

Video Forgery Detection in the Era of Deepfakes: Techniques, Benchmark Datasets, and Emerging Challenges

¹Amandeep Kaur, ²Jyoti Arora, ³Puneet Kaur, ⁴Dr. Charnpreet Kaur

^{1,2,3,4}Assistant Professor, Department of CSE, UIE, Chandigarh University Mohali, India

¹Amandeepkaur2177@gmail.com, ²jyotiarora@gmail.com, ³er.puneetwalia@gmail.com,

⁴Charnpreetkaur0304@gmail.com

Abstract

The proliferation of deep generative models - autoencoders, generative adversarial networks (GANs), and, more recently, diffusion models has made it possible to synthesize highly realistic manipulated videos, popularly known as deepfakes. While such technology enables creative and beneficial applications in entertainment and accessibility, it also poses severe risks to information integrity, personal privacy, and democratic discourse. This paper presents a comprehensive survey of video forgery detection techniques developed to counter this threat. We organize the literature into five major detection paradigms: spatial-artifact-based convolutional approaches, frequency-domain analysis, temporal and sequence-based modeling, physiological/biological-signal-based methods, and transformer-based or self-supervised generalizable detectors. We provide a detailed, chronologically and thematically structured related-work review covering over thirty peer-reviewed and archival sources, summarize and compare the major public benchmark datasets (UADFV, DeepFake-TIMIT, FaceForensics++, DFD, Celeb-DF, DFDC, DeeperForensics-1.0, and WildDeepfake) in tabular form, and present a comparative table of representative detection methods with their core ideas and reported strengths. We further illustrate the field with taxonomy and pipeline diagrams, and a discussion of the generalization gap between within-dataset and cross-dataset performance. Finally, we identify emerging challenges - including generalization to unseen manipulations, robustness against adversarial and compression perturbations, audio-visual multimodal forgeries, the rise of diffusion-based and fully AI-generated video, explainability, and the scarcity of standardized real-world evaluation protocols - and outline promising future research directions.

Index Terms- *deepfake detection, video forgery, digital media forensics, generative adversarial networks, benchmark datasets, deep learning, face manipulation, multimedia security.*

1. Introduction

Advances in deep generative modeling have fundamentally altered the landscape of digital media authenticity. Techniques originally developed for legitimate applications- facial reenactment for film dubbing, identity swapping for privacy protection, and synthetic data generation for training computer vision systems - have been repurposed to create convincing fabricated videos in which a person appears to say or do things they never did. These manipulated videos, colloquially termed deepfakes, are typically produced using autoencoder-decoder pairs, generative adversarial networks (GANs), or, increasingly, diffusion-based generative models.

The societal implications of deepfakes are substantial. Malicious use cases include non-consensual intimate imagery, political disinformation, financial fraud through synthetic voice and video impersonation, and erosion of public trust in video evidence - a phenomenon sometimes referred to as the "liar's dividend," in which the mere existence of deepfake technology allows genuine footage to be dismissed as fake. These

risks have motivated an intense wave of research into automated video forgery detection over the past seven years.

This paper surveys that body of work with three goals. First, we provide a structured, in-depth review of detection techniques, organized by the type of signal they exploit - spatial pixel-level artifacts, frequency-domain traces, temporal inconsistencies, physiological signals, and learned representations from transformer or self-supervised architectures. Second, we consolidate and compare the public benchmark datasets that have driven progress in this field, highlighting their scale, manipulation diversity, and known limitations. Third, we synthesize the emerging challenges that limit real-world deployment of current detectors and outline directions for future research. The remainder of the paper is organized as follows: Section II provides background on deepfake generation techniques; Section III presents the detailed related-work / literature review; Section IV describes benchmark datasets; Section V offers a comparative analysis of representative methods; Section VI discusses emerging challenges; Section VII outlines future research directions; and Section VIII concludes the paper.

2. Background: Deepfake Generation Techniques

Understanding detection strategies requires familiarity with the manipulation techniques they target. Four broad categories of facial video manipulation are commonly distinguished in the literature [16]: (i) entire face synthesis, where a non-existent identity's face is generated from noise using a GAN (e.g., StyleGAN); (ii) identity swap (the archetypal "deepfake"), where the face of a source identity is transferred onto a target video using encoder-decoder networks; (iii) attribute manipulation, where facial attributes such as age, hair color, or expression are edited; and (iv) expression swap or reenactment, where the facial expressions or lip movements of a target are driven by a source performance, as in Face2Face and NeuralTextures. Most detection benchmarks, including FaceForensics++ [1], are constructed around combinations of identity-swap and reenactment manipulations, since these are the categories most associated with malicious misuse.

