*International Journal of Computer Applications (0975 - 8887)*
*Volume 187 - No.82, February 2026*

# A Survey on Deepfake Video Detection using Hybrid Multimodal Features

Adithya Anil
Department of Computer Science and Engineering
Federal Institute of Science and Technology (FISAT), India

Afeefa M.S.
Department of Computer Science and Engineering
Federal Institute of Science and Technology (FISAT), India

Akshara Raghu
Department of Computer Science and Engineering
Federal Institute of Science and Technology (FISAT), India

Anamika Sudheer
Department of Computer Science and Engineering
Federal Institute of Science and Technology (FISAT), India

Shimy Joseph
Assistant Professor
Department of Computer Science and Engineering
Federal Institute of Science and Technology (FISAT), India

## ABSTRACT

The rapid evolution of deep generative models has enabled the creation of highly realistic deepfake videos, posing significant threats to digital media authenticity, privacy and public trust. Modern deepfake generation techniques based on Generative Adversarial Networks (GANs) [1], autoencoders [2] and neural rendering models [3] can synthesize facial expressions, lip movements and identities with remarkable realism, making manual detection increasingly unreliable. This paper presents a comprehensive survey of deepfake video detection techniques with a focus on hybrid multimodal approaches that integrate spatial, temporal and physiological features [4, 5]. Existing methods based on visual artifacts, temporal inconsistencies, frequency-domain analysis and biological signal extraction such as remote photoplethysmography (rPPG) [5, 10] are systematically reviewed. The survey further examines hierarchical fusion architectures, benchmark datasets [2, 3], evaluation protocols and real-world deployment challenges. Key limitations and open research directions are identified to guide the development of robust, generalizable and real-time deepfake detection systems.

## General Terms

Security, Artificial Intelligence, Multimedia Processing

## Keywords

Deepfake Detection,Multimodal Fusion, rPPG Signal, Transformer Models, Video Forgery Detection

## 1. INTRODUCTION

The rapid growth of digital media platforms and social networking services has dramatically transformed the creation, distribution and consumption of multimedia content. Videos, in particular, have become one of the most influential forms of communication due to their perceived realism and credibility. However, recent advances in artificial intelligence and deep learning have enabled the creation of highly realistic synthetic media, commonly known as deepfakes [1, 7], which challenge the fundamental notion of visual authenticity. Deepfake videos involve the manipulation or complete synthesis of facial appearance, expressions or speech of individuals using deep generative models, often making it difficult for human observers to distinguish between real and fabricated content [2, 3].
Deepfake generation techniques have evolved rapidly with the development of Generative Adversarial Networks (GANs) [1], variational autoencoders [2] and neural rendering methods [3]. Early approaches were limited in quality and often produced visible artifacts such as blurred facial boundaries, unnatural color transitions and inconsistent illumination. These imperfections enabled the use of handcrafted features and traditional machine learning classifiers for detection [6, 8]. However, modern deepfake models employ advanced architectures, high-resolution training data and temporal smoothing strategies, resulting in visually convincing videos that closely resemble authentic recordings [9].
The increasing realism of deepfake videos poses serious societal, ethical and security concerns. Malicious applications include political misinformation, defamation, identity impersonation, financial fraud and social engineering attacks [7]. In sensitive domains such as journalism, law enforcement and digital forensics, the inability to reliably verify video authenticity can lead to severe consequences. As a result, deepfake detection has emerged as a critical research area within multimedia forensics and computer vision [2, 3].
Early deepfake detection methods primarily focused on spatial artifact analysis. These approaches relied on detecting inconsistencies in facial textures, blending artifacts and abnormal pixel distribu-

tions using handcrafted features or shallow convolutional networks [6, 8]. While effective against early-generation deepfakes, such methods demonstrated limited robustness when applied to newer synthesis techniques. The introduction of deep learning-based detectors, particularly Convolutional Neural Networks (CNNs) [2, 9], marked a significant shift toward automated feature learning. Architectures such as XceptionNet and ResNet achieved high accuracy on benchmark datasets by learning discriminative spatial representations directly from raw video frames [2].

Despite their success, purely spatial deep learning models suffer from several limitations. Their performance often degrades under common video transformations such as compression, resizing, reencoding and noise addition, which are frequently applied by social media platforms [3]. Moreover, spatial-only models exhibit poor generalization across datasets, as they tend to overfit to dataset-specific artifacts rather than learning intrinsic properties of deepfake manipulation [2]. Transformer-based models and attention mechanisms have been proposed to address some of these issues, but generalization remains a major challenge [9].

To overcome the limitations of frame-level analysis, researchers have explored temporal modeling techniques that capture inconsistencies across consecutive frames [8]. Temporal approaches analyze facial dynamics, eye blinking patterns, head movements and lip synchronization using recurrent neural networks, temporal convolutional networks and attention-based architectures [8, 9]. By modeling motion coherence, these methods improve robustness against static artifact suppression. However, recent deepfake generation pipelines increasingly preserve temporal consistency, reducing the effectiveness of purely temporal cues and increasing the need for more intrinsic detection signals [3].

