International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 10 | Oct - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

A Comprehensive Survey of Deepfake Video Detection: Methods, Datasets, Challenges, and Future Directions

Abhishek Sharma
Department of Computer Science and Engineering
Rajasthan Institute of Engineering & Technology (RIET)
Jaipur, Rajasthan, India
abhishek.abhi567@gmail.com

Mr. Amit Kumar Sharma
Assistant Professor
Department of Computer Science and Engineering
Rajasthan Institute of Engineering & Technology (RIET)
Jaipur, Rajasthan, India
amitsharmacse09@gmail.com

***Abstract*—Deepfake technology has evolved from a niche research curiosity into a mainstream tool capable of producing synthetic media indistinguishable from authentic content. This survey provides a comprehensive and structured review of deepfake video detection methods published between 2018 and 2024. We organise existing work into five thematic categories: spatial artifact detection using convolutional neural networks, temporal inconsistency detection using recurrent and hybrid architectures, attention-guided and transformer-based methods, frequency-domain analysis, and biological cue exploitation. For each category we analyse the core mechanism, representative methods, reported performance, and limitations. We also survey the major benchmark datasets—FaceForensics++, Celeb-DF, DFDC, WildDeepfake, and FaceShifter—and discuss how dataset quality and diversity affect generalisation. A consolidated comparison table covering 25 representative papers is provided. We then identify the most critical open challenges: robustness to compression and post-processing, cross-dataset generalisation, real-time detection on edge devices, zero-shot detection against unseen generation methods, and interpretability. The survey concludes with a forward-looking discussion on how diffusion-model-based deepfakes demand a fundamental rethink of current detection paradigms, and what research directions are most likely to yield practical progress.**

***Index Terms*—Deepfake detection, face forgery, deep learning, GAN, diffusion models, FaceForensics++, temporal modeling, attention mechanism, cross-dataset generalisation, survey, video forensics**

I. Introduction

A. Background and Motivation

A deepfake is a synthetic or manipulated video in which the face, voice, or body of a person has been replaced or altered using a deep learning model. The term was coined in 2017 when an anonymous user on Reddit began posting face-swap videos produced with autoencoder-based generative models. Since then, the technology has advanced at a remarkable pace. Today, a skilled user with a consumer GPU and freely available open- source software can produce a convincing face-swap video in a matter of hours. Professional-grade tools can produce results in minutes. High-end tools generate output of such quality that trained forensic experts consistently struggle to identify them as manipulated in controlled experiments [1].

The word "deepfake" itself is a portmanteau of "deep learning" and "fake". While early face manipulation techniques relied on manual compositing in video editing software, the current generation of tools is entirely automated. A user provides a source video (containing the face to be transferred) and a target video (containing the scene and background), and the system handles face detection, alignment, synthesis, and blending automatically. The barrier to entry has fallen so dramatically that deepfake creation is no longer restricted to well-funded research laboratories or film studios; it is accessible to anyone with a moderately powerful personal computer and an internet connection.

B. Societal Impact

The societal damage from deepfakes is already measurable and documented across multiple domains. In the **political domain**, fabricated videos of heads of state and political candidates have been distributed to spread disinformation during election campaigns in multiple countries. The "cheapfake" problem—using simple video editing to misrepresent politicians—predates deepfakes, but the addition of convincing face-swap and lip-sync manipulation makes the problem qualitatively more dangerous. In the **financial domain**, criminals have successfully used voice cloning and face synthesis to impersonate senior executives during video conference calls, authorising fraudulent wire transfers. Documented losses from such attacks have totalled hundreds of millions of dollars globally [13]. In the **personal domain**, non-consensual intimate deepfakes—synthetic pornographic content produced by placing a real person's face on a performer's body—constitute a severe form of image-based sexual abuse. This form of abuse has been documented in dozens of countries and disproportionately affects women. In the **legal domain**, deepfake video has been submitted as evidence in court proceedings, raising

fundamental questions about the evidentiary value of video recordings that were previously considered reliable.

These applications share a common thread: they exploit the longstanding human tendency to trust video evidence. The phrase "seeing is believing" has underpinned the credibility of video in journalism, law enforcement, and personal communication for over a century. Deepfake technology directly attacks this trust. Restoring it requires detection systems that are fast, accurate, and deployable at scale.

C. The Detection Problem

Automated detection of deepfakes is not an academic exercise—it is an infrastructure requirement for a society that depends on video as evidence, communication, and news. Social media platforms process hundreds of millions of video uploads per day; no human moderation team can review this volume. Deep learning has emerged as the only scalable approach to this detection problem.

Convolutional neural networks (CNNs) trained on large labelled datasets of real and manipulated video frames learn to identify the visual artifacts that manipulation methods leave behind. These artifacts include unnatural blending at the face boundary, inconsistencies in skin texture and lighting, abnormal blinking patterns, and periodic patterns in the frequency domain introduced by GAN upsampling operations. Temporal models extend frame-level analysis to sequences of frames, detecting inconsistencies in motion, texture change, and expression dynamics that are invisible when each frame is examined alone. More recently, attention mechanisms, vision transformers, and frequency-domain analysis have pushed detection accuracy above 96% on controlled benchmarks such as FaceForensics++ [5].

Yet benchmark accuracy does not translate reliably to real-world deployment. Models trained on one dataset consistently fail on others when the generation method or compression level differs—a phenomenon known as the generalisation gap. The emergence of diffusion-model-based face synthesis since 2022 has introduced an entirely new class of deepfakes with different artifact signatures that existing detectors were not designed to handle. Furthermore, real-world deepfakes almost always undergo significant compression before they are encountered by a detection system—through WhatsApp, Instagram, YouTube, or other platforms—and this compression erases many of the fine- grained pixel-level artifacts that current detectors rely on.

D. Scope and Contributions of This Survey

This survey provides a comprehensive, structured, and critical review of deepfake video detection methods published between 2018 and 2024. Unlike earlier surveys that focus primarily on a single detection paradigm, this work covers all five major detection categories under a unified analytical framework: spatial artifact detection, temporal inconsistency detection, attention and transformer-based methods, frequency- domain analysis, and biological cue exploitation. For each category, we analyse the core detection mechanism, representative methods, reported benchmark performance, and key limitations.

The specific contributions of this survey are as follows:

1) A structured **taxonomy of deepfake generation methods** covering autoencoder-based face swap, GAN-based methods, 3D graphics-based replacement, and the emerging class of diffusion-model-based synthesis.
2) A comparative **survey of six major benchmark datasets** including FaceForensics++, Celeb-DF, DFDC, WildDeepfake, FaceShifter, and DGM4, analysing how dataset quality and diversity affect generalisation.
3) A **consolidated comparison table** of 25 representative detection papers, organised by year, architecture, benchmark dataset, reported accuracy, and primary limitation.
4) An analysis of **six critical open challenges** that limit real-world deployment: compression robustness, cross-dataset generalisation, real-time edge deployment, zero-shot detection, adversarial robustness, and interpretability.
5) A **forward-looking discussion** of five future research directions, with particular emphasis on the implications of diffusion-model-based deepfakes for the next generation of detection systems.

This survey aims to serve both researchers entering the field and practitioners designing deployment systems, by providing not just a catalogue of methods but a clear analytical picture of what works, what fails, and why.