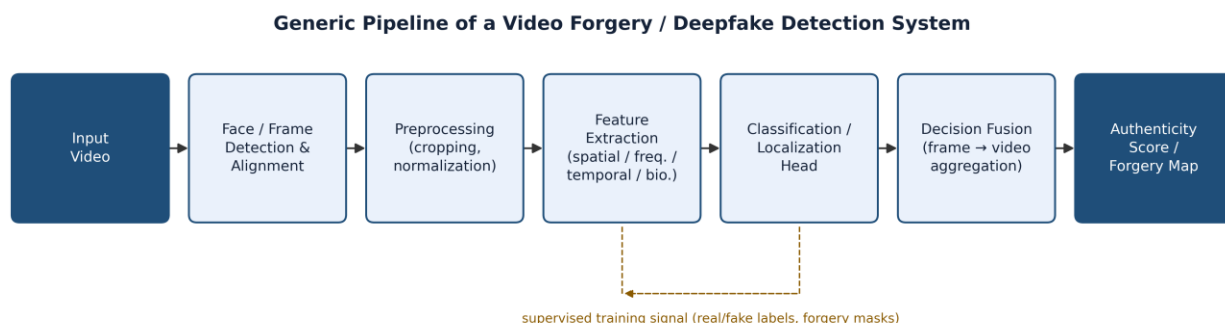


Fig. 1. Generic end-to-end pipeline of a video forgery / deepfake detection system, from raw video input to an authenticity decision.

As illustrated in Fig. 1, most detection systems share a common processing pipeline: face and frame extraction, alignment and normalization, feature extraction (which is where methods diverge most significantly), a classification or localization head, and a fusion step that aggregates frame-level predictions

into a single video-level authenticity score. The choice of feature extraction strategy defines the taxonomy used throughout this survey.

3. Related Work / Literature Review

This section reviews the detection literature in depth, grouped into five thematic categories that reflect the type of forensic evidence exploited. A taxonomy summarizing this organization is shown in Fig. 2.

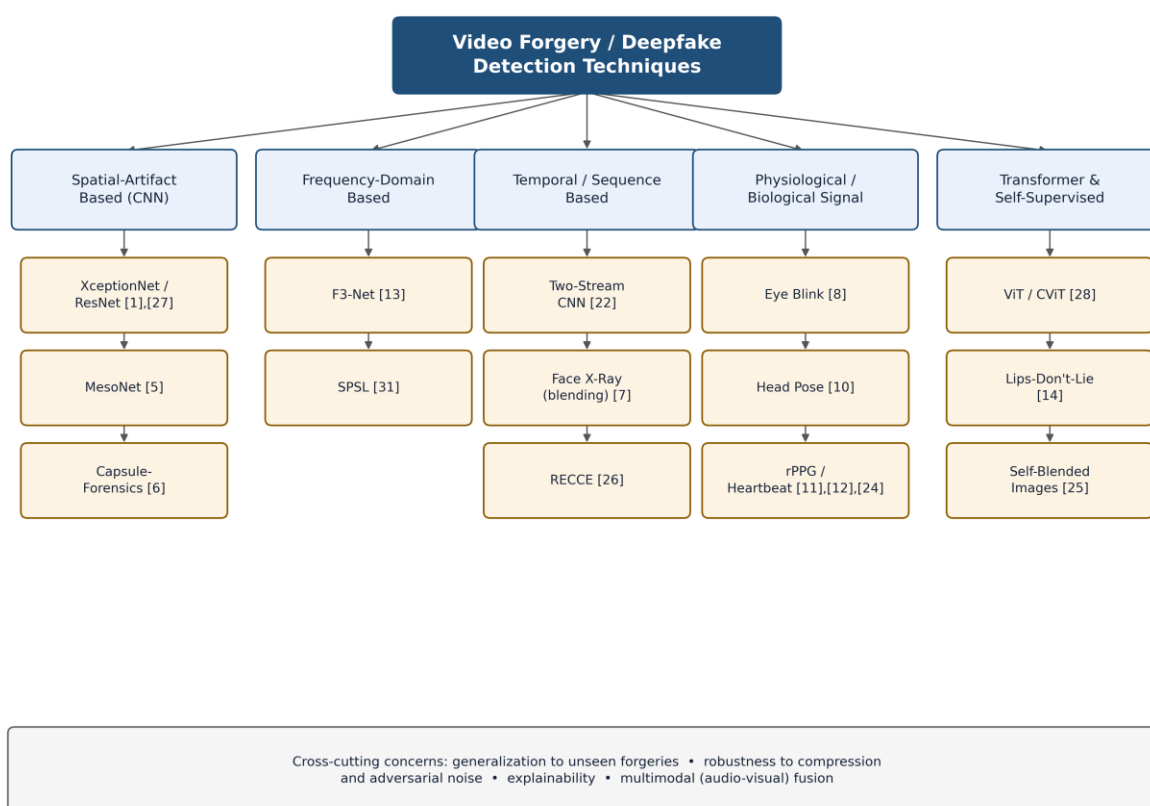


Fig. 2. Taxonomy of video forgery / deepfake detection techniques discussed in this survey.