Physiological signal-based detection has recently emerged as a promising direction due to its strong theoretical foundation [4, 5]. Remote photoplethysmography (rPPG) enables the extraction of subtle cardiovascular signals from facial skin regions by analyzing periodic color variations caused by blood flow [5, 10]. These biological signals are inherently present in real videos but are difficult to accurately reproduce in synthetic videos generated by deepfake models. As a result, rPPG-based detection methods have demonstrated improved cross-dataset generalization and robustness compared to visual-only approaches [5]. However, physiological signal extraction is sensitive to motion, lighting variations and occlusions, necessitating careful signal processing and robust modeling techniques [4, 5].

Given the complementary strengths and weaknesses of spatial, temporal and physiological cues, recent research has shifted toward hybrid multimodal deepfake detection frameworks [10]. These approaches integrate multiple feature modalities to improve detection reliability under real-world conditions. In particular, hierarchical and nested fusion architectures process each modality independently before progressively combining them, preserving modality-specific information while enabling effective cross-modal interaction Such architectures offer improved interpretability, scalability and robustness compared to early or late fusion strategies .

This survey focuses on deepfake video detection using hybrid multimodal features, with special emphasis on nested hierarchical models and physiological signal analysis using rPPG [4, 5, 10]. The paper systematically reviews existing detection approaches, categorizes them based on feature modality and learning strategy, and analyzes commonly used datasets and evaluation protocols [2, 3]. Furthermore, it identifies key limitations in current systems and outlines future research directions aimed at developing robust, generalizable and real-time deepfake detection solutions suitable for deployment in real-world scenarios.Broader media forensic perspectives highlight the growing importance of adaptive deepfake detection frameworks in evolving digital ecosystems [14].

The remainder of this paper is organized as follows. Section II reviews deepfake generation techniques, Section III discusses existing detection methodologies, Section IV summarizes datasets and evaluation metrics, and Section V concludes the survey with future research directions.

## 2. TAXONOMY OF DEEPFAKE DETECTION APPROACHES

Deepfake detection techniques can be systematically categorized based on the types of cues they exploit, the level at which inconsistencies are analyzed and the learning paradigms they employ. Broadly, existing approaches rely on visual artifacts, temporal inconsistencies, physiological signals or combinations of multiple modalities [10]. While early methods focused on detecting visible synthesis artifacts, recent research increasingly emphasizes multimodal fusion strategies to improve robustness and generalization under real-world conditions [5]. This section presents a structured taxonomy of deepfake detection approaches, highlighting their underlying principles, strengths and limitations.

### 2.1 Visual Artifact-Based Detection

Visual artifact-based detection methods analyze frame-level inconsistencies introduced during the face manipulation or synthesis process [6, 8].Face warping artifacts introduced during generative manipulation remain a reliable cue for identifying forged facial regions [15]. Artifact-based detection methods also examine geometric distortions such as face warping inconsistencies introduced during synthesis [15]. These approaches primarily focus on spatial distortions that arise due to imperfect blending, resolution mismatch or limitations of generative models [1, 2].

—**Spatial CNN-Based Methods**: Spatial CNN-based approaches employ deep convolutional neural networks such as XceptionNet, ResNet, VGGNet and EfficientNet to learn discriminative visual features directly from individual video frames [2, 9]. These models are trained to identify subtle artifacts including unnatural facial textures, inconsistent skin tones, boundary blending errors and abnormal lighting patterns. Such methods have demonstrated strong performance on benchmark datasets like FaceForensics++, Celeb-DF and DFDC [2, 3]. However, their effectiveness often decreases under heavy video compression, resizing and post-processing [3]. Additionally, these models tend to overfit dataset-specific artifacts, resulting in limited cross-dataset generalization [2].

—**Frequency Domain-Based Methods**: Frequency-based detection techniques analyze artifacts in the spectral domain that are often introduced by convolutional upsampling and interpolation operations used in deepfake generation models [1]. By transforming video frames into the frequency domain using Discrete Cosine Transform (DCT), Fast Fourier Transform (FFT) or wavelet decomposition, these methods capture abnormal frequency distributions that are less perceptible in the spatial domain [6]. Frequency-based approaches improve robustness against visual post-processing and compression to some extent [6]. However, recent generative models increasingly incorporate frequency-aware training, which reduces the effectiveness of purely frequency-based detectors [1].Recent studies show that frequency-domain analysis can expose subtle synthesis artifacts that are difficult to observe in spatial representations [12].

### 2.2 Temporal Consistency-Based Detection

Temporal consistency-based methods exploit inconsistencies across consecutive video frames by analyzing facial motion patterns and dynamic behavior [8]. Unlike spatial approaches that treat frames independently, temporal models aim to capture sequential dependencies that are difficult for deepfake generators to replicate accurately [9].