This survey is organised as follows. Section II introduces the taxonomy of deepfake generation methods. Section III surveys major benchmark datasets. Sections IV–VIII cover the five detection paradigms. Section IX presents the consolidated 25-paper comparison table. Section X analyses open challenges. Section XI discusses future directions. Section XII concludes.

II. Taxonomy of Deepfake Generation

Understanding how deepfakes are generated is essential for understanding what detection methods look for. We organise generation methods into four categories.

A. Autoencoder-Based Face Swap

The original deepfake technique uses two autoencoders that share a common encoder but have separate decoders. Both autoencoders are trained on faces of two different individuals. After training, the encoder from one autoencoder is paired with the decoder of the other to reconstruct one person's face with the expressions of another. The Deepfakes method in FaceForensics++ [5] uses this approach. The key visual artifact it produces is a blending boundary between the swapped face region and the original video background, which is exploited by boundary-detection methods [8].

B. GAN-Based Methods

Generative adversarial networks [2] introduced a more powerful generation paradigm. A generator network learns to produce realistic face images by competing against a

TABLE I
MAJOR DEEPFAKE DETECTION BENCHMARK DATASETS

Dataset	Yr	Videos	Meth.	Key Feature
FF++ [5]	2019	5,000	4	Multi-compression
Celeb-DF [10]	2020	6,229	1	High quality
DFDC [13]	2020	128,154	8	Largest scale
Wild-DF [24]	2020	7,314	Multi	In-the-wild
FaceShift. [35]	2020	1,000	1	High fidelity
DGM4 [36]	2023	230,000	4	Multi-modal

discriminator that tries to distinguish real from fake images. Face2Face [25] uses expression transfer via a rendering pipeline guided by 3D morphable models. NeuralTextures [26] goes further by modifying only facial texture while preserving the original face geometry, producing manipulations with no geometric boundary artifacts. StyleGAN [27] and its successors produce entirely synthetic faces of people who do not exist, with no source video required. GAN-based methods are the dominant generation paradigm in the FaceForensics++ dataset and most of the literature from 2018 to 2022.

C. 3D Graphics-Based Replacement

FaceSwap [5] uses a traditional computer graphics pipeline: a 3D face model is fitted to each frame, the face region is rendered in the target identity, and the result is composited onto the original video. This method produces distinct geometric and lighting artifacts near the face boundary and ears that CNN-based detectors find relatively easy to identify [4].

D. Diffusion-Model-Based Synthesis

Since 2022, diffusion models have challenged GANs as the state-of-the-art in image synthesis [28]. Unlike GANs, which generate images in a single forward pass, diffusion models produce images through an iterative denoising process that tends to produce more photorealistic results with fewer characteristic artifacts. DiffFace, DiffSwap, and related methods demonstrate that face-swap results produced by diffusion models can fool detectors trained on GAN-based deepfakes, because the frequency-domain fingerprints of diffusion models differ from those of GANs. This represents the most significant challenge currently facing the deepfake detection community.

III. BENCHMARK DATASETS

The quality and diversity of the training dataset is one of the most important factors determining how well a detection model will generalise. Table I summarises the major benchmark datasets used in the literature.

FaceForensics++ (FF++) [5] is the most widely used benchmark. It contains 1,000 original YouTube videos and manipulated versions produced by four methods: Deepfakes, Face2Face, FaceSwap, and NeuralTextures. Three compression levels are provided (raw, c23, c40), enabling controlled study of compression robustness. Its widespread use makes cross-paper comparison tractable, but its relatively controlled production conditions mean that models trained only on FF++ may not generalise to in-the-wild deepfakes.

Celeb-DF [10] was created specifically to expose the generalisation weakness of models trained on FF++. It contains 590 real and 5,639 fake celebrity videos produced with a refined face-swap pipeline that minimises the obvious visual artifacts present in FF++ deepfakes. Models that achieve over 95% accuracy on FF++ typically drop to 65–75% on Celeb-DF without fine-tuning, revealing that those models had learned dataset-specific artifacts rather than general deepfake features.

The DeepFake Detection Challenge (DFDC) [13] dataset is the largest publicly available benchmark, with over 128,000 video clips from more than 3,000 paid actors. It includes a wide variety of manipulation methods, lighting conditions, and compression levels, and is consequently the most difficult benchmark for current detectors. The top submission in the original challenge achieved only 65.18% AUC on the holdout test set, illustrating how much headroom remains.

WildDeepfake [24] collects face sequences from deepfake videos found in the wild on the internet rather than produced under controlled conditions. This makes it a more realistic test of deployment-scenario generalisation.

IV. SPATIAL ARTIFACT DETECTION

The earliest and still most widely used approach to deepfake detection treats each video frame independently and trains a CNN to classify it as real or fake. The CNN learns to detect spatial artifacts in the face region that are characteristic of the manipulation pipeline used to produce the fake.

A. Lightweight CNN Methods

Afchar et al. [4] proposed MesoNet, the first CNN specifically designed for deepfake detection. The model targets mesoscopic spatial features—those at a scale intermediate between individual pixels and the full face. Two variants were proposed: Meso-4, a simple four-layer CNN, and MesoInception-4, which replaces two of the four convolutional layers with inception modules to capture features at multiple spatial scales simultaneously. MesoNet is the smallest model in the deepfake detection literature (fewer than 30,000 parameters) and achieves over 95% accuracy on early face-swap deepfakes. On the more challenging FaceForensics++ benchmark under c23 compression, it achieves approximately 83%, reflecting its limited capacity for high-quality manipulations. Its primary advantage is speed: it can classify hundreds of frames per second on consumer hardware, making it suitable as a first-stage filter in a cascaded detection pipeline.

B. Large-Scale CNN Methods

Rossler et al. [5] released FaceForensics++ and conducted the first systematic comparison of CNN architectures for deepfake detection. They found that XceptionNet [6] outperformed all other tested architectures. XceptionNet uses depthwise separable convolutions throughout, factorising each standard $k \times k$ convolution into a depthwise spatial convolution followed

by a pointwise 1×1 convolution. This reduces parameters by approximately $8\times$ compared to a standard convolution while maintaining strong representational capacity. Xception-Net achieves around 95.7% accuracy on FF++ c23 and has become one of the standard baselines in the literature.

Tan and Le [20] introduced EfficientNet, which instead of manually choosing between depth, width, and resolution scaling, optimises all three simultaneously using a compound coefficient ϕ:

$$d = \alpha^{\phi}, \quad w = \beta^{\phi}, \quad r = \gamma^{\phi} \tag{1}$$

where constants α, β, γ are found by grid search. EfficientNet-B4 ($\phi = 4$) achieves around 96.1% on FF++ c23 and shows better compression robustness than XceptionNet, maintaining 85.3% accuracy at c40 compression [23].

C. Artifact-Specific Methods

Li and Lyu [8] observed that most face-swap deepfake pipelines use affine warping to align the swapped face onto the target, and that this process leaves distinct geometric artifacts at the face boundary. A CNN trained specifically on crops of the face boundary region detects these warping artifacts with high accuracy on face-swap deepfakes, though the method is less effective on texture-only manipulations like NeuralTextures.

Yang et al. [9] exploited a different artifact: 3D head pose inconsistency. Many reenactment deepfakes (including Face2Face) produce subtle errors in the rendered 3D head pose because the driving and target faces have different physical dimensions. A classifier trained on estimated 3D head pose features across consecutive frames detects these inconsistencies reliably for reenactment methods.

