2023, 18: 2596-2608.

[9] Sahidullah S, et al. Shortcut Learning in Binary Classifier Black Boxes: Applications to Voice Anti-Spoofing and Biometrics. arXiv preprint arXiv:2306.00044, 2023.

[10] Du Y, Wang J, Zhang X, et al. Cross-modal Fusion with Distillation for DeepFake Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

[11] Yao W D, Li P C, Zhao Y, et al. Review of research on face deepfake detection methods. Journal of Image and Graphics, 2025, 30(7): 2343-2363.

[12] Li C X. Will deepfake bring a new "epistemic crisis"? Studies in Dialectics of Nature, 2024, 40(8).

[13] Wang L. Research on Attack and Defense Techniques for Facial Deepfake. Ph.D. dissertation, Shandong University, 2024.

[14] Yao W D, Li P C, Zhao Y, et al. Review of research on face deepfake detection methods. Journal of Image and Graphics, 2025, 30(7): 2343-2363.

[15] Li J, Wang J, Zhang X, et al. Generalization problem in video deepfake detection: methods, challenges and technical progress. Big Data Research, 2025.

[16] Zhang Z, Zhang X, Qian X, et al. Survey on deepfake audio generation and authentication technology. Big Data, 2025.

[17] Wang L. Research on Attack and Defense Techniques for Facial Deepfake. Ph.D. dissertation, Shandong University, 2024.

[18] Wu H. Research on Anti-Forensic Adversarial Attack Methods against Deepfake Video Detection. M.S. thesis, UESTC, 2024.

[19] Mo Y, Zhang Z, Chen Y. Design and Implementation of Face Anti-spoofing Verification System Based on Big Data Analysis. Modern Computer, 2024, 30(16).