—**Attention-Based Temporal Models**: Attention-based temporal detection approaches utilize transformer architectures, temporal convolutional networks or recurrent neural networks to model long-range dependencies across video frames [9]. By dynamically assigning higher importance to temporally inconsistent regions, these models can detect abnormal motion patterns related to blinking, lip synchronization and head movement [6, 8]. Temporal attention mechanisms improve robustness in videos with complex motion and background variation [9]. However, advanced deepfake generation techniques increasingly enforce temporal smoothness, reducing the discriminative power of motion-based cues when used in isolation [3].

### 2.3 Physiological Signal-Based Detection

Physiological signal-based methods leverage intrinsic biological signals that naturally occur in real human videos but are difficult for generative models to accurately reproduce [5, 10]. These approaches provide strong theoretical motivation and improved generalization capabilities [4].

—**rPPG-Based Detection**: Remote photoplethysmography (rPPG) techniques extract subtle cardiovascular signals from facial skin regions by analyzing periodic color variations caused by blood circulation [5, 10]. Since deepfake synthesis pipelines primarily focus on visual realism and often ignore underlying physiological consistency, the extracted rPPG signals from fake videos exhibit abnormal frequency patterns, phase inconsistencies or reduced signal strength . rPPG-based detection has demonstrated strong cross-dataset performance and robustness against visual compression [5]. Nevertheless, accurate rPPG extraction remains challenging under significant head motion, illumination changes and low-resolution video conditions [4].

—**Micro-Expression and Eye Movement Analysis**: These approaches analyze involuntary facial responses such as eye blinking frequency, gaze direction changes and micro-expressions that occur naturally in real videos [6]. Deepfake models may generate unnatural blinking patterns or fail to preserve subtle muscle movements around the eyes and mouth [6, 8]. By modeling these physiological cues using temporal or statistical analysis, such methods provide complementary information to visual artifact-based detectors [4]. However, their reliability can be affected by occlusions, eyewear and rapid head movement [6].

### 2.4 Hybrid Multimodal and Hierarchical Models

Hybrid multimodal approaches integrate multiple complementary cues to overcome the limitations of single-modality detectors [10]. By jointly modeling spatial, temporal and physiological information, these methods aim to achieve higher robustness and generalization [5]. Transformer-driven forensic architectures further enhance representation learning and cross-domain generalization in deepfake detection [16].

—**Early and Late Fusion Models**: Early fusion approaches combine features from different modalities at the input or feature extraction stage, allowing joint representation learning . In contrast, late fusion techniques aggregate modality-specific predictions using ensemble strategies or decision-level fusion . While both approaches improve detection performance compared to unimodal models, early fusion may suffer from feature dominance issues, whereas late fusion often fails to capture intermodal dependencies effectively .

—**Nested Hierarchical Architectures**: Nested hierarchical models process spatial, temporal and physiological features independently through specialized subnetworks before progressively fusing them across multiple hierarchical levels . This design preserves modality-specific information while enabling structured cross-modal interaction [10]. Hierarchical fusion improves interpretability, robustness to noise and cross-dataset generalization, making such architectures particularly suitable for real-world deepfake detection . These models form the foundation of recent state-of-the-art hybrid detection frameworks [10].

## 3. SURVEY OF EXISTING WORKS

Early research in deepfake detection primarily focused on identifying spatial artifacts introduced during the face synthesis and manipulation process [6, 8]. These initial approaches relied heavily on handcrafted features such as texture descriptors, color inconsistencies and facial landmark deviations, which were then classified using traditional machine learning algorithms [6]. While such methods were effective against early-generation deepfakes, their reliance on manually designed features limited their scalability and robustness [6].

The introduction of deep learning-based detection methods marked a significant shift in the research landscape [2, 9]. Convolutional Neural Networks (CNNs) enabled automated feature learning from raw image data, eliminating the need for manual feature engineering [2]. Architectures such as XceptionNet and ResNet achieved high accuracy on benchmark datasets by learning discriminative spatial representations directly from raw video frames [2, 11, 13]. However, subsequent cross-dataset evaluations revealed a critical limitation: models trained on a specific dataset often failed to generalize to unseen manipulation techniques and datasets [2, 3]. This performance degradation highlighted the tendency of CNN-based detectors to overfit dataset-specific artifacts rather than learning intrinsic properties of deepfake content [2].Capsule-network based forensic models further improve detection by preserving hierarchical spatial relationships in manipulated faces [13].

To address the limitations of frame-level spatial analysis, researchers began exploring temporal modeling techniques that leverage motion-based inconsistencies across video frames [8]. Temporal approaches analyze facial dynamics such as eye blinking frequency, head motion, lip synchronization and expression transitions using recurrent neural networks, temporal convolutional networks and attention-based architectures [8, 9]. Several studies demonstrated that incorporating temporal information improves robustness to visual noise, compression and partial occlusions [8]. However, as deepfake generation pipelines evolved, temporal coherence became increasingly well-preserved, reducing the effectiveness of motion-based detection methods when used independently [3].