Dang et al. [16] showed that gradient-based attention maps (similar to Grad-CAM) can localise the specific facial regions that a CNN uses to make its classification decision. By training a model with an attention branch that forces the network to focus on manipulated regions, they improve both accuracy and interpretability. This dual-objective approach also makes the model more robust to unseen manipulation types, because it learns to look at where artifacts plausibly appear rather than memorising specific artifact patterns.

V. Temporal Inconsistency Detection

Examining video frames in isolation ignores a rich source of information: the temporal dynamics of facial expressions, head movements, and texture changes across frames. A deepfake that looks convincing in every individual frame may still exhibit unnatural temporal patterns when examined as a sequence.

A. CNN-LSTM Approaches

Guera and Delp [3] were the first to apply recurrent neural networks to deepfake detection. They extracted CNN features from individual frames and fed the resulting sequence of feature vectors to an LSTM, which learned to identify temporal inconsistencies characteristic of fake videos. The method showed clear improvement over single-frame classification on their UADFV dataset (97.1% vs 84.3%). The key insight was that while individual frames may look realistic, the LSTM can detect the subtle jitter in texture, lighting, and expression transitions that deepfake pipelines introduce when processing consecutive frames independently.

Masi et al. [14] extended this with a two-branch recurrent network. One branch processes RGB appearance features using a standard CNN-LSTM pipeline. The second branch processes optical flow between consecutive frames, capturing motion-based inconsistencies that are invisible in appearance features. The combined model outperforms single-branch approaches, particularly on reenactment deepfakes where expression motion is the primary manipulation signal.

B. Bidirectional and Attention-Augmented Temporal Models

Standard LSTMs process sequences from past to future only. Bidirectional LSTMs [21] process the sequence in both directions, allowing each time step to incorporate both past and future context. For deepfake detection, this means the model can detect inconsistencies that only become apparent when comparing a frame to both its predecessors and successors.

A further limitation of plain CNN-LSTM models is that the CNN feature extractor assigns equal importance to all spatial feature channels before passing them to the LSTM. However, not all channels are equally informative for detecting manipulation: channels encoding face-boundary features are more informative for face-swap deepfakes, while skin-texture channels are more informative for NeuralTextures. Sharma [23] proposed an Attention-CNN-LSTM architecture that inserts a channel-wise spatial attention (CWSA) module—inspired by Squeeze-and- Excitation networks [22]—between the CNN feature extractor and the bidirectional LSTM. Given feature map $\mathbf{F} \in \mathrm{R}^{H \times W \times C}$, the CWSA module computes:

$$\mathbf{z} = \frac{1}{HW} \sum_{i,j} \mathbf{F}_{ij}, \quad \mathbf{a} = \sigma(W_2\, \mathrm{ReLU}(W_1 \mathbf{z} + b_1) + b_2) \tag{2}$$

$$\tilde{\mathbf{F}} = \mathbf{F} \otimes \mathbf{a} \tag{3}$$

The attended features $\tilde{\mathbf{F}}$ are then passed to the bidirectional LSTM. This achieves 96.8% accuracy on FF++ c23, outperforming the plain CNN-LSTM by 2.6 percentage points while adding fewer than 0.3M parameters.

C. Vision Transformer and Hybrid Methods

Coccomini et al. [17] combined EfficientNet with a Vision Transformer (ViT) to capture both local spatial detail and global long-range dependencies within each frame. The EfficientNet branch captures fine-grained artifacts at the pixel level, while the ViT branch models relationships between distant facial regions (e.g. detecting inconsistencies between the eye region and the skin region that span the full width of the face). On the DFDC dataset, this combination outperforms either component alone.

Haliassos et al. [30] took a different temporal approach by exploiting lip movement irregularities. Real speech produces

highly constrained lip dynamics governed by phoneme sequences; deepfake face swap disrupts these dynamics because the swapped face was trained on a different speaker. A model trained to detect irregular lip movements achieves strong generalisation because it exploits a biological constraint rather than a generation artifact.

VI. Attention-Guided and Transformer Methods

Attention mechanisms allow a neural network to selectively weight the most informative spatial regions or feature dimensions, rather than treating all parts of the input with equal importance. Their application to deepfake detection is motivated by a fundamental observation: different manipulation methods concentrate artifacts in different facial regions, and a detector that can learn where to look will generalise better than one that processes the entire face uniformly. An autoencoder-based face swap, for example, typically leaves its strongest artifacts at the face boundary; a GAN-based texture modification spreads its artifacts across the skin surface; and a reenactment method introduces inconsistencies primarily around the mouth and eyes. Attention mechanisms provide a principled way to exploit this spatial heterogeneity.

A. Multi-Region Spatial Attention

Zhao et al. [12] proposed multi-attentional deepfake detection, the first method to apply multiple independent attention modules to different facial sub-regions simultaneously. The five modules focus respectively on the eyes, nose, mouth, forehead, and jawline. Each module produces a region-specific feature map that captures the artifact pattern characteristic of that facial zone. These maps are then aggregated using a learned textittexture enhancement module that amplifies subtle artifact signals before passing them to the final classifier. The method achieves 97.6% on FF++ and demonstrates improved robustness to compression compared to single-attention methods, because even when compression erases the artifact in one facial region, another region may retain enough information for detection. This spatial redundancy is a key advantage over global pooling approaches.

The multi-attention design also provides a form of interpretability: by inspecting which attention map activates most strongly, an analyst can identify which facial region contains the strongest manipulation evidence. This is practically valuable for forensic investigation, where the detection decision must be explainable and not just a confidence score.

B. Identity-Aware Attention

Cozzolino et al. [15] proposed ID-Reveal, which approaches deepfake detection from a fundamentally different angle: rather than searching for visual artifacts, it verifies whether the *identity* of the person in the video is consistent with what is known about them. An identity encoder is pre-trained on a large face dataset to produce an embedding that captures the expected appearance of any given individual. At test time, the model compares the identity features extracted from each frame of the query video against a set of reference embeddings extracted from authentic footage of the same person. Inconsistencies between the query embedding and the reference embeddings—measured as distances in the embedding space—flag the video as potentially manipulated.

This identity-centric approach offers two practical advantages. First, it is agnostic to the specific generation method: regardless of how the deepfake was produced, if the output does not match the genuine identity of the target person, the detector will flag it. This makes ID-Reveal significantly more robust to unseen generation methods than artifact-based detectors. Second, it is especially well-suited to detecting deepfakes of public figures—politicians, executives, and celebrities—for whom large volumes of authentic reference footage are publicly available. The primary limitation is that the method requires reference footage of the specific target identity and cannot be applied in a fully open-world setting where the target person is unknown.

C. Full Transformer Architectures

Wodajo and Atnafu [32] proposed one of the first pure Vision Transformer (ViT) architectures for deepfake detection, motivated by the observation that CNN- based models have limited receptive fields—a convolution kernel can only attend to a small neighbourhood of pixels at each layer—and therefore may miss long-range spatial dependencies between distant facial regions. A transformer, by contrast, computes self-attention globally across all image patches in a single operation, allowing it to directly model relationships between, for example, the eye region and the chin, or the left and right cheeks.