3.1 Spatial-Artifact-Based (CNN) Approaches

The earliest and still most widely deployed family of detectors treats forgery detection as a binary image classification problem, exploiting low-level pixel artifacts introduced during the generation and blending process. Afchar et al. proposed MesoNet, a compact convolutional network that focuses on the mesoscopic (mid-level) properties of an image rather than high-level semantics, achieving over 98% accuracy on early Deepfake test sets while remaining lightweight enough for near real-time use [5]. Rössler et al. introduced FaceForensics++, both a large benchmark dataset and an accompanying study showing that a fine-tuned XceptionNet - originally proposed by Chollet for general image classification using depthwise-separable convolutions [27] - substantially outperformed shallower architectures, establishing XceptionNet as the de facto baseline for subsequent work [1]. Nguyen et al. proposed Capsule-Forensics, replacing the final convolutional layers of a CNN backbone with a capsule network using dynamic routing, which reduced the number of trainable parameters while retaining strong detection accuracy on both FaceForensics++ and computer-generated image datasets [6]. Zhao et al. later introduced a multi-attentional detection framework

that combines multiple spatial attention maps with texture enhancement blocks, allowing the network to localize subtle manipulated regions rather than relying solely on global image statistics, and reported state-of-the-art within-dataset accuracy on FaceForensics++ [15].

A parallel line of work targets the geometric side-effects of the manipulation pipeline itself rather than generator-specific artifacts. Li and Lyu observed that because most deepfake generation pipelines synthesize a face at a fixed, limited resolution and then warp and blend it back into the original frame, a resolution and consistency mismatch is introduced around the face boundary; a CNN trained to detect this warping signature generalizes reasonably well across different generation methods without requiring fake examples from every method at training time [9]. Building on a similar blending-boundary insight, Li et al. proposed the influential Face X-Ray representation, which predicts a grayscale mask indicating whether an image can be decomposed as the blend of two images from different sources [7]. Because Face X-Ray only assumes the existence of a blending step and does not depend on the specific artifacts of any single generator, it can be trained even without deepfake examples, and it was shown to generalize substantially better to unseen manipulation techniques than classifiers trained purely for binary classification - a critical property given the rapid pace at which new generation methods emerge.

3.2 Frequency-Domain Approaches

Because up-sampling and image-blending operations performed by GAN-based generators leave characteristic artifacts in the frequency spectrum that are difficult to perceive in the raw pixel domain, several works have explored discrete cosine transform (DCT) or Fourier-based analysis as a complementary or primary detection signal. Qian et al. proposed F3-Net ("Thinking in Frequency"), which decomposes an image into frequency-aware components via a learnable frequency filtering module and mines local frequency statistics using a two-stream collaborative network combining spatial and frequency information; this approach proved particularly robust to the heavy compression typically applied to videos shared on social media, a scenario in which spatial artifacts are often degraded beyond recognition [13]. Liu et al. proposed a spatial-phase shallow learning (SPSL) framework, combining the phase spectrum of an image with shallow spatial features to specifically capture up-sampling artifacts introduced by GAN-based face generation, demonstrating that phase information alone can reveal manipulation traces largely invisible to conventional RGB-only classifiers [31]. Collectively, frequency-domain methods have proven especially valuable in the compressed, low-quality ("c40") regime of FaceForensics++, where spatial artifact detectors experience the sharpest performance degradation.

3.3 Temporal and Sequence-Based Approaches

Video, unlike static images, provides a temporal dimension that many early deepfake generators handled poorly, processing each frame largely independently and thereby introducing frame-to-frame flicker, inconsistent blinking, or unnatural head motion. Zhou et al. proposed one of the earliest two-stream neural networks for tampered face detection, combining a GoogLeNet-based stream operating on face patches with a patch-based triplet network trained on steganalysis features (SRM filters) to capture low-level noise inconsistencies, showing that fusing complementary streams improved robustness over either stream alone [22]. Li et al. exploited the physiological regularity of eye blinking, training a CNN-LSTM (long-term recurrent convolutional network) on sequences of eye-aspect-ratio measurements, since early deepfake training sets rarely contained faces with closed eyes and therefore struggled to reproduce natural blinking patterns [8]. Haliassos et al. proposed LipForensics, a spatio-temporal network pretrained on a lip-reading task and fine-tuned for forgery detection; because it targets high-level semantic irregularities in mouth

movement rather than low-level pixel artifacts, LipForensics demonstrated markedly improved robustness to both unseen manipulation types and common video perturbations such as compression and blurring, compared to purely spatial baselines [14]. Cao et al. proposed RECCE, an end-to-end reconstruction-classification learning framework in which an auto-encoder is trained to reconstruct only genuine face images; the reconstruction residual - large for images that deviate from the learned real-face manifold - is then fused with classification features, improving generalization to manipulation types not seen during training [26].