More recent research has shifted toward physiological signal-based detection, motivated by the observation that deepfake synthesis models primarily focus on visual realism while neglecting underlying biological signals [4, 5]. Remote photoplethysmography (rPPG) techniques extract subtle cardiovascular signals from facial

skin regions by analyzing periodic color variations over time . Multiple studies have shown that real videos exhibit stable and consistent heart rate patterns, whereas deepfake videos often display disrupted or inconsistent rPPG signals . These methods demonstrate strong cross-dataset generalization and improved robustness against visual post-processing, making physiological cues a promising direction for deepfake detection [4, 5]. Nonetheless, accurate rPPG extraction remains challenging in the presence of head motion, illumination variation and low-quality video inputs [4].
Building upon the strengths of individual detection paradigms, recent state-of-the-art systems adopt hybrid multimodal frameworks that integrate spatial, temporal and physiological features within unified architectures [10]. These approaches typically employ separate subnetworks for each modality, followed by structured fusion mechanisms to combine complementary information . Hierarchical and nested fusion strategies have been shown to preserve modality-specific representations while enabling effective cross-modal interaction . Experimental results consistently indicate that such hybrid models outperform unimodal detectors across multiple datasets and manipulation techniques . However, these systems introduce additional computational complexity, increased training requirements and synchronization challenges between modalities, highlighting the need for efficient and scalable multimodal designs .
Overall, the evolution of deepfake detection research reflects a gradual transition from artifact-specific analysis toward holistic, multimodal understanding [10]. Comprehensive forensic surveys emphasize the need for scalable and adaptable detection pipelines for evolving manipulation techniques [18]. While hybrid frameworks represent the current state of the art, challenges related to generalization, real-time deployment and robustness under unconstrained conditions remain open research problems .

## 4. COMPREHENSIVE EVALUATION OF DEEPFAKE VIDEO DETECTION METHODS

Deepfake detection research has been evaluated using multiple benchmark datasets, performance metrics, and experimental protocols. Table 4 summarizes representative studies...
It is evident that many studies rely primarily on single-dataset evaluation, which may overestimate detection accuracy in controlled environments. This suggests that cross-dataset testing is essential for measuring real-world robustness. A key observation is that large-scale datasets such as DFDC introduce greater attack diversity, exposing limitations in earlier detection methods. However, these results also reveal inconsistencies in evaluation protocols across studies, emphasizing the need for standardized benchmarking frameworks.
The evaluation comparison highlights several important observations. Many early studies primarily relied on single-dataset training and testing, resulting in inflated accuracy values that do not reflect real-world performance. Cross-dataset evaluation, which measures the ability of models to generalize to unseen manipulation methods, is still inconsistently applied across studies. Large-scale datasets such as DFDC introduced greater attack diversity, revealing that many previously high-performing detectors experience significant performance degradation when exposed to novel synthesis techniques. Physiological signal-based approaches and hybrid multimodal frameworks demonstrate stronger cross-dataset robustness, suggesting that intrinsic and multi-cue representations improve real-world reliability.

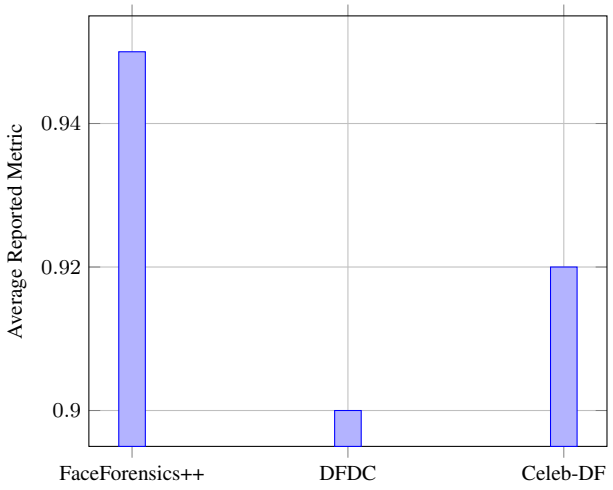


Fig. 1. Dataset-wise reported performance distribution for deepfake detection methods.

### 4.1 Limitations of Existing Evaluation Practices

Despite rapid progress in deepfake detection research, current evaluation practices exhibit several weaknesses. Many studies rely heavily on benchmark datasets collected under controlled conditions, which do not fully represent real-world variations in lighting, compression, camera motion, and demographic diversity. Furthermore, performance metrics such as accuracy alone may provide misleading conclusions when dataset class imbalance exists, making AUC, F1-score, and cross-dataset validation more reliable indicators. Another limitation is the insufficient evaluation against emerging deepfake generation pipelines, as most benchmarks contain only a subset of possible manipulation techniques. Additionally, standardized protocols for testing robustness under compression, adversarial perturbations, and social-media processing are still lacking, making fair comparison between detection systems difficult.

