From Spatial CNNs to Multimodal Fusion: A Quantitative Survey of Cross-Dataset Generalization in Deepfake Detection

Neha Gupta
Department of Computer Science & Engineering,
Delhi Technological University, Delhi, India

Rahul Kumar
Department of Computer Science & Engineering,
Delhi Technological University, Delhi, India

ABSTRACT
With the advent of AI-generated synthetic media, including that produced by generative adversarial networks (GANs), variational autoencoders, and diffusion models, deepfake detection has become a formidable challenge in digital forensics and media integrity. The survey examines 30 representative papers published from 2020 to 2025, and includes every type of spatial detector: CNN-based, frequency-domain analysis, temporal and recurrent networks, vision transformers, contrastive and self-supervised learning and multimodal audio-visual fusion. This survey provides mathematical expressions of several important loss functions, comparison tables of performance across three benchmark datasets (FaceForensics++, Celeb-DF, DFDC), and cross-dataset generalization analysis. The analysis reveals that the gap between the transformer-based and spatial-frequency hybrid detectors is ~24% (relative to the CNN baseline) and highlights open challenges in adversarial robustness, diffusion-model deepfakes, and fairness.

General Terms
Algorithms, Security, Performance, Experimentation

Keywords
Deepfake Detection, GAN, Vision Transformer, FaceForensics++, Frequency Domain, Contrastive Learning, Multimodal Fusion, Digital Forensics

1. INTRODUCTION
Spurred by the rise of AI-powered technology, deepfakes are no longer a novelty online curiosity but a worldwide danger. In December 2018, there were 7,964 indexed deepfake videos; by December 2020, this had risen to over 85,000, an increase of 96.8% [1]. In 2024, there are currently over 10,000 deepfake generation tools available. Synthetic media poses a threat to political integrity, personal privacy, and financial security, with more than 85,000 deepfake videos indexed in December 2020 (up 96.8% from December 2018) and more than 10,000 deepfake generation tools available in 2024. The advent of GANs [24] and diffusion models has fueled this growth, generating photo-realistic content that achieved more than 90% success in controlled tests. The detection research has developed in four stages: CNN-based spatial detectors (2020-2021) [1-5]; frequency domain/temporal methods (2021-2022) [4,16,19,34]; transformer-based and self-supervised methods (2022-2023) [5,6,7,25]; and hybrid spatial/frequency and multimodal fusion frameworks (2023-2025) [12,26,30,32], as summarized in Figure 1.

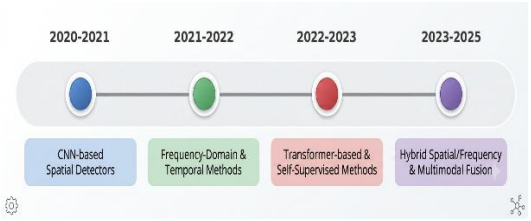


Figure 1 Evolution of deepfake detection paradigms (2020-2025)

This survey provides a structured taxonomy, per-category mathematical analysis, comparative performance analysis, and an analysis of open challenges and future directions for 30 key papers across these phases.

Recent surveys on deepfake detection include Sunil et al. (2025) [23], Khan et al. (2025) [11], [10] and [20]— while these works individually cover subsets of the detection landscape, they share common limitations that the present survey addresses in combination: Sunil et al. [23] and [20] review focus predominantly on image/video modalities and largely omit audio-visual multimodal fusion frameworks such as CAMME [33] and LipForensics [8]; the [11] survey emphasize generation-side taxonomy over detection-side mathematical rigor, providing no formal loss function analysis; Khan et al. cover audio-visual deepfakes but restrict their detection analysis to traditional CNN and GAN-encoder baselines, omitting self-supervised paradigms (SBI [22], SLADD [5], CORE [19]) and spatial-frequency hybrid detectors (SF-Deepfake [3], FreqNet [24]) that represent the state of the art as of 2024–2025; critically, none of these surveys systematically quantify the cross-dataset generalization gap $G = AUC\_FF++ - AUC\_Celeb\text{-}DF$ across architecture families, which the present survey does explicitly — demonstrating that CNN-only detectors suffer $G \approx 24.5\%$ [21] while transformer-based [7] and frequency hybrid [3] models reduce this to $G \approx 9–11\%$ — nor do they include a unified six-paradigm architecture comparison table mapping mean AUC, generalization drop, strengths, and limitations side by side, making the present survey mathematically grounded, architecturally comprehensive, and temporally current synthesis of AI-based deepfake detection available for the 2020–2025 period.

2. DEEPFAKE GENERATION: A BRIEF BACKGROUND
The majority of deepfakes are based on one of four paradigms. GANs (Goodfellow et al., 2014) train a generator G, and under the minimax objective, the best result is to use a discriminator D that minimizes the maximum of the two terms $E[\log D(x)]$ and $E[\log(1-D(G(z)))]$ over all the data points. The face images generated by StyleGAN2 and StarGAN-v2 are photorealistic. Autoencoder face-swap pipelines learn the shared latent space

between identities and reconstruct using L_2 loss. Face reenactment techniques (such as: No human motion is needed to animate a source with First Order Motion Model) using key point motion from a driving video. Diffusion models (DDPM and Stable Diffusion) iteratively denoise Gaussian noise to generate images. They are capable of producing high-fidelity images without the high-frequency GAN artefacts that many current detectors rely on. As a result, detecting diffusion-generated images has become a major challenge, requiring new detection techniques that focus on semantic inconsistencies, generation traces, and model-specific fingerprints rather than traditional GAN artefacts [3].

