Deep Learning Approaches for Real-Time Deepfake Detection: A Review of CNN-Based, Temporal, Transformer-Based, and Frequency-Domain Methods

Dr. Sanjeev Kumar Shukla[1], Abhishek Gupta[2], Alok Singh[3], Ankit Dubey[4]
[1]*HOD, Computer Application Department, Axis Colleges*
[2,3,4]*Axis Institute of Higher Education*

***Abstract*—Deepfake videos are becoming more realistic and harder to detect as artificial intelligence and deep learning techniques continue to advance. Early deepfakes exhibited visible flaws such as poor face swapping and unnatural eye movement, but modern generation methods can produce highly convincing videos that are increasingly misused for misinformation, identity fraud, fraudulent video calls, and non-consensual content. This paper presents a structured review of deep learning approaches for real-time deepfake video detection. Following a screening process that narrowed an initial pool of 103 candidate articles to 73 relevant studies, detection methods are organized into four families – CNN-based spatial detection, temporal/sequence models, Transformer-based approaches, and frequency-domain analysis – and are further considered in combination with multi-modal audio-visual fusion. Each family is evaluated with respect to detection accuracy, computational cost, and suitability for real-time deployment. The review finds that highly accurate models are frequently too computationally expensive for real-time use, while lightweight models trade some accuracy for speed, and that real-world conditions such as compression, low resolution, and previously unseen manipulation techniques continue to reduce detector reliability outside benchmark settings. The paper concludes by highlighting the need for lightweight, generalizable, and multi-modal detection systems capable of operating reliably in real-time environments.**

***Index Terms*—Deepfake Detection, Deep Learning, Real-Time Detection, Convolutional Neural Networks, LSTM, Transformers, Video Forensics, Frequency-Domain Analysis**

I. INTRODUCTION

The term “deepfake” refers to synthetic audio-visual content produced using deep generative models – most commonly autoencoders and generative adversarial networks (GANs) that convincingly replaces, reenacts, or fabricates a person's likeness or voice [1], [4]. Deepfake technology can generate realistic face swaps, facial expressions, lip movements, and even cloned voices, and as the realism of this content improves, the gap between what a human viewer can reliably judge as authentic continues to narrow.

This technology has legitimate applications in entertainment, filmmaking, and education, but it is also increasingly misused to spread misinformation, impersonate individuals, defraud organizations, damage reputations, and violate personal privacy [4], [21]. As deepfakes become harder to distinguish from genuine footage, they also erode general trust in video as reliable evidence, which has turned automated detection into an urgent research problem.

Detecting deepfakes reliably is difficult for several interlocking reasons. Generation methods evolve quickly, so detectors tuned to the artifacts of one generator often fail to generalize to another [7], [13]. Real-world video is rarely pristine – compression and resolution loss strip away many of the subtle statistical cues that detectors depend on [4], [9]. Detection is also inherently adversarial in nature, since forgers can deliberately post-process manipulated content to evade known detectors. Finally, most detectors are still evaluated in narrow, dataset-specific settings that do not reflect the diversity of manipulations encountered in practice [24].

Earlier detection methods relied mainly on hand-crafted features, which do not scale well to modern generation techniques. Deep learning instead learns discriminative patterns directly from data: CNNs examine individual frames for spatial artifacts [1], [2], [3]; temporal models such as LSTM analyze relationships across consecutive frames [10]; Transformer-based models capture longer-range dependencies across a video; and frequency-domain methods surface hidden generation artifacts that are difficult for the human eye to notice [7], [8], [9].

This paper reviews CNN-based, temporal, Transformer-based, and frequency-domain approaches to real-time deepfake detection, with a focus on the trade-off between accuracy and processing speed. Section 2 describes the review methodology. Section 3 provides background on how deepfakes are generated. Section 4 reviews visual-based detection methods. Section 5 reviews multi-modal and audio-visual approaches. Section 6 discusses cross-cutting challenges and benchmark datasets. Section 7 discusses future directions, and Section 8 concludes the paper.

II. REVIEW METHODOLOGY

This review follows a structured screening process to identify and evaluate deep learning approaches used for real-time deepfake detection. Relevant literature was collected from major academic sources in artificial intelligence, deep learning, computer vision, and video forensics, using search terms including "deepfake detection," "real-time deepfake detection," "CNN-based deepfake detection," "Transformer-based deepfake detection," and "video forgery detection."

Candidate articles were screened in successive passes. In the first pass, titles and abstracts were reviewed to exclude work that was off-topic, such as deepfake generation papers with no detection component. In the second pass, remaining articles were checked to remove duplicates, extended-abstract versions of already-included papers, and studies whose methodology was not clearly reproducible.Table 1 summarizes this screening funnel.