The method divides the face image into non-overlapping patches, treats each patch as a token, and computes multi-head self-attention across all tokens simultaneously. On FF++, performance is comparable to strong CNN baselines, but the transformer architecture offers better scalability to higher image resolutions and shows stronger generalisation when pre-trained on large-scale unlabelled face datasets. Coccomini et al. [17] extended this by combining EfficientNet for local spatial artifact extraction with a ViT for global context modeling, obtaining the best of both paradigms. The hybrid outperforms either component alone on the DFDC benchmark, where the combination of local and global analysis provides complementary evidence that neither approach captures individually.

A key insight from this line of work is that the choice of backbone is less important than the choice of what features the model attends to. Transformers offer more flexible attention patterns than CNNs, but they require larger training datasets to realise their potential. For datasets the size of FF++ (around 1,000 videos), CNN-based approaches often remain competitive because the inductive biases of convolution (locality and translation equivariance) are well- matched to the local nature of most deepfake artifacts.

VII. Frequency-Domain Methods

Most deepfake detection methods operate in the spatial (pixel) domain, directly processing RGB image values. An al-

ternative and complementary paradigm analyses images in the frequency domain. This approach is motivated by the observation that deep generative models—particularly those that use repeated upsampling operations to produce high- resolution output from low-resolution latent codes—leave characteristic periodic patterns in the Fourier spectrum of generated images. These spectral fingerprints are invisible to the human eye but are reliably detectable by a trained classifier.

A. Fourier Spectrum Analysis

Frank et al. [11] provided the foundational analysis of frequency-domain artifacts in GAN-generated images. They showed that upsampling operations commonly used in GAN generators—including nearest-neighbour interpolation, bilinear interpolation, and transposed convolution—introduce characteristic periodic patterns in the 2D Discrete Fourier Transform (DFT) of the generated image. Specifically, these patterns manifest as a grid of equally-spaced spectral peaks in the magnitude spectrum. The spacing of these peaks is determined by the upsampling factor: a 2× upsampling layer introduces peaks at half the Nyquist frequency, a 4× layer at quarter the Nyquist frequency, and so on. Because real camera images do not contain this regular grid structure—their spectra are dominated by the natural 1/f frequency distribution of natural scenes—a simple linear classifier trained on the Fourier magnitude spectrum achieves strong performance on GAN-generated faces.

A key finding of Frank et al. is that this spectral fingerprint generalises surprisingly well across different GAN architectures. Because the periodic artifact is a property of the upsampling operation itself (not of any specific GAN design), a classifier trained on ProGAN-generated images retains non-trivial detection performance on images from StyleGAN, BigGAN, and other architectures that use the same upsampling scheme. This cross-architecture generalisation is a significant advantage over spatial methods, which tend to be more architecture-specific.

The primary limitation, however, is sensitivity to image compression. JPEG compression and video codecs such as H.264 apply a DCT-based compression step that partially destroys the fine spectral structure. When deepfake videos are compressed—as they almost always are in practice—the spectral grid pattern is attenuated, reducing classification accuracy significantly. This is one of the most important practical limitations of the frequency-domain approach.

B. DCT and Wavelet Analysis

Qian et al. [29] proposed F3-Net (Frequency in Face Forgery detection Network), which addresses the compression sensitivity of raw Fourier analysis by operating in the DCT domain rather than the DFT domain, and by combining frequency features with spatial features through a cross-attention mechanism. The method uses two frequency- analysis branches: a *frequency-aware image decomposition* branch that learns to decompose the input image into frequency subbands in the DCT domain, and a *local frequency statistics* branch that computes statistical summaries of the frequency content in local image patches. These two branches capture artifact information at different spatial scales, and a cross-attention module aligns the complementary information from both branches before passing it to the classifier.

The spatial features extracted by a CNN are combined with the DCT features through the cross-attention mechanism, allowing the model to learn when to rely on spatial artifact evidence and when to rely on frequency evidence. F3-Net achieves state-of-the-art performance on FF++ and shows notably better cross-compression robustness than purely spatial methods, because the DCT-domain features are partially preserved even after JPEG or H.264 compression.

Wavelet-based decomposition has also been explored as an alternative to DCT. Wavelets provide a time-frequency localisation that DCT lacks: rather than decomposing the entire image uniformly, a wavelet decomposition produces representations that are localised in both space and frequency. This makes wavelet features potentially more robust to the local, spatially-concentrated artifacts that characterise face-boundary manipulations.

C. Limitations of Frequency-Domain Methods

Despite their theoretical appeal, frequency-domain methods face two fundamental challenges that limit their practical deployment. First, their effectiveness depends critically on the spectral fingerprint of the specific generation method used during training. When a new generation architecture is introduced, the spectral patterns change and detectors trained on older methods fail. This is a manifestation of the same generalisation gap that affects spatial methods, but it is particularly acute in the frequency domain because the spectral fingerprint is so directly tied to the implementation details of the upsampling pipeline.

Second, and most importantly, the emergence of diffusion-model-based deepfakes has rendered the frequency analysis of GAN-based detectors largely ineffective. Diffusion models generate images through an iterative denoising process that does not involve the repeated upsampling operations characteristic of GANs. As a result, diffusion-generated images do not exhibit the periodic spectral grid pattern that Frank et al. identified for GANs. Corvi et al. [37] demonstrated empirically that detectors trained on GAN-based deepfakes perform at or near random chance on diffusion-generated images. This represents the most significant failure mode currently facing the frequency-domain detection paradigm and motivates the development of generation-agnostic spectral features that characterise synthetic media regardless of the underlying synthesis model.

VIII. Biological Cue Methods

A fundamentally different approach to deepfake detection avoids searching for generation-specific visual artifacts altogether. Instead, it exploits the biological and physiological constraints that govern the appearance of real human faces in video. The core insight is that while a generative model

can produce a photorealistic face image, it is much harder for it to correctly model the subtle physiological signals that are present in authentic human video—signals such as the periodic colour change caused by the cardiac pulse, the precise biomechanics of blinking, or the tight coupling between speech articulators and lip movement. Biological cue methods exploit these constraints to detect deepfakes in a way that is inherently more independent of the specific generation pipeline.

A. Eye Blinking Patterns

Li et al. [7] were among the first to exploit a specific biological constraint: the pattern of eye blinking. Natural human blinking follows a well-characterised statistical pattern—blinks occur approximately 15–20 times per minute, each lasting 100–400 milliseconds, and the pattern is largely involuntary and difficult to suppress. The key observation was that most early GAN training datasets consisted of photographs or carefully curated video frames where the subject's eyes are open, because frames with closed eyes are less visually useful for training a face generator. As a result, GAN-trained face generators were implicitly biased toward producing faces with open eyes, and the generated videos exhibited unnaturally infrequent and short blinks.

Li et al. trained a Long Short-Term Memory (LSTM) network on sequences of eye-region features extracted from video frames. The LSTM learned to distinguish the temporal pattern of natural blinking from the unnaturally sparse blinking of early GAN-generated faces. The method achieved high accuracy on the UADFV dataset, the primary benchmark available at the time. The fundamental limitation is generalizability: as generation methods improved and began training on video data that includes natural blinking, the blinking artifact was corrected. Newer generation methods exhibit blinking patterns that are statistically indistinguishable from real faces, making this cue ineffective against modern deepfakes. Nevertheless, the blinking method remains historically significant as the first biological-cue approach to deepfake detection and as an illustration of the principle that artifacts introduced by training data bias can be detectable even when pixel-level artifacts are not.