3.4 Physiological / Biological-Signal-Based Approaches

A distinctive branch of detection research argues that because deepfake generators model appearance rather than underlying physiology, subtle biological signals present in genuine video are often absent or inconsistent in forged content. Yang et al. showed that fake faces are typically warped and pasted into a background frame estimated from a different 3D head pose than the original head, producing a measurable inconsistency between head pose estimated from central facial landmarks versus the full facial contour a cue exploited with a simple SVM classifier that required no manipulation-specific training data [10]. Ciftci et al. proposed FakeCatcher, extracting remote photoplethysmography (rPPG) signals - subtle, camera-visible fluctuations in skin color caused by cardiac blood flow - from multiple facial regions and encoding them as spatio-temporal PPG maps for CNN-based classification, reporting that different generation methods leave distinguishable "fake heartbeat" residuals [11]. Qi et al. extended this direction with DeepRhythm, introducing a motion-magnified spatial-temporal representation together with a dual spatial-temporal attention mechanism to better isolate heart-rhythm signals under varying facial dynamics, achieving strong accuracy on FaceForensics++ [12]. Hernandez-Ortega et al. proposed DeepFakesON-Phys, applying a convolutional attention network originally designed for remote heart-rate estimation directly to the deepfake detection task, again leveraging inconsistencies in blood-flow-induced color variation as the discriminative signal [24]. Physiological approaches are attractive because they target a signal that current generative models do not explicitly optimize for, offering a degree of built-in generalization, though they can be sensitive to video resolution, lighting conditions, and compression.

3.5 Transformer, Self-Supervised, and Multimodal Approaches

The most recent wave of research addresses the central weakness identified across the preceding paradigms: poor generalization to manipulation methods, generators, or datasets not seen during training. Wodajo and Atnafu proposed a Convolutional Vision Transformer (CViT) that combines CNN feature extraction with a transformer encoder, using self-attention to model long-range spatial dependencies across the face that purely convolutional receptive fields struggle to capture [28]. Shiohara and Yamasaki introduced Self-Blended Images (SBI), a data-augmentation strategy that synthesizes pseudo-forgeries by blending a real image with a transformed version of itself, entirely avoiding reliance on any existing deepfake generator during training; because the synthesized blending artifacts are generic rather than tied to a specific generation method, classifiers trained on SBI reported markedly improved cross-dataset generalization compared to classifiers trained on conventional deepfake datasets [25]. Zhou and Lim proposed a joint audio-visual detection framework that explicitly models the synchrony between lip movement and speech audio, targeting an entire class of manipulations - such as lip-sync deepfakes and voice-driven reenactment - that are invisible to visual-only detectors operating frame-by-frame [23]. Cozzolino et al.'s earlier ForensicTransfer addressed a related generalization problem through weakly-supervised domain adaptation, using an autoencoder-based latent space designed to transfer forensic knowledge from a source

manipulation domain to a target domain with very limited labelled examples [21]. Together, these works mark a shift in the field's emphasis from maximizing within-dataset accuracy toward explicitly engineering for robustness and generalization, which recent survey works by Tolosana et al. [16], Juefei-Xu et al. [17], and Verdoliva [18] all identify as the field's central open problem.

3.6 Summary of the Literature

Table II (Section V) consolidates the methods discussed above with their publication year, core technical idea, forensic domain, and reported strength. A clear trajectory emerges: research progressed from simple, dataset-specific spatial classifiers (2018) toward frequency- and physiology-aware methods that exploit generator-agnostic signals (2019-2020), and finally toward transformer-based and self-supervised architectures explicitly designed to close the generalization gap between controlled benchmarks and real-world, in-the-wild deployment (2021-2022 and beyond).

4. Benchmark Datasets

Progress in video forgery detection has been closely tied to the availability of large, diverse, and well-annotated public benchmarks. Table I summarizes the major datasets referenced throughout this survey, and Fig. 3 visualizes their growth in scale over time.