### 4.2 Evaluation-Focused Future Directions

Future deepfake detection research should prioritize standardized and rigorous evaluation methodologies. Large-scale multi-source datasets reflecting real-world video conditions, including diverse compression levels, illumination changes, demographic variation, and device heterogeneity, must be developed. Cross-dataset validation should become a mandatory evaluation component to ensure model generalization beyond controlled training environments. Benchmark protocols should also incorporate adversarial testing scenarios, manipulation diversity analysis, and robustness measurement under post-processing operations such as resizing, filtering, and re-encoding. Furthermore, unified reporting standards combining multiple metrics including AUC, F1-score, precision–recall curves, and computational efficiency will enable more transparent and comparable assessment of detection performance. Establishing such evaluation frameworks will be essential for advancing reliable and deployable deepfake detection systems. As shown in Fig. 1, performance varies across datasets... This suggests that dataset characteristics strongly influence reported detection results. A key observation is that large-scale and diverse datasets such as DFDC often produce lower average scores, reflecting increased manipulation diversity. However, these results also reveal poten-

Table 1. Evaluation-Centric Comparison of Deepfake Detection Studies

| Study | Dataset | Metrics Used | Cross-Dataset Eval | Attack Diversity | Key Results |
|---|---|---|---|---|---|
| Li et al. (Face X-ray) | FaceForensics++ | AUC, Accuracy | Limited | Moderate manipulation types | High detection accuracy for known forgery patterns |
| Rössler et al. (FaceForensics++) | Multiple datasets | Accuracy, ROC | Partial | Multiple synthesis methods | Highlighted generalization limitations of CNN detectors |
| DFDC Challenge Models | DFDC dataset | AUC, Log-loss | No | Large-scale manipulation diversity | Performance dropped significantly on unseen videos |
| Ciftci et al. (FakeCatcher) | Celeb-DF, DFDC | AUC, Cross-dataset accuracy | Yes | Physiological signal analysis | Improved robustness across datasets using rPPG cues |
| Recent Hybrid Multimodal Studies | Multiple datasets | F1-score, Accuracy, AUC | Yes | Broad manipulation coverage | Highest reliability achieved through multimodal fusion |

tial dataset bias, highlighting the need for cross-dataset evaluation to ensure reliable real-world deployment.

## 5. EXPERIMENTAL RESULTS AND COMPARATIVE ANALYSIS IN LITERATURE

Several studies have reported experimental performance of deepfake detection methods across different benchmark datasets and evaluation protocols. Table 2 summarizes representative works from the literature, including datasets used, detection methodology, feature modality, evaluation metric, and reported performance values.

It is evident that spatial CNN-based methods achieve strong performance on benchmark datasets, particularly when trained and tested under controlled conditions. This suggests that visual artifact detection remains effective for known manipulation techniques. A key observation is that temporal and physiological models demonstrate improved robustness when handling motion inconsistencies and intrinsic biological patterns. However, these results also reveal that performance may degrade under cross-dataset scenarios, highlighting the importance of multimodal detection strategies for real-world applications.

The tabulated results show that spatial CNN-based methods achieve strong accuracy on controlled datasets such as FaceForensics++, but their performance often decreases under cross-dataset evaluation. Temporal models improve detection reliability by capturing motion-based inconsistencies across frames, though they introduce higher computational requirements. Physiological signal-based approaches demonstrate improved robustness by exploiting intrinsic biological patterns that are difficult for generative models to reproduce. Recent multimodal fusion architectures consistently report the highest overall performance, indicating that combining multiple feature modalities is currently the most effective strategy for reliable deepfake detection.

## 6. COMPARATIVE ANALYSIS

Table 3 presents a structured comparison of representative deepfake detection approaches based on feature modality, datasets, strengths, limitations, and major findings.

It is evident that spatial-only approaches provide high frame-level accuracy but often suffer from limited generalization across datasets. This suggests that incorporating temporal and physiological cues improves detection robustness. A key observation is that hybrid multimodal architectures consistently outperform unimodal models by integrating complementary information sources. However, these results also reveal increased computational complexity, indicating a trade-off between detection accuracy and deployment feasibility.

The comparison demonstrates that spatial CNN-based approaches provide strong frame-level detection performance but often fail to generalize across datasets. Temporal models improve robustness by incorporating sequential motion patterns, although their computational complexity is higher. Physiological signal-based methods offer improved generalization due to their reliance on intrinsic biological cues, but they remain sensitive to environmental variations such as lighting and head motion. Hybrid multimodal architectures integrate spatial, temporal, and physiological information, consistently achieving the most reliable performance across datasets, though at the cost of increased model complexity and processing requirements.