B. Remote Photoplethysmography

Remote photoplethysmography (rPPG) is a non-contact method for measuring the cardiac pulse from variations in skin colour captured by a standard video camera. Each heartbeat causes a small but measurable change in the volume of blood in the superficial facial vessels, which in turn causes a subtle periodic variation in the brightness and chrominance of skin pixels. The rPPG signal is typically invisible to the naked eye but can be recovered by averaging colour values over a skin region across many frames and applying appropriate signal processing.

Ciftci et al. [31] made the key observation that this rPPG signal is absent in deepfake videos. Because the generative model processes each frame independently (or uses only a short temporal window for temporal consistency), it does not model the long-range temporal colour variation caused by the cardiac pulse. The output frames therefore contain no coherent rPPG signal: the colour variation across frames is random rather than periodic at the cardiac frequency (∼1 Hz for a resting heart rate of 60 bpm).

A detector based on rPPG estimation works as follows. A signal processing pipeline estimates the rPPG signal from a video by computing a weighted average of the R, G, and B channel values over a skin mask across frames. A spectral analysis then checks whether this signal contains a significant peak in the physiologically plausible frequency range (0.7–4 Hz). If no such peak is found, the video is flagged as fake. Ciftci et al. demonstrated that this approach achieves strong performance on several benchmarks and—crucially—generalises better across datasets than artifact-based methods, because the absence of the rPPG signal is a property of all current generative models, not just those used to produce the training data.

The primary limitation of rPPG-based detection is sensitivity to video quality and frame rate. Recovering a reliable rPPG signal requires at least 10–15 seconds of video at a frame rate of 25 fps or higher. Short clips, heavily compressed video, or video with significant head movement or illumination change can all corrupt the rPPG estimate and lead to false negatives. The method is therefore most reliable for longer video clips under controlled conditions.

C. Head Pose and Facial Geometry

Yang et al. [9] exploited a geometric inconsistency specific to reenactment deepfakes. In a face reenactment pipeline, the driving face (whose expressions and head movements are being transferred) and the target face (whose identity is being preserved) are typically different people with different face shapes, sizes, and 3D structure. The reenactment process must therefore perform a 3D transformation to map the driving face's pose and expression onto the target face's geometry. This transformation introduces subtle but detectable inconsistencies in the rendered 3D head pose.

Specifically, Yang et al. showed that the 3D head pose estimated from the swapped face region and the 3D head pose estimated from the background/context of the same frame are often inconsistent in reenactment deepfakes. In a real video, the face and its surrounding context must be geometrically consistent—the face can only be at one location in 3D space. In a reenactment deepfake, the swapped face may have a slightly different 3D orientation than what is geometrically consistent with the background, because the reenactment model does not perfectly account for the scene geometry. A classifier trained on the discrepancy between these two 3D pose estimates achieves reliable detection of reenactment deepfakes.

The head pose method has two notable advantages. First, it is robust to compression because 3D pose estimation operates on coarse geometric structure rather than fine pixel detail. Second, it is specific enough to focus on reenactment deepfakes while not generating false positives on face-swap deepfakes

(which do not involve pose transfer). The limitation is the mirror image of this specificity: the method is ineffective for face-swap and texture-based manipulation types where head pose is not the primary manipulation signal.

IX. Consolidated Comparison of Methods

Table II presents a consolidated comparison of 25 representative deepfake detection methods, covering architecture, dataset, reported accuracy, and primary limitation. All accuracy figures are reported on the primary benchmark used in the respective paper; where multiple benchmarks were tested, the FF++ c23 figure is used if available.

Several trends emerge from this comparison. First, accuracy on FF++ has largely plateaued above 95% for spatial methods and above 96% for hybrid methods, suggesting that FF++ is becoming too easy for controlled benchmarking. Second, there is a persistent and large gap between FF++ accuracy and performance on harder datasets: DFDC top performance is around 65% AUC and Celeb-DF performance for models trained on FF++ is typically 65–72%. Third, no single method dominates across all four axes of accuracy, generalisation, compression robustness, and computational efficiency; the best choice depends on the deployment scenario.

X. Open Challenges

Despite rapid progress on controlled benchmarks, deepfake detection faces a set of fundamental unsolved challenges that limit the practical deployment of current methods. This section analyses six of the most important.

A. Cross-Dataset Generalisation

The most consistently documented and serious weakness in the deepfake detection literature is the failure of trained models to generalise across datasets. A model trained on FF++ and evaluated on Celeb-DF or DFDC without fine-tuning loses 20–30 percentage points in accuracy. The best models on FF++ (above 96% accuracy) typically achieve only 65–75% on Celeb-DF and around 65% AUC on DFDC. This gap is not simply a matter of dataset difficulty: it reveals that current models learn the specific visual signature of the generation pipeline used to produce the training data, rather than universal features that characterise all synthetic faces.

The root cause of this generalisation failure is what can be called *dataset bias entanglement*: the model conflates the deepfake artifact it is supposed to detect with other statistical regularities that happen to differ between the training and test datasets. For example, the video quality, resolution, and background diversity of FF++ differ systematically from those of Celeb-DF, and a model may learn to classify based on these confounders rather than the actual manipulation signal.

Addressing cross-dataset generalisation requires one or more of the following: (a) much larger and more diverse training datasets that span multiple generation methods, quality levels, and recording conditions; (b) domain adaptation techniques such as adversarial domain invariance training [38], which explicitly penalises features that are discriminative for the source domain but not the target domain; (c) self-supervised pretraining on large unlabelled video datasets followed by fine-tuning on labelled deepfake data, which may produce more general low- level feature representations; or (d) meta-learning approaches that optimise for fast adaptation to new domains with minimal labelled examples.

B. Compression and Post-Processing Robustness

Real-world deepfake videos almost always undergo significant compression before they are seen by a detection system. Social media platforms recompress uploaded videos using H.264 or H.265 codecs at bitrates that substantially reduce the fine-grained pixel-level information in each frame. WhatsApp reduces video quality aggressively to limit data usage. YouTube re-encodes all uploaded video at standardised quality levels. This compression is not a minor perturbation: it can entirely erase the subtle blending boundary artifacts and texture inconsistencies that many detectors rely on.

Quantitative evidence from the literature makes this impact clear. On the FF++ benchmark, moving from raw (lossless) to c40 (low quality) compression reduces MesoNet accuracy from 89.0% to 61.2%, XceptionNet from 97.0% to 71.1%, and EfficientNet-B4 from 97.2% to 85.3% [23]. These are not marginal drops; they represent the difference between a usable detector and one that is barely better than random on a majority of real-world deepfakes.

Two complementary strategies for improving compression robustness have been explored in the literature. The first is *compression-aware training*: including examples at multiple compression levels in the training set, and applying random compression as a data augmentation technique during training. This teaches the model to detect artifacts that persist even after compression, at the cost of reduced accuracy on uncompressed video. The second strategy is *compression-invariant feature design*: using features—such as head pose estimates, rPPG signals, or identity embeddings—that are inherently robust to compression because they operate at a coarser level of representation than pixel-level artifacts. A fully satisfactory solution to compression robustness remains an open problem.