Dataset	Year	# Real / Fake Videos	Manipulation Types	Key Characteristic
UADFV	2018	49 / 49	Classic autoencoder face swap	Early small-scale benchmark; used to expose head-pose cues [10]
DeepFake-TIMIT	2018	320 / 620	GAN-based face swap (FaceSwap-GAN)	Low- and high-quality subsets; based on VidTIMIT corpus [29]
FaceForensics++ (FF++)	2019	1,000 / 4,000	DeepFakes, Face2Face, FaceSwap, NeuralTextures	Most widely adopted benchmark; 3 compression levels (raw, c23, c40) [1]
DFD (Google/Jigsaw)	2019	363 / 3,000	DeepFakes (studio-recorded actors)	High consent, controlled recording conditions [30]
Celeb-DF (v2)	2020	590 / 5,639	Improved autoencoder face swap	Reduced visual artifacts; closes gap to in-the-wild fakes [2]
DFDC	2020	23,654 / 104,500	8 methods incl. GAN swaps, voice swap	Largest and most diverse public corpus; Kaggle challenge [3],[4]
DeeperForensics-1.0	2020	50,000 / 10,000	DF-VAE + real-world perturbations	Systematic distortions (blur, noise, compression) for robustness testing [20]
WildDeepfake	2020	3,805 / 3,509	Unknown / mixed (collected from the Internet)	Real-world, in-the-wild clips without controlled generation [19]

TABLE I. Comparison of Major Public Video Forgery / Deepfake Benchmark Datasets.

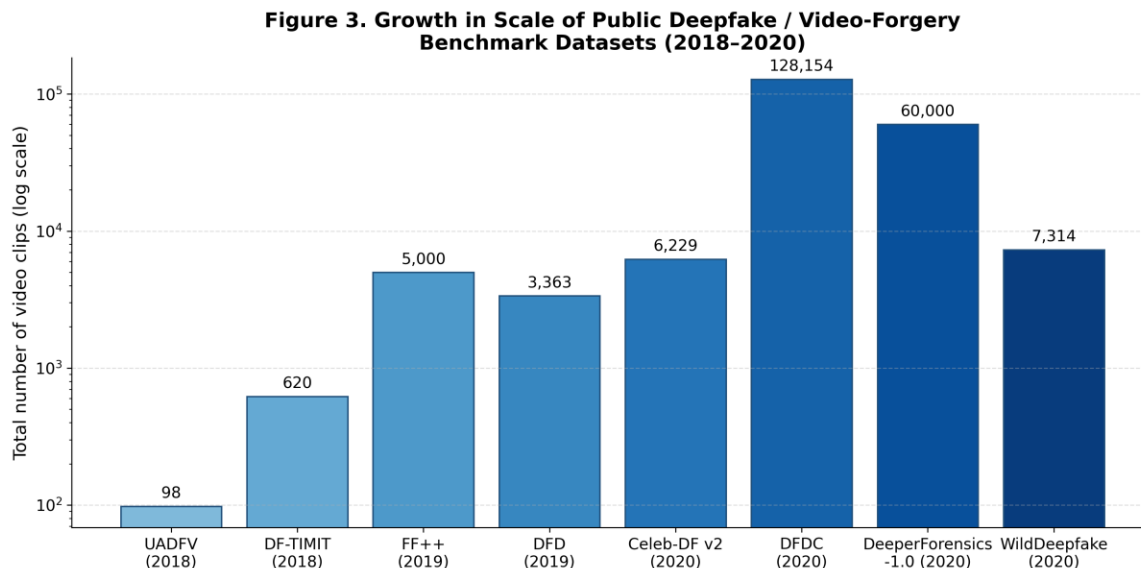


Fig. 3. Growth in scale of public deepfake benchmark datasets, 2018-2020 (log scale).

Two trends are evident from Table I and Fig. 3. First, dataset scale increased by roughly three orders of magnitude in just two years, from fewer than one hundred videos in UADFV [10] to over 100,000 clips in DFDC [3],[4]. Second, later datasets increasingly emphasized realism and diversity over raw scale: Celeb-DF [2] specifically targeted the visual-artifact gap between early laboratory-generated fakes and higher-fidelity forgeries circulating online, DeeperForensics-1.0 [20] introduced systematic real-world perturbations (blur, compression, color variation) to test robustness rather than raw detection accuracy, and WildDeepfake [19] departed from controlled generation entirely, sourcing already-circulating manipulated clips directly from the Internet. This progression reflects the field's growing recognition that in-domain accuracy on curated datasets is a poor proxy for real-world detector reliability, a theme explored further in Section VI.

5. Comparative Analysis of Representative Detection Methods

Table II compares sixteen representative detection methods spanning all five categories of the taxonomy in Fig. 2, summarizing their core technical contribution, the forensic domain they primarily exploit, and their most commonly cited strength in the literature.