### 6.1 Trend Analysis

An examination of recent research trends reveals a clear evolution in deepfake detection methodologies. Early studies primarily relied on spatial artifact detection using handcrafted features and CNN-based frame analysis. Later approaches incorporated temporal modeling techniques to capture motion-based inconsistencies across video frames. More recent work increasingly focuses on physiological signal extraction and hierarchical multimodal fusion frameworks. This progression indicates a shift from surface-level artifact detection toward intrinsic and cross-modal feature modeling aimed at improving robustness in real-world scenarios.

### 6.2 Research Gaps and Open Challenges

Despite significant advances, several research gaps remain unresolved. First, many detection models exhibit strong performance on benchmark datasets but demonstrate limited generalization when applied to real-world social media videos with compression, noise, and resolution variations. Second, physiological signal extraction methods such as rPPG remain sensitive to motion, illumination changes, and occlusions, reducing reliability under unconstrained conditions. Third, hybrid multimodal architectures introduce substantial computational complexity, limiting their deployment on resource-constrained platforms. Additionally, most existing systems lack interpretability, making it difficult to explain detection decisions in forensic or legal contexts. Addressing these limitations is essential for developing practical, scalable, and trustworthy deepfake detection systems.

### 6.3 Future Directions Grounded in Analysis

Future research should focus on designing lightweight hierarchical multimodal models capable of real-time deployment while pre-

Table 2. Reported Experimental Results in Deepfake Detection Literature

| Study | Dataset | Method | Features | Metric | Reported Performance |
|---|---|---|---|---|---|
| Li et al. (Face X-ray) | FaceForensics++ | CNN-based forgery detector | Spatial artifacts | AUC | $\approx 0.95$ |
| Rössler et al. | FaceForensics++ | XceptionNet | Spatial deep features | Accuracy / AUC | $> 90\%$ accuracy |
| Guera and Delp | DFDC / Video datasets | CNN–LSTM | Spatio-temporal motion cues | Accuracy | $\approx 90\%$ |
| Ciftci et al. (FakeCatcher) | Celeb-DF, DFDC | rPPG-based detection | Physiological signals | AUC | $\approx 0.97$ |
| Recent Multimodal Studies | Multiple datasets | Hierarchical fusion networks | Spatial + temporal + physiological | F1-score | $\approx 0.93$ |

Table 3. Comparative Analysis of Deepfake Detection Methods

| Feature Type | Model | Dataset | Strengths | Limitations | Key Findings |
|---|---|---|---|---|---|
| Spatial | XceptionNet | FaceForensics++ | High frame accuracy | Poor cross-dataset generalization | Detects visible manipulation artifacts effectively |
| Temporal | CNN–LSTM | DFDC | Captures motion inconsistencies | High computational cost | Improves sequence-level detection reliability |
| Physiological | FakeCatcher (rPPG) | Celeb-DF | Strong generalization capability | Sensitive to lighting and motion | Biological signals provide intrinsic detection cues |
| Hybrid Multimodal | Hierarchical Fusion Models | Multiple datasets | Best overall robustness | Complex architecture | Combining modalities yields state-of-the-art performance |

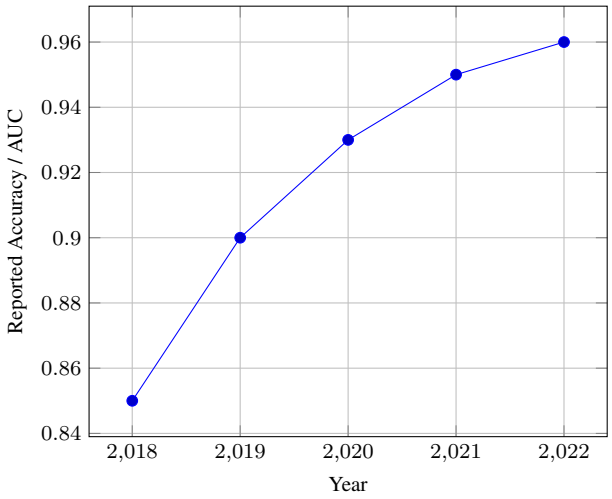


Fig. 2. Performance trends in deepfake detection methods over time.

serving detection robustness. Self-supervised and domain-adaptive learning strategies can help improve cross-dataset generalization and reduce dependence on large labeled datasets. Robust physiological signal extraction techniques incorporating motion compensation and illumination normalization should be further explored. Moreover, explainable AI mechanisms that highlight the spatial, temporal, or physiological cues influencing model predictions will be critical for adoption in sensitive applications such as digital forensics, journalism, and online content verification. Finally, the development of large-scale, diverse, and standardized datasets reflecting real-world video conditions will play a crucial role in advancing reliable deepfake detection research.