C. Real-Time Detection on Edge Devices

The social and regulatory case for deepfake detection requires deployment at scale: on mobile devices that individuals use to receive and share video, on browser extensions that flag suspicious content, and on social media servers that process hundreds of millions of video uploads per day. Current high-accuracy detectors (XceptionNet, EfficientNet-B4, Attention-CNN-LSTM) require a GPU with several gigabytes of VRAM for real-time inference. This is simply not available on a typical smartphone processor or in the inference infrastructure used by most social media platforms.

The trade-off between model size and accuracy is well documented. Bansal et al. [18] show that efficient CNN architectures can achieve around 95% accuracy on FF++ at real-time frame rates on consumer hardware, but a 1.5–2

TABLE II
CONSOLIDATED COMPARISON OF 25 DEEPFAKE DETECTION METHODS

Reference	Year	Architecture	Dataset	Acc. (%)	Primary Limitation
Korshunov & Marcel [1]	2018	Face recognition baseline	Custom	20.0	Designed as attack, not detector
Afchar et al. [4]	2018	MesoNet (compact CNN)	Custom	95.1	Drops to 83% on FF++ c23
Guera & Delp [3]	2018	CNN + LSTM	UADFV	97.1	Only tested on one dataset
Li et al. [7]	2018	Eye blink LSTM	UADFV	99.0	Fails on newer GAN methods
Rossler et al. [5]	2019	XceptionNet	FF++	95.7	Single frame, no temporal
Li & Lyu [8]	2019	Warp artifact CNN	FF++	92.3	Face-swap only
Yang et al. [9]	2019	Head pose LSTM	FF++	90.1	Reenactment deepfakes only
Tan & Le [20]	2019	EfficientNet-B4	FF++	96.1	Cross-dataset drops to 72%
Li et al. [10]	2020	Multiple (benchmark)	Celeb-DF	65.5	Exposed cross-dataset failure
Frank et al. [11]	2020	Frequency spectrum CNN	FF++	90.3	Fails under JPEG compression
Dang et al. [16]	2020	Attention + CNN	FF++	95.8	High compute cost
Dolhansky et al. [13]	2020	Multiple (challenge)	DFDC	65.2	Best AUC on large-scale set
Masi et al. [14]	2020	Two-branch RNN	FF++	94.3	Optical flow adds latency
Qian et al. [29]	2020	F3-Net (freq+spatial)	FF++	97.5	Complex two-branch design
Haliassos et al. [30]	2021	Lip sync LSTM	FF++	96.3	Audio-visual sync required
Zhao et al. [12]	2021	Multi-attention CNN	FF++	97.6	Purely spatial, no temporal
Cozzolino et al. [15]	2021	ID-Reveal (identity)	FF++	95.4	Requires reference frames
Ciftci et al. [31]	2021	rPPG biological cue	FF++	93.2	Sensitive to low frame rate
Wodajo & Atnafu [32]	2021	Patch transformer	FF++	94.2	High memory cost
Coccomini et al. [17]	2022	EfficientNet + ViT	DFDC	96.5	Large model size
Zheng et al. [33]	2021	Patch forensics CNN	FF++	96.1	Limited temporal modeling
Luo et al. [34]	2021	Local relation (LRL)	FF++	96.4	Complex multi-stage pipeline
Bansal et al. [18]	2023	Real-time CNN	FF++	95.1	Only tested at single compression
Akhtar [19]	2023	Survey (multi-method)	Multiple	—	Not a detection method
Sharma [23]	2024	Attention-CNN-LSTM	FF++	**96.8**	Cross-dataset drops to 70%

percentage point accuracy gap remains compared to the best-performing models. In a high-stakes detection scenario, this gap matters: a 2% false negative rate on a platform processing 100 million videos per day corresponds to 2 million missed deepfakes.

Several techniques from the model compression literature are underexplored in the deepfake detection context. Knowledge distillation trains a small student network to mimic the output distribution of a large teacher network, potentially producing a compact model that retains much of the teacher's accuracy. Post-training quantisation reduces the numerical precision of model weights from 32-bit floating point to 8-bit or even 4-bit integers, dramatically reducing memory and compute requirements with modest accuracy loss. Neural architecture search (NAS) can discover compact architectures that are better matched to the detection task than hand-designed models. All three directions deserve more systematic investigation in the deepfake detection domain.

D. Zero-Shot Detection of Unseen Generation Methods

The deepfake generation landscape changes rapidly. Between 2017 and 2024, the dominant generation paradigm shifted from autoencoder-based face swap to GAN-based synthesis to diffusion-model-based synthesis—three fundamentally different technical approaches, each producing deepfakes with different artifact signatures. Each transition rendered previously trained detectors less effective, because the new generation method did not produce the artifacts that the old detector was trained to find.

This is the *zero-shot detection problem*: when a new generation method is released, a deployed detector has no training data from it and must generalise from what it has seen before. Empirical evidence shows that current detectors do not generalise well in this setting. Corvi et al. [37] showed that detectors trained on GAN- generated deepfakes perform near random chance on diffusion- generated images. The fundamental question is whether there exist features that are universally characteristic of synthetic media regardless of how the synthesis was performed.

Biological cue methods offer the most promising direction for zero-shot generalisation, because they exploit physiological constraints (cardiac pulse, blinking, lip dynamics) that no current generation method correctly models. Identity-consistency methods such as ID- Reveal [15] are similarly generation-agnostic: they detect manipulation by comparing the claimed identity against reference footage, regardless of how the manipulation was performed. However, neither biological cues nor identity consistency cover the full range of deepfake types, and a fully general zero-shot detector does not yet exist.

E. Adversarial Attacks and Evasion

Deepfake detection models are machine learning classifiers, and as such they are vulnerable to adversarial attacks: carefully crafted perturbations, imperceptible to human observers, that cause the classifier to output an incorrect prediction with high confidence. In the deepfake detection context, this means an adversary who is both creating deepfakes and aware of the detector being used can add a small adversarial perturbation to the generated video frames that causes the detector to classify the fake as real.

Neves et al. [39] demonstrated this vulnerability concretely, showing that gradient-based adversarial attacks can reliably fool state-of-the-art deepfake detectors with perturbations that are indistinguishable from video compression noise. This is not a hypothetical concern: sophisticated actors who deploy deepfakes for malicious purposes (state-level disinformation campaigns, financial fraud) have both the motivation and the technical capability to apply such perturbations.

Adversarially robust detection requires approaches that go beyond standard empirical risk minimisation. Adversarial training—augmenting the training data with adversarial examples generated against the current model—is the most established defence, but it provides only empirical robustness guarantees and may not defend against adaptive attacks. Certified defences based on randomised smoothing provide provable robustness within a bounded perturbation radius, but at a significant cost in clean accuracy. Detection methods that operate on features that are inherently hard to perturb—such as biological signals or geometric measurements—offer an alternative path toward adversarial robustness.

F. Interpretability and Forensic Validity

Current deepfake detectors are essentially black boxes. They accept a video frame as input and produce a probability score as output. This is sufficient for automated filtering on social media platforms, where the goal is high-throughput triage. It is insufficient for forensic investigation, journalism, or legal proceedings, where the detection decision must be explainable: which specific facial region was altered, what artifact was detected, and why the detector is confident that this artifact is inconsistent with authentic content.

Post-hoc explanation methods such as Grad-CAM and LIME can produce visualisations that highlight the image regions most influential for the classification decision. These are useful starting points, but they have two significant limitations. First, they are not guaranteed to reflect the true computational mechanism by which the model reached its decision: a model can produce a plausible- looking attention map while actually basing its decision on entirely different features. Second, they are not robust to adversarial manipulation: an adversary can craft a deepfake that produces a misleading attention map while still being classified as fake by a less sophisticated detector.