Stage	Criterion Applied	Articles Remaining
Initial retrieval	Keyword search across major databases	103
Title/abstract screening	Topical relevance to real-time deepfake detection	~87
Duplicate / extended-abstract removal	Removal of redundant or superseded versions	~79
Full-text screening	Reproducibility and clear methodology	73

Table 1. Article screening funnel used in this review.

The 73 retained studies were organized into CNN-based, temporal (RNN/LSTM), Transformer-based, and frequency-domain categories, and compared on detection accuracy, computational requirements, processing speed, and robustness, with particular attention paid to real-time suitability.

III. BACKGROUND: HOW DEEPFAKES ARE GENERATED

Detection methods are, in large part, a response to the specific artifacts left behind by generation pipelines, so a brief overview of generation techniques is useful context. Video deepfakes are commonly grouped into a small number of manipulation categories: face swapping, where the identity of a source face is transferred onto a target video while preserving the target's expressions and pose; facial reenactment, where identity is preserved but expressions, gaze, or head pose are driven by a source performance; attribute manipulation, which alters attributes such as age, hair, or expression without a full identity swap; and lip-sync / audio-driven generation, where mouth movements are synthesized to match a different or synthetic audio track.

Underlying these manipulations are autoencoder-based face-swapping pipelines [1], generative adversarial networks – including high-fidelity StyleGAN-family architectures – and, more recently, diffusion-based generators. Each generation family tends to leave characteristic traces: blending boundaries at the swapped-face contour, unnatural or missing physiological signals such as blink rate [5], spectral irregularities introduced by up-sampling operations in GAN decoders [9], or subtle temporal

inconsistencies between synthesized mouth motion and accompanying audio [14]. Detection research has largely evolved by identifying and modeling one or more of these traces.

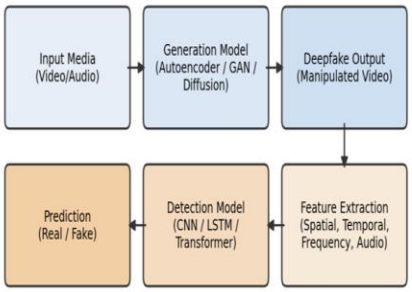


Fig. 1. General pipeline linking deepfake generation to detection: manipulated output is analyzed via spatial, temporal, frequency, and/or audio-visual features prior to classification.

IV. VISUAL-BASED DETECTION METHODS

4.1 Spatial CNN-Based Detection

Most early and still widely used deepfake detectors treat detection as a binary frame-classification problem solved with CNNs operating in the pixel domain. Compact architectures such as MesoNet [1] were designed to exploit mesoscopic image properties to flag facial tampering with relatively few parameters, making them efficient for near real-time use. Deeper backbones such as XceptionNet [2], with its depth-wise separable convolutions, became a de facto baseline after strong performance on the FaceForensics++ benchmark [4], and EfficientNet-style compound-scaling architectures have been used to balance accuracy against computational cost. Capsule-network-based detectors [3] have also been proposed to model part-whole spatial relationships that convolutional pooling tends to discard.

A recurring finding is that spatial CNNs, while highly accurate within the manipulation type, they are trained on, generalize poorly to novel generators and unfamiliar compression levels [7], [13]. This has motivated temporal extensions – CNN–LSTM combinations or 3D convolutions that capture inter-frame inconsistencies such as unnatural blinking or boundary jitter [10] – as well as attention-based architectures that direct the network toward manipulation-sensitive regions such as the eyes, nose, and mouth. Biological-signal-based methods form a related branch, exploiting physiological cues such as blink frequency [5] or remote photoplethysmography that generative models struggle to reproduce faithfully [21]; these cues tend to generalize somewhat better across manipulation types but remain sensitive to video quality and lighting.

4.2 Frequency-Domain Detection

A second major branch of visual detection operates in the frequency domain. GAN-based generators, which typically rely on transposed-convolution up-sampling, introduce periodic spectral artifacts that are difficult to perceive visually but appear as irregularities in the Fourier or Discrete Cosine Transform spectrum [8], [9]. The most widely cited method in this space, F3-Net [7], uses a dual-branch design combining a Frequency-Aware Image Decomposition branch with a Local Frequency Statistics branch to expose subtle forgery patterns. Related work has combined phase-spectrum information with spatial features [12], used wavelet transforms for improved robustness to compression, and fused high-frequency residuals with RGB features to improve generalization [11], [13].