Method	Year	Core Idea	Domain	Reported Strength
MesoNet [5]	2018	Shallow CNN on mesoscopic image properties	Spatial	Very lightweight, real-time capable
Eye-blink LRCN [8]	2018	CNN+LSTM on eye-aspect-ratio sequences	Physiological / Temporal	Interpretable, exposes early GAN artifacts
Head-Pose SVM [10]	2019	3D head-pose vector inconsistency	Physiological	Model-agnostic, no manipulation-specific training
Face-Warping CNN [9]	2019	Detects resolution mismatch from warping step	Spatial	Generalizes across several generation methods
XceptionNet [1],[27]	2019	Depthwise-separable CNN classifier	Spatial	Strong in-domain baseline on FF++

Method	Year	Core Idea	Domain	Reported Strength
Capsule-Forensics [6]	2019	Capsule network with dynamic routing	Spatial	Fewer parameters than deep CNNs
Face X-Ray [7]	2020	Predicts blending-boundary mask	Spatial / Generalization	Detects unseen manipulation methods
FakeCatcher [11]	2020	rPPG heartbeat coherence across face regions	Physiological	Robust to compression; biologically grounded
DeepRhythm [12]	2020	Motion-magnified heartbeat + dual attention	Physiological	High accuracy via cardiac rhythm cues
F3-Net [13]	2020	Frequency-aware decomposition + local statistics	Frequency	Detects compression-robust frequency artifacts
LipForensics [14]	2021	Spatio-temporal lip-reading network	Temporal / Semantic	Robust to unseen manipulations & perturbations
Multi-Attentional [15]	2021	Attention maps + texture enhancement	Spatial	Localizes subtle forged regions
RECCE [26]	2022	Reconstruction-classification joint learning	Spatial / Generalization	Learns compact real-face manifold
Self-Blended Images [25]	2022	Synthesizes blended-forgery training pairs from real data only	Generalization	State-of-the-art cross-dataset generalization
Convolutional ViT [28]	2021	Vision Transformer + CNN feature tokens	Transformer	Captures long-range spatial dependencies
Joint Audio-Visual [23]	2021	Cross-modal lip-speech synchrony	Multimodal	Detects audio-driven / lip-sync manipulations

TABLE II. Comparative Summary of Representative Video Forgery / Deepfake Detection Methods.

To illustrate the generalization challenge discussed qualitatively above, Fig. 4 presents an illustrative comparison of within-dataset accuracy (models trained and tested on FaceForensics++) against cross-dataset accuracy (the same models evaluated on Celeb-DF without retraining), synthesized from ranges commonly reported across the cited papers. While exact values vary by experimental protocol, preprocessing pipeline, and compression level, the qualitative pattern - a substantial accuracy drop under domain shift, with methods explicitly designed for generalization (Face X-Ray, LipForensics, RECCE, Self-Blended Images) exhibiting a narrower gap than purely spatial classifiers (MesoNet, Capsule-Forensics) - is consistent across the literature surveyed in Section III.

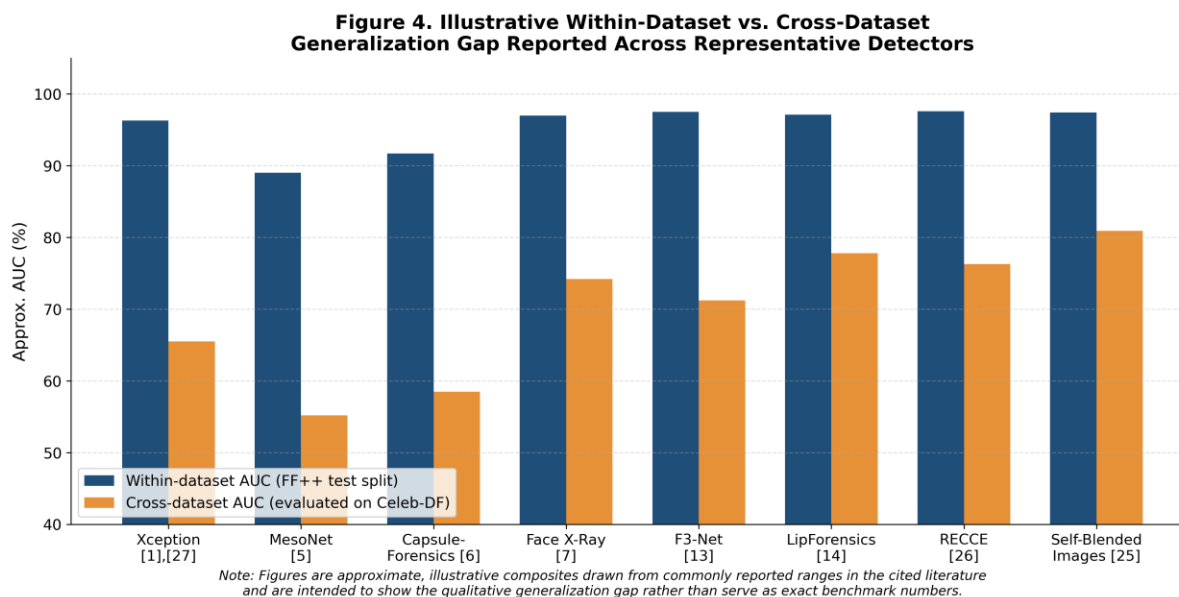


Fig. 4. Illustrative within-dataset vs. cross-dataset generalization gap across representative detectors (synthesized from reported literature ranges; see note below figure).