As illustrated in Fig. 2, reported detection performance has steadily improved over time.... This suggests that recent approaches incorporating temporal and physiological features contribute significantly to improved robustness. A key observation is that performance gains become gradual in later years, indicating increasing maturity of detection architectures. However, these results also reveal that high benchmark accuracy does not always guarantee real-world generalization.

## 7. LIMITATIONS AND OPEN CHALLENGES

Despite significant advancements in deepfake video detection using hybrid multimodal approaches [10], several challenges continue to hinder their large-scale, real-world deployment. One of the foremost challenges is *dataset bias and limited generalization* [2, 3, 10]. Most existing detection models are trained on datasets such as FaceForensics++, DFDC, and Celeb-DF, which do not fully capture the diversity of real-world conditions [2, 3]. Differences in lighting, camera resolution, facial demographics, video compression, and motion patterns can substantially degrade cross-dataset performance [10]. Models that perform well in controlled experimental settings often fail to generalize to unseen deepfake generation methods or videos collected from social media platforms [3].Detection systems must also address increasingly realistic machine-generated content that closely resembles authentic facial patterns [17].

Another significant limitation is *sensitivity to video quality and compression* [3, 5]. Many spatial and temporal detection methods rely on subtle frame-level artifacts, which can be easily diminished by lossy compression, resizing, or re-encoding applied by online platforms [3]. Physiological signal-based methods such as rPPG extraction are particularly affected, as small color variations necessary for heart rate estimation may be obscured under low resolution or compression artifacts, leading to false negatives [5].

*Environmental factors and illumination variations* further constrain reliability [4, 5]. Variations in lighting, shadows, reflections, and background movement can introduce noise into both visual and physiological features [4]. For instance, rPPG signals can be significantly distorted by changes in ambient illumination or rapid head movement, reducing detection accuracy in unconstrained real-world videos [5].

*Motion and occlusion challenges* also limit performance [6, 8]. Large head rotations, partial facial occlusions (e.g., glasses, masks, hair), and fast camera movements can disrupt temporal coherence and physiological feature extraction [4, 5]. Even advanced temporal models and attention-based networks struggle to maintain robustness under such conditions, which are common in social media and user-generated content [8].

*Computational complexity and real-time constraints* pose another challenge [10]. Nested hierarchical multimodal architectures integrate multiple subnetworks for spatial, temporal, and physiological feature processing, followed by fusion layers . While highly accurate, these models are computationally expensive, hindering deployment on resource-limited devices such as mobile phones, laptops, or streaming platforms . Efficient design and optimization are essential for practical real-time applications [10].

*Adversarial vulnerability and robustness* remains an open problem [10]. Deepfake detection systems are susceptible to adversarial perturbations, where slight modifications to input frames can fool models without noticeable changes to human observers [10]. Attackers can exploit this vulnerability to bypass detection, raising security and trust concerns for forensic and automated verification systems [10].

Finally, there is a *lack of explainability and interpretability* in most deep learning-based detectors [10]. Black-box models do not provide insight into which cues (spatial artifacts, temporal inconsistencies, or physiological signals) drive the final decision [10]. This lack of transparency can limit their adoption in sensitive applications such as law enforcement, digital forensics, and journalism [10].

## 8. FUTURE RESEARCH DIRECTIONS

Future research should prioritize the development of *robust, generalizable, and lightweight hierarchical models* capable of real-time deployment [10]. Optimizing nested fusion architectures for computational efficiency without compromising multimodal feature integration will enable deployment on mobile and edge devices while maintaining high detection accuracy [10].

Another promising direction involves *self-supervised and few-shot learning* approaches [10]. Leveraging unlabeled videos, temporal consistency, and cross-modal correlations can reduce dependence on large annotated datasets, enabling models to adapt to unseen deepfake generation techniques with minimal supervision [10].

*Robust physiological signal extraction* under unconstrained conditions is crucial [5, 10]. Techniques such as motion compensation, illumination normalization, and adaptive region-of-interest tracking can improve rPPG-based detection even in videos with significant head movement, facial occlusions, or variable lighting [5].

*Cross-dataset and cross-generation generalization* remains a critical research focus [10]. Integrating domain adaptation, adversarial training, and multimodal consistency constraints can enhance model resilience to novel deepfake methods, diverse video resolutions, and social media compression artifacts [10].

*Explainable and interpretable detection frameworks* are essential for trust and adoption in forensic applications [10]. Visualizations, attention maps, and modality-specific confidence scores can provide insights into which features contribute to the detection decision, facilitating accountability and transparency [10].

*Adversarial defense mechanisms* must be systematically explored [10]. Techniques such as adversarial training, noise-robust feature extraction, and anomaly detection in physiological signals can improve resistance to intentionally manipulated inputs designed to evade detection [10].

Finally, *standardized, large-scale, and diverse datasets* encompassing multiple demographics, video qualities, facial expressions, and real-world conditions are critical [2, 3]. Such datasets will enable fair benchmarking, improve cross-dataset performance, and accelerate progress toward practical, reliable deepfake detection systems .