A more principled approach to interpretable detection would produce pixel-level evidence of specific manipulations—for example, a segmentation mask indicating exactly which pixels were synthesised, or a difference image showing the detected boundary artifact. Such outputs would be directly useful for forensic analysis and would make the detection claim independently verifiable. Developing detection methods that produce this level of interpretable output without sacrificing detection accuracy is an important and largely open research challenge.

XI. Future Directions

The open challenges identified in the previous section point toward several specific research directions that are most likely to yield practical progress in the near term. This section discusses five of these directions in detail.

A. Diffusion-Model-Aware Detection

The emergence of diffusion-model-based face synthesis since 2022 represents the most significant paradigm shift in the field since GAN-based methods replaced autoencoder-based face swap in 2018–2019. Diffusion models generate images through an iterative denoising process: a random noise image is progressively refined over hundreds of denoising steps, guided by a learned score function, until a photorealistic image is produced. This process does not involve the repeated upsampling operations characteristic of GAN generators, and therefore does not produce the periodic spectral grid artifacts that frequency-domain detectors rely on. Furthermore, the iterative refinement process tends to produce more spatially coherent faces with fewer of the blending boundary artifacts that CNN-based spatial detectors exploit.

Early work by Corvi et al. [37] demonstrated empirically that state-of-the-art detectors trained on GAN-based deepfakes perform at or near random chance when evaluated on diffusion-generated images. This is not a marginal performance drop; it is a complete failure of generalisation. Conversely, a detector specifically trained on diffusion-generated images achieves high accuracy on diffusion content but generalises poorly to GAN content.

Several research directions are needed to address this fundamental challenge. The first is the creation of large- scale datasets of diffusion-generated deepfakes, which are necessary for training and evaluating diffusion-aware detectors. The second is the development of detection features that characterise synthetic faces regardless of the generation paradigm—candidates include artifacts in the high-frequency detail of synthetic textures, inconsistencies in the noise floor statistics of synthetic images, and subtle regularities in local patch statistics that differ between camera-captured and synthesis-generated images. The third direction is *unified detection*: a single detector that handles both GAN-based and diffusion-based deepfakes, potentially by combining generation-agnostic biological cue analysis with generation-specific artifact detection through an ensemble or mixture-of-experts design.

B. Domain Adaptation and Universal Detection

Transfer learning and domain adaptation techniques that have been successful in other areas of computer vision have not been systematically applied to deepfake detection. The

standard pipeline—pretraining on ImageNet, fine-tuning on a labelled deepfake dataset—produces models that work well within their training domain but fail when the test domain differs. More sophisticated adaptation strategies are needed.

Adversarial domain adaptation trains a feature extractor to produce representations that are discriminative for real vs. fake classification but invariant to which dataset (domain) the video came from. This is achieved by adding a domain discriminator that tries to predict the source domain from the extracted features, while the feature extractor is trained adversarially to fool the domain discriminator. The result is a feature space where the real/fake decision boundary is learned from the manipulation signal itself, not from domain-specific confounders.

Meta-learning approaches such as Model-Agnostic Meta-Learning (MAML) optimise the model parameters not for performance on a single dataset but for fast adaptation to new datasets with minimal examples. A meta-learned detector can be updated to detect a new generation method by fine-tuning on as few as five to ten labelled examples, because the meta-learning procedure has specifically optimised the initial parameters for rapid task adaptation.

Source-free domain adaptation addresses the practically important setting where the target domain data is available at test time but the source domain training data cannot be retained (due to privacy regulations or storage constraints). Methods in this category adapt the detector using only unlabelled target-domain data, without access to the original labelled training set.

C. Multimodal Audio-Visual Detection

The vast majority of existing deepfake detectors operate exclusively on the visual modality. This is a significant limitation, because many deepfake videos involve not just face manipulation but also audio manipulation (voice cloning to produce a fake voice) or audio-visual inconsistency (the face has been swapped but the original audio has been retained, creating a mismatch between lip movements and spoken phonemes).

Zhou et al. [40] showed that a detector which jointly analyses face video and audio achieves substantially better performance than a visual-only detector on the same benchmark, and is more robust to attacks that fool the visual modality alone. The audio stream provides complementary evidence: voice cloning models have their own characteristic artifacts in the frequency domain, and the temporal alignment between the audio phoneme sequence and the visual lip movement sequence is a powerful consistency signal that is difficult for a deepfake pipeline to perfectly replicate.

Future work in multimodal detection needs to address two practical challenges. First, many deepfake videos circulate without synchronised audio (image sequences, silent video, or videos with music rather than speech), so a robust detector must work effectively in unimodal settings as well. Second, audio and video undergo different compression pipelines when shared on social media, so the detector must be robust to audio compression (MP3, AAC) and video compression (H.264, H.265) independently and in combination.

D. Proactive Watermarking and Content Authentication

Passive deepfake detection—searching for artifacts in a suspicious video after the fact—is inherently reactive. By the time a deepfake is flagged by a detector, it may already have been viewed millions of times and have caused the intended harm. A proactive approach embeds verifiable authentication information into authentic content at the point of creation, enabling downstream verification that does not depend on finding artifacts.

Cryptographic content authentication systems attach a digital signature to video metadata at the time of capture. Platforms and users can verify this signature to confirm that the content was captured by a certified device and has not been altered since capture. The C2PA (Coalition for Content Provenance and Authenticity) standard is one such initiative, involving major technology companies and camera manufacturers. The limitation is that this requires widespread adoption of signing infrastructure across all cameras, editing software, and distribution platforms—a coordination challenge that is as much regulatory and commercial as it is technical.

Imperceptible digital watermarking embeds a hidden signal in the pixels of an authentic video at the time of capture or broadcast. This watermark can survive moderate compression and editing, and its presence (or absence, if the content has been manipulated in a way that disrupts the watermark) can be verified by a detector. Watermarking is complementary to passive detection: if a video lacks the expected watermark, it is either unauthenticated or has been manipulated.

E. Large-Scale Self-Supervised Pretraining

The paradigm of large-scale self-supervised pretraining followed by task-specific fine-tuning—which has driven major advances in natural language processing (BERT, GPT) and general vision (DINO, MAE)—has not been systematically explored for deepfake detection. The potential benefit is significant: a model pretrained to predict masked patches, reconstruct temporal sequences, or maintain temporal consistency across large amounts of unlabelled authentic video would develop rich representations of natural face appearance and dynamics. Fine-tuning such a pretrained model on a labelled deepfake dataset would start from a much stronger representational foundation than fine-tuning from ImageNet weights, potentially producing a detector that is more robust to distribution shift.

Specific pretraining objectives that may be relevant for deepfake detection include: *temporal consistency prediction* (predicting whether two frames could plausibly belong to the same video), *masked face reconstruction* (predicting masked portions of a face frame from visible context), and *contrastive identity learning* (learning representations where frames from the same person are close and frames from different people are far apart in embedding space). Each of these objectives encourages the model to learn aspects of natural face appearance

that deepfake generation methods fail to replicate perfectly, potentially producing pretraining representations that serve as strong initialisation for fine-tuned deepfake detectors.