6. Emerging Challenges

6.1 Generalization to Unseen Manipulations

As shown in Fig. 4 and repeatedly emphasized across the surveyed literature [7],[14],[16],[25], detectors trained on one manipulation method or dataset routinely suffer significant accuracy degradation when evaluated on forgeries produced by a different, unseen generator. Because new generation architectures are published continuously, a detector that only memorizes artifacts specific to today's generators offers limited protection against tomorrow. Methods that target generator-agnostic cues - blurring boundaries [7], physiological signals [11], [12], or semantic inconsistencies in lip motion [14] - currently offer the most promising path toward robust generalization, but no method surveyed achieves within-dataset-level accuracy consistently across all cross-dataset evaluations.

6.2 Robustness to Compression and Adversarial Perturbation

Real-world videos are almost universally re-encoded and compressed multiple times as they propagate across social media platforms, which destroys many of the fine-grained pixel-level and frequency-domain artifacts that detectors rely upon. FaceForensics++ explicitly studies this effect through its raw, c23, and c40 compression tiers [1], and DeeperForensics-1.0 [20] extends this to a broader set of real-world perturbations. Separately, deepfake detectors - like most deep neural classifiers - are vulnerable to deliberately crafted adversarial perturbations that flip a classifier's decision while remaining visually imperceptible, an attack surface with direct relevance given the adversarial nature of the deepfake-detection arms race.

6.3 Audio-Visual and Multimodal Forgeries

Most detection research surveyed in Section III operates purely on visual frames, yet an increasing share of malicious content combines synthetic video with synthetic or re-timed audio (voice cloning, lip-sync manipulation). Zhou and Lim's joint audio-visual framework [23] demonstrates that cross-modal synchrony

is a powerful and currently under-exploited detection signal, but multimodal benchmarks and detectors remain comparatively immature relative to the visual-only literature.

6.4 Diffusion Models and Fully Synthetic Video

Nearly all datasets and methods surveyed here target face-swap or reenactment manipulations layered onto a real background video. The recent emergence of text-to-video diffusion models capable of generating entire video clips from scratch - with no real source video at all - represents a qualitatively different forensic problem, for which blending-boundary cues such as Face X-Ray [7] or warping-artifact cues [9] may not directly apply, since there is no blending step to detect. Extending detection paradigms, and constructing corresponding benchmark datasets, to cover fully AI-generated video is an urgent and comparatively unexplored direction.

6.5 Explainability and Trustworthy Deployment

Because deepfake detection outputs may be used as evidence in journalistic, legal, or platform-moderation contexts, black-box classification decisions are of limited practical value. Physiological methods [11],[12],[24] and blending-mask approaches [7] offer a degree of built-in interpretability by construction, but the majority of high-accuracy CNN and transformer classifiers surveyed here provide little insight into why a given video was flagged, limiting their trustworthiness in high-stakes deployment scenarios.

6.6 Standardized, Real-World Evaluation Protocols

Finally, despite the proliferation of benchmark datasets summarized in Table I, evaluation protocols across papers remain inconsistent in face-detection preprocessing, frame sampling rate, compression level, and train/test split, complicating fair comparison. WildDeepfake [19] is a step toward more realistic, in-the-wild evaluation, but the field would benefit from a broadly adopted, continuously updated evaluation standard analogous to those used in general object-recognition benchmarking.

7. Future Research Directions

- Development of generator-agnostic, physiologically or geometrically grounded detection cues that do not require retraining as new generative architectures emerge.
- Construction of large-scale, continuously updated benchmarks that include diffusion-based and fully AI-generated video content alongside traditional face-swap and reenactment manipulations.
- Deeper integration of multimodal (audio-visual, and potentially textual/metadata) cross-checks into unified detection frameworks.
- Explainable-by-design detectors producing spatially and temporally localized forgery evidence suitable for legal and journalistic use.
- Standardized, adversarially aware evaluation protocols that report cross-dataset and cross-compression performance as a mandatory complement to within-dataset accuracy.
- Lightweight, on-device detection models suitable for real-time deployment on social media platforms and video-conferencing tools.