Frequency-based methods are particularly attractive for compressed video, since compression itself operates largely in the DCT domain, giving these detectors a natural point of leverage. However, frequency-only detectors can be confounded by lighting variation and textured backgrounds unrelated to manipulation, which is why recent architectures increasingly fuse spatial and frequency cues rather than relying on either domain alone.

Family	Representative Techniques	Key Strength	Key Limitation
CNN (spatial)	MesoNet, XceptionNet, EfficientNet, capsule networks	High in-dataset accuracy; efficient in compact variants	Weak cross-generator generalization; degrades under compression
Temporal (LSTM/3D-CNN)	CNN–LSTM, 3D convolutions	Captures inter-frame motion cues	Higher latency; sequence-length sensitive
Transformer-based	Cross-attention / long-range encoders	Models' long-range dependencies; strong accuracy	Computationally expensive for real-time use

Frequency-domain	F3-Net (FAD/LFS), wavelet, high-frequency fusion	Exposes up-sampling artifacts; robust to compression	Sensitive to lighting/texture confounds
Multi-modal / audio-visual	SyncNet-style sync models, cross-modal fusion	Best reported cross-manipulation generalization	Higher cost; needs aligned audio and video

Table 2. Summary comparison of the detection method families reviewed in this paper.

V. MULTI-MODAL AND AUDIO-VISUAL METHODS

Multi-modal detectors combine visual and audio evidence, most commonly by modeling the temporal synchronization between lip movement and speech. The underlying premise is that even highly realistic generators struggle to perfectly align synthesized mouth motion with an independently generated or retained audio track, so measurable lag or inconsistency between the two modalities can serve as a forgery signal that is largely independent of the specific visual generator used [14]. Foundational synchronization-based detectors have been shown to achieve stronger cross-manipulation generalization than unimodal visual detectors trained on a single manipulation type [15]. Building on this idea, subsequent systems combine CNN–LSTM temporal encoders with dedicated audio-visual synchronization networks to jointly analyze lip-motion synchronization and blink patterns, fusing modality-specific scores into a single confidence estimate.

Other strategies extract features from each modality independently and fuse them at different stages of the network, including affective-cue mismatch detection that compares whether perceived emotion in the voice matches that expressed on the face [16]. More recent transformer-based ensembles apply cross-modal attention across audio and visual embeddings to model longer-range temporal dependencies than convolutional or recurrent approaches typically capture. Overall, multi-modal fusion tends to outperform unimodal detectors on cross-manipulation and cross-dataset generalization benchmarks, at the cost of increased architectural complexity, higher computational requirements, and a dependency on both modalities being present and of sufficient quality – conditions not always met in compressed or low-bandwidth real-world video.

VI. CROSS-CUTTING CHALLENGES AND BENCHMARKS

6.1 Compression, Resolution, and Generalization

Video shared on social platforms is almost always re-encoded, which removes or distorts many of the fine-grained spatial and frequency artifacts that detectors rely on; low input resolution compounds this effect by further degrading texture and blending-boundary cues. Perhaps the most consistently reported limitation across the reviewed literature is poor generalization: detectors trained on one generation method or benchmark dataset frequently show large performance drops when evaluated on manipulations from different generators or “in-the-wild” content [7], [13], [24].

6.2 Adversarial Attacks and Computational Cost

Because detection is fundamentally adversarial, forgers with or without knowledge of a detector's architecture can add imperceptible perturbations specifically designed to flip a detector's decision, placing a robustness burden on detection systems that most models evaluated on clean benchmark data do not satisfy. Separately, real-time detection must process video quickly, but the most accurate models tend to be the most computationally expensive, creating a persistent accuracy-speed trade-off that efficient model design must address.

6.3 Benchmark Datasets

Face Forensics++ [4] remains the most widely used benchmark for spatial and frequency-domain visual detectors, offering multiple manipulation types at controlled compression levels. The DeepFake Detection Challenge (DFDC) dataset [23] is commonly used to evaluate detectors under more diverse identities, lighting, and manipulation techniques, and is frequently paired with FaceForensics++ in cross-dataset studies. More recent benchmarking efforts such as Deepfake Bench [24] re-implement published methods under a common evaluation pipeline to produce more directly comparable generalization results, an effort motivated

by inconsistent evaluation practices noted throughout the reviewed literature.