XII. Conclusion

This survey has reviewed deepfake video detection methods from 2018 to 2024, covering five major paradigms: spatial CNN methods, temporal recurrent and hybrid methods, attention-guided and transformer methods, frequency-domain methods, and biological cue methods. Key findings are as follows.

Detection accuracy on the FaceForensics++ benchmark has largely plateaued above 95% for frame-level methods and above 96% for temporal models, suggesting that FF++ is approaching saturation as a challenging benchmark. EfficientNet-B4 and the proposed Attention-CNN-LSTM achieve the highest FF++ accuracies (96.1% and 96.8%) while maintaining reasonable compression robustness.

Cross-dataset generalisation remains the most significant unsolved problem. Models trained on FF++ lose 20–30 percentage points on Celeb-DF or DFDC without fine-tuning. Biological cue methods (rPPG, lip sync) and identity-aware methods show the best cross-dataset generalisation because they exploit constraints independent of the specific generation pipeline.

The emergence of diffusion-model-based deepfakes since 2022 has invalidated the frequency-domain assumptions of many existing detectors and motivates the development of generation-agnostic detection approaches. Future progress will most likely come from domain adaptation and universal feature learning, multimodal audio-visual detection, proactive watermarking infrastructure, and large-scale self-supervised pretraining—directions that move beyond fitting a classifier to a single controlled dataset toward building detection systems that remain effective as generation technology continues to evolve.

References

[1] P. Korshunov and S. Marcel, "Deepfakes: A new threat to face recognition? Assessment and detection," *arXiv:1812.08685*, 2018.
[2] I. Goodfellow *et al.*, "Generative adversarial networks," *Proc. NeurIPS*, vol. 27, 2014.
[3] D. Guera and E. J. Delp, "Deepfake video detection using recurrent neural networks," *Proc. IEEE AVSS*, 2018.
[4] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," *Proc. IEEE WIFS*, pp. 1–7, 2018.
[5] A. Rossler *et al.*, "FaceForensics++: Learning to detect manipulated facial images," *Proc. IEEE/CVF ICCV*, pp. 1–11, 2019.
[6] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *Proc. IEEE CVPR*, pp. 1251–1258, 2017.
[7] Y. Li, M. Chang, and S. Lyu, "In ictu oculi: Exposing AI generated fake face videos by detecting eye blinking," *Proc. IEEE WIFS*, 2018.
[8] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," *Proc. IEEE CVPR Workshops*, 2019.
[9] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," *Proc. IEEE ICASSP*, pp. 8261–8265, 2019.
[10] Y. Li *et al.*, "Celeb-DF: A large-scale challenging dataset for deepfake forensics," *Proc. IEEE/CVF CVPR*, pp. 3207–3216, 2020.
[11] J. Frank *et al.*, "Leveraging frequency analysis for deep fake image recognition," *Proc. ICML*, pp. 3247–3258, 2020.
[12] H. Zhao *et al.*, "Multi-attentional deepfake detection," *Proc. IEEE/CVF CVPR*, pp. 2185–2194, 2021.
[13] B. Dolhansky *et al.*, "The deepfake detection challenge (DFDC) dataset," *arXiv:2006.07397*, 2020.
[14] I. Masi *et al.*, "Two-branch recurrent network for isolating deepfakes in the wild," *Proc. ECCV*, pp. 667–684, 2020.
[15] D. Cozzolino *et al.*, "ID-Reveal: Identity-aware deepfake video detection," *Proc. IEEE/CVF ICCV*, pp. 15108–15117, 2021.
[16] H. Dang *et al.*, "On the detection of digital face manipulation," *Proc. IEEE/CVF CVPR*, pp. 5781–5790, 2020.
[17] D. A. Coccomini, N. Messina, C. Gennaro, and F. Falchi, "Combining EfficientNet and vision transformers for video deepfake detection," *Proc. ICIAP*, pp. 219–229, 2022.
[18] N. Bansal *et al.*, "Real-time advanced computational intelligence for deep fake video detection," *Applied Sciences*, vol. 13, no. 5, p. 3095, 2023.
[19] Z. Akhtar, "Deepfakes generation and detection: A short survey," *Journal of Imaging*, vol. 9, no. 1, p. 18, 2023.
[20] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," *Proc. ICML*, pp. 6105–6114, 2019.
[21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[22] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *Proc. IEEE/CVF CVPR*, pp. 7132–7141, 2018.
[23] A. Sharma, "Deepfake video detection using deep learning: A comparative study with a novel Attention-CNN-LSTM architecture," *Dept. of CSE, RIET Jaipur*, 2024.
[24] B. Zi *et al.*, "WildDeepfake: A challenging real-world dataset for deepfake detection," *Proc. ACM MM*, pp. 2382–2390, 2020.
[25] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Niessner, "Face2Face: Real-time face capture and reenactment of RGB videos," *Proc. IEEE CVPR*, pp. 2387–2395, 2016.
[26] J. Thies, M. Zollhofer, and M. Niessner, "Deferred neural rendering: Image synthesis using neural textures," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 1–12, 2019.
[27] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *Proc. IEEE/CVF CVPR*, pp. 4401–4410, 2019.
[28] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *Proc. IEEE/CVF CVPR*, pp. 10684–10695, 2022.
[29] Y. Qian *et al.*, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," *Proc. ECCV*, pp. 86–103, 2020.
[30] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalizable and robust approach to face forgery detection," *Proc. IEEE/CVF CVPR*, pp. 5039–5049, 2021.
[31] U. A. Ciftci, I. Demir, and L. Yin, "FakeCatcher: Detection of synthetic portrait videos using biological signals," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 1–13, 2020.
[32] D. Wodajo and S. Atnafu, "Deepfake video detection using convolutional vision transformer," *arXiv:2102.11126*, 2021.
[33] M. Zheng *et al.*, "Exploring temporal coherence for more general video face forgery detection," *Proc. IEEE/CVF ICCV*, pp. 15044–15054, 2021.
[34] Y. Luo *et al.*, "Generalizing face forgery detection with high-frequency features," *Proc. IEEE/CVF CVPR*, pp. 16317–16326, 2021.
[35] L. Li *et al.*, "FaceShifter: Towards high fidelity and occlusion aware face swapping," *Proc. IEEE/CVF CVPR*, pp. 37–46, 2020.
[36] S. Shao *et al.*, "DGM4: Detecting and grounding multi-modal media manipulation," *Proc. IEEE/CVF CVPR*, pp. 10526–10536, 2023.
[37] R. Corvi *et al.*, "On the detection of synthetic images generated by diffusion models," *Proc. IEEE ICASSP*, pp. 1–5, 2023.
[38] H. Shimodaira, "Improving predictive inference under covariate shift by weighting the log-likelihood function," *J. Statistical Planning and Inference*, vol. 90, no. 2, pp. 227–244, 2000.
[39] J. C. Neves, R. Tolosana, R. Vera-Rodriguez, V. Lopes, H. Proenca, and J. Fierrez, "GANprintR: Improved fakes and evaluation of the state of the art in face manipulation detection," *IEEE J. Sel. Top. Signal Process.*, vol. 14, no. 5, pp. 1038–1048, 2020.
[40] Y. Zhou, T. Chen, W. Wang, and H. Li, "Joint audio-visual deepfake detection," *Proc. IEEE/CVF ICCV*, pp. 14800–14809, 2021.