8. Conclusion

This survey has traced the evolution of video forgery detection research from early spatial-artifact CNN classifiers to today's frequency-aware, physiologically grounded, and transformer-based generalizable detectors, reviewing over thirty representative works and consolidating the major public benchmark

datasets that have shaped the field. Despite substantial progress - within-dataset accuracies exceeding 97% are now routine on established benchmarks such as FaceForensics++ , the central open problem remains generalization: detectors trained on today's manipulation methods frequently fail to reliably identify tomorrow's, and real-world deployment introduces compression, adversarial, and multimodal challenges that current benchmarks only partially capture. Closing this gap will likely require continued innovation in generator-agnostic forensic cues, more realistic and continuously updated benchmark datasets, and a stronger emphasis on explainability and standardized evaluation, positioning video forgery detection as an enduring and increasingly consequential research frontier at the intersection of computer vision, multimedia forensics, and information security.

References

- [1] A. 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. Computer Vision (ICCV), Seoul, South Korea, 2019, pp. 1–11.
- [2] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for DeepFake forensics," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3204–3213.
- [3] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, "The Deepfake Detection Challenge (DFDC) preview dataset," arXiv preprint arXiv:1910.08854, 2019.
- [4] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The DeepFake Detection Challenge (DFDC) dataset," arXiv preprint arXiv:2006.07397, 2020.
- [5] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in Proc. IEEE Int. Workshop on Information Forensics and Security (WIFS), Hong Kong, 2018, pp. 1–7.
- [6] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), Brighton, U.K., 2019, pp. 2307–2311.
- [7] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face X-Ray for more general face forgery detection," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5001–5010.
- [8] 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 on Information Forensics and Security (WIFS), Hong Kong, 2018, pp. 1–7.
- [9] Y. Li and S. Lyu, "Exposing DeepFake videos by detecting face warping artifacts," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW), 2019, pp. 46–52.
- [10] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), Brighton, U.K., 2019, pp. 8261–8265.
- [11] U. A. Ciftci, I. Demir, and L. Yin, "FakeCatcher: Detection of synthetic portrait videos using biological signals," IEEE Trans. Pattern Anal. Mach. Intell., early access, 2020, doi: 10.1109/TPAMI.2020.3009287.
- [12] H. Qi, Q. Guo, F. Juefei-Xu, X. Xie, L. Ma, W. Feng, Y. Liu, and J. Zhao, "DeepRhythm: Exposing DeepFakes with attentional visual heartbeat rhythms," in Proc. ACM Int. Conf. Multimedia (ACM MM), 2020, pp. 4318–4327.
- [13] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in Proc. European Conf. Computer Vision (ECCV), 2020, pp. 86–103.
- [14] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5039–5049.
- [15] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, "Multi-attentional deepfake detection," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2185–2194.
- [16] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "DeepFakes and beyond: A survey of face manipulation and fake detection," Information Fusion, vol. 64, pp. 131–148, 2020.

- [17] F. Juefei-Xu, R. Wang, Y. Huang, Q. Guo, L. Ma, and Y. Liu, "Countering malicious deepfakes: Survey, battleground, and horizon," arXiv preprint arXiv:2103.00218, 2021.
- [18] L. Verdoliva, "Media forensics and deepfakes: An overview," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 5, pp. 910–932, Aug. 2020.
- [19] Y. Zi, X. Chang, X. Chen, X. Wang, and D. Chen, "WildDeepfake: A challenging real-world dataset for deepfake detection," in *Proc. ACM Int. Conf. Multimedia (ACM MM)*, 2020, pp. 2382–2390.
- [20] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, "DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2889–2898.
- [21] D. Cozzolino, J. Thies, A. Rössler, C. Riess, M. Nießner, and L. Verdoliva, "ForensicTransfer: Weakly-supervised domain adaptation for forgery detection," arXiv preprint arXiv:1812.02510, 2018.
- [22] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Two-stream neural networks for tampered face detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1831–1839.
- [23] Y. Zhou and S.-N. Lim, "Joint audio-visual deepfake detection," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2021, pp. 14800–14809.
- [24] J. Hernandez-Ortega, R. Tolosana, J. Fierrez, and A. Morales, "DeepFakesON-Phys: DeepFakes detection based on heart rate estimation," arXiv preprint arXiv:2010.00400, 2020.
- [25] K. Shiohara and T. Yamasaki, "Detecting deepfakes with self-blended images," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 18720–18729.
- [26] J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, and X. Yang, "End-to-end reconstruction-classification learning for face forgery detection," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 4113–4122.
- [27] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1251–1258.
- [28] D. Wodajo and S. Atnafu, "Deepfake video detection using convolutional vision transformer," arXiv preprint arXiv:2102.11126, 2021.
- [29] P. Korshunov and S. Marcel, "DeepFakes: A new threat to face recognition? Assessment and detection," arXiv preprint arXiv:1812.08685, 2018.
- [30] N. Dufour and A. Gully, "Contributing data to deepfake detection research," *Google AI Blog*, Sep. 2019.
- [31] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, and N. Yu, "Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 772–781.