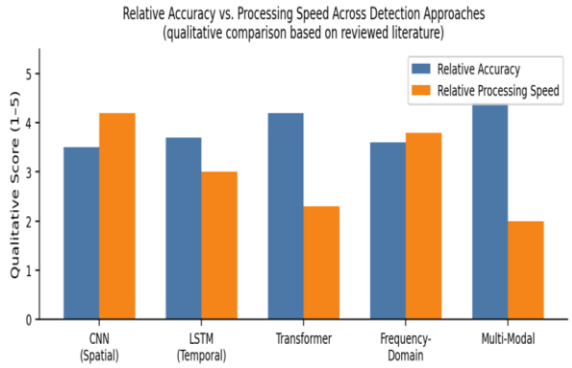


Fig. 2. Qualitative accuracy-versus-speed comparison across detection method families, based on trends discussed in the reviewed literature.

VII. DISCUSSION AND FUTURE DIRECTIONS

Synthesizing across the reviewed studies, three patterns stand out. First, no single-modality detector reviewed here demonstrates consistently strong generalization across generators, datasets, and compression levels; reported gains are usually specific to the benchmark on which a method was tuned. Second, frequency-domain and biological-signal cues appear to offer more transferable evidence than raw spatial pixel statistics, but are best used as complements to, rather than replacements for, spatial CNN features. Third, multi-modal audio-visual synchronization methods show the most promising generalization behavior, consistent with the intuition that combining independent evidence sources raises the cost of successful evasion, though this comes with added computational and data-alignment overhead.

Future work should prioritize generalization – models that learn transferable features rather than artifacts tied to specific generation methods – and efficiency, building lightweight models suitable for resource-constrained, real-time deployment in social media, video conferencing, and identity verification. Continued investment in multi-modal fusion architectures that combine spatial, frequency, and audio-visual synchronization evidence, ideally with gating mechanisms that adapt to missing or degraded modalities, is also warranted. Additional open problems include detectors that remain effective against diffusion-based generators, which are less reliant on the up-sampling operations that motivate much of the current frequency-domain literature, as well as standardized benchmarking protocols and interpretable systems that explain which regions of a video appear suspicious rather than returning a bare real/fake label.

VIII. CONCLUSION

Deepfake technology has advanced rapidly alongside AI and deep learning, offering value in entertainment and digital media while carrying real risks of misinformation, fraud, privacy violation, and reputational harm. This review examined CNN-based, temporal, Transformer-based, frequency-domain, and multi-modal approaches to real-time deepfake detection. Each family offers distinct strengths, but real-world factors – compression, low resolution, unseen manipulation techniques, adversarial attacks, and computational cost – continue to limit performance outside benchmark conditions. Effective real-time detection requires balancing accuracy, speed, and computational efficiency, and combining visual, frequency, and audio-visual information tends to improve robustness.

No single approach solves every challenge. Future research should focus on lightweight, accurate, and generalizable models capable of handling both existing and emerging deepfake techniques, helping real-time detection systems play a stronger role in safeguarding the security and reliability of digital content.

REFERENCES

[1] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, “MesoNet: A compact facial video forgery detection network,” in *Proc. IEEE Int. Workshop Inf. Forensics Security (WIFS)*, 2018.

[2] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1251–1258.

[3] H. H. Nguyen, J. Yamagishi, and I. Echizen, “Capsule-forensics: Using capsule networks to detect forged images and videos,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2019, pp. 2307–2311.

[4] Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 1–11.

[5] Y. Li, M.-C. Chang, and S. Lyu, "In Ictu Oculi: Exposing AI created fake videos by detecting eye blinking," in *Proc. IEEE Int. Workshop Inf. Forensics Security (WIFS)*, 2018.

[6] R. Caldelli, L. Galteri, I. Amerini, and A. Del Bimbo, "Optical flow-based CNN for detection of unlearnt deepfake manipulations," *Pattern Recognit. Lett.*, vol. 146, pp. 31–37, 2021.

[7] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020.

[8] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging frequency analysis for deep fake image recognition," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020.

[9] R. Durall, M. Keuper, and J. Keuper, "Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020.

[10] Masi, A. Killekar, R. M. Mascarenhas, S. P. Gurudatt, and W. AbdAlmageed, "Two-branch recurrent network for isolating deepfakes in videos," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020.

[11] Li, H. Xie, J. Li, Z. Wang, and Y. Zhang, "Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021.

[12] H. Liu, X. Li, W. Zhou, Y. Chen, H. Xue, W. Zhang, and N. Yu, "Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021.

[13] Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing face forgery detection with high-frequency features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021.

[14] Y. Zhou and S.-N. Lim, "Joint audio-visual deepfake detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 14800–14809.

[15] Chugh, P. Gupta, A. Dhall, and R. Subramanian, "Not made for each other: Audio-visual dissonance-based deepfake detection and localization," in *Proc. ACM Int. Conf. Multimedia*, 2020.

[16] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "Emotions don't lie: An audio-visual deepfake detection method using affective cues," in *Proc.