**Summary:** Addressing these challenges through efficient hierarchical models, self-supervised learning, robust physiological extraction, explainability, and standardized datasets is vital for building reliable, scalable, and trustworthy deepfake detection systems suitable for deployment in diverse real-world scenarios [10].

## 9. CONCLUSION

This survey presented a comprehensive analysis of deepfake video detection methods, with a particular focus on hybrid multimodal frameworks integrating spatial, temporal, and physiological cues [10]. Detection approaches based on visual artifacts, temporal inconsistencies, and remote photoplethysmography (rPPG) were systematically reviewed to highlight their underlying principles, performance characteristics, and practical limitations [2, 4, 5, 6, 8]. The discussion emphasized the evolution from traditional frame-level analysis and handcrafted features to deep learning–driven hierarchical and multimodal architectures, which significantly improve robustness and cross-dataset generalization .

Despite notable advances, real-world deployment of deepfake detection systems remains constrained by several critical factors, including dataset bias, sensitivity to video compression, illumination variability, motion and occlusion challenges, computational complexity, and vulnerability to adversarial manipulation . High-performing hybrid models often rely on multi-stage pipelines and nested hierarchical architectures that increase computational overhead, while lightweight models frequently trade robustness for efficiency . These trade-offs underline the need for balanced system design that simultaneously addresses accuracy, scalability, and real-time applicability .

The survey further indicates that future deepfake detection systems will benefit from unified multimodal learning frameworks capable of jointly leveraging spatial, temporal, and physiological information . Advances in self-supervised and few-shot learning, robust rPPG signal extraction, explainable AI, and adversarial defense mechanisms are expected to play a central role in improving reliability under diverse and unconstrained real-world conditions [5, 10]. Additionally, the creation of large-scale, standardized, and diverse datasets will be essential for fair benchmarking, improved generalization, and accelerated progress in the field [2, 3].

Overall, addressing the identified challenges through integrated system design, robust learning strategies, and standardized evaluation protocols will be crucial for advancing scalable, accurate, and trustworthy deepfake video detection systems suitable for deployment in practical applications, including media verification, digital forensics, and online content moderation .

## 10. REFERENCES

[1] Y. Li, X. Yang, P. Sun, H. Qi and S. Lyu, “Face x-ray for more general face forgery detection,” in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2020.

[2] A. Roessler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies and M. Niessner, “FaceForensics++: Learning to detect manipulated facial images,” in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2019.

[3] B. Dolhansky et al., “The deepfake detection challenge dataset,” in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops, 2020.

[4] J. Hernandez-Ortega et al., “DeepFakesON-Phys: Deepfake detection based on heart rate estimation,” IEEE Transactions on Information Forensics and Security, 2021.

[5] T. Wang et al., “rPPG-based deepfake detection using physiological signals,” IEEE Access, 2022.

[6] C. Ciftci, I. Demir and L. Yin, “FakeCatcher: Detection of synthetic portrait videos using biological signals,” IEEE Transactions on Information Forensics and Security, 2020.

[7] P. Korshunov and S. Marcel, “Deepfakes: A new threat to face recognition? Assessment and detection,” in Proc. IEEE Int. Conf. Biometrics: Theory, Applications and Systems (BTAS), 2018.

[8] D. Guera and E. J. Delp, “Deepfake video detection using recurrent neural networks,” in Proc. IEEE Int. Conf. Advanced Video and Signal Based Surveillance (AVSS), 2018.

[9] Z. Zhao et al., “Multi-attentional deepfake detection,” in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2021.

[10] S. Mittal et al., "Emotions don't lie: An audio-visual deepfake detection method," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops, 2020.

[11] K. Shiohara and T. Yamasaki, "Detecting deepfakes with self-blended images," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1876–1885.

[12] J. Frank, T. Eisenhofer, L. Schonherr, A. Fischer, D. Kolossa and T. Holz, "Leveraging frequency analysis for deepfake image recognition," in Proc. International Conference on Machine Learning (ICML), 2020, pp. 3247–3258.

[13] H. H. Nguyen, F. Fang, J. Yamagishi and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 2307–2311.

[14] L. Verdoliva, "Media forensics and deepfakes: An overview," IEEE Journal of Selected Topics in Signal Processing,part vol. 14, no. 5, pp. 910–932, 2020.

[15] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," in Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2019, pp. 46–52.

[16] Z. Wang, X. Luo, Y. Qiu and Y. Zhang, "Transforensics: Image forgery detection through vision transformer," in Proc. IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4261–4270.

[17] S. Tariq, S. Lee, H. Kim, Y. Shin and S. S. Woo, "Detecting both machine and human created fake face images in the wild," in Proc. ACM International Conference on Multimedia, 2018, pp. 1042–1050.

[18] 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.