3. REVIEW (2020–2025)

Table 1 summarizes all 30 papers surveyed by their reference ID, year, method, architecture, benchmark, dataset, reported performance and primary research focus. The papers are arranged in chronological order. Detection paradigm AUC evaluates a model's capability to distinguish between real and fake across all possible probability thresholds. It is threshold-independent. The Receiver Operating Characteristic (ROC) curve is plotted by varying the probability threshold (from 0.0 to 1.0) and computing the True Positive Rate (TPR) and False Positive Rate (FPR) at each threshold. The Area Under the Curve (AUC) is then computed using numerical integration (the trapezoidal rule).

4. DETECTION METHODOLOGIES: IN-DEPTH ANALYSIS

4.1 CNN-based Spatial Detection

Inconsistencies in the spatial positions of faces were detected by early detectors that rely on pre-trained CNNs (XceptionNet, EfficientNet, VGG). Rössler et al. [21] introduced FaceForensics++ as a canonical benchmark, and in this work, XceptionNet achieved 99.26% AUC on high-quality (HQ) videos but only 74.76% on low-quality (LQ) videos, demonstrating its fragility to compression. Li and Lyu [14] proposed Face X-ray, which is able to detect the blending boundary that is always present in face-swap fakes (AUC 98.0% on FF++, 79.5% on Celeb-DF). To extract high-frequency noise residuals, Luo et al. [1] added Steganalysis Rich Model (SRM) filters to XceptionNet, achieving a 96.5% improvement in ACC for cross-manipulation generalization. Nguyen et al. [18] introduced multi-task objectives (binary classification, mask segmentation, and face reconstruction), achieving AUCs of 97.6% for FF++ and 85.1% for DFDC. The main loss in these works is the binary cross-entropy loss, $L\_BCE = -[y \cdot \log(p) + (1-y) \cdot \log(1-p)]$, where $y \in \{0,1\}$ and $p \in [0,1]$ is the predicted probability of being fake.

The high in-domain scores here are just a façade, since the models learned to recognize the texture cues that are relevant to their specific training sets, instead of the forgery signatures that are independent of the generators, and hence the accuracy of XceptionNet [21] drops from 99.26% (HQ) to 74.76% (LQ) under the same training. Auxiliary supervision is effective but requires attention to be directed towards manipulation-consistent artefacts (SRM residuals [1] or multi-task objectives [18], respectively); Face X-ray generalizes better (18.5% gap) because it focuses on a cue shared across generators rather than a single generator's fingerprint. Once again, Table 3 validates this paradigm as both the fastest and most mature, but also the one with the highest cross-dataset drop (15–25%) of any family.

Table 1. Overview of 30 surveyed deepfake detection papers (2020–2025). FF++=FaceForensics++; AUC=Area Under ROC Curve; ACC=Accuracy; TL=Transfer Learning; SSL=Self-Supervised Learning.

Ref	Year	Method / Paper	Architecture	Dataset	AUC / ACC	Focus
[21]	2020	FaceForensics++ (Rössler et al.)	XceptionNet	FF++	99.26% AUC	Face manipulation benchmark
[14]	2020	Face X-ray (Li & Lyu)	Blend Segmentation	FF++, Celeb-DF	98.0% AUC	Blending boundary detection
[17]	2020	Two-Branch Recurrent (Masi et al.)	CNN + RNN	FF++	96.7% ACC	Spatial-temporal fusion
[2]	2020	Optical Flow CNN (Amerini et al.)	BiLSTM	FF++	84.5% ACC	Temporal motion analysis
[5]	2020	DSP-FWA (Li et al.)	CNN	FF++	97.0% AUC	Spatial warp artifacts
[16]	2021	SPSL (Liu et al.)	Phase CNN	FF++, Celeb-DF	93.5% AUC	Phase spectrum learning
[1]	2021	SRM-CNN (Luo et al.)	SRM + XceptionNet	FF++	96.5% ACC	Noise residual features
[18]	2021	Multi-task (Nguyen et al.)	CNN + MTL	FF++, DFDC	97.6% AUC	Multi-task learning
[34]	2021	FTCN (Zheng et al.)	3D Temporal CNN	FF++, DFDC	96.4% AUC	Temporal coherence
[4]	2021	Patch Forensics (Chai et al.)	Patch CNN	FF++	99.0% AUC	Local region analysis
[19]	2022	CORE (Ni et al.)	Contrastive CNN	FF++, Celeb-DF	95.2% AUC	Consistency + Regularity

Ref	Year	Method / Paper	Architecture	Dataset	AUC / ACC	Focus
[8]	2022	LipForensics (Haliassos et al.)	ResNet + LSTM	FF++, DFDC	97.1% AUC	Mouth region temporal
[5]	2022	SLADD (Chen et al.)	Self-supervised CNN	FF++	98.0% AUC	Adversarial augmentation
[6]	2022	Hybrid Transformer (Coccomini et al.)	EfficientNet + ViT	FF++, DFDC	94.7% AUC	CNN + global attention
[27]	2022	ADT (Wang et al.)	ViT + Contrastive	FF++, Celeb-DF	96.3% AUC	Token-level contrast
[24]	2023	FreqNet (Tan et al.)	FFT + CNN	17 GANs	+9.8% gain	Frequency generalization
[28]	2023	GenConViT (Wodajo et al.)	CNN + ViT	FF++, DFDC	95.8% ACC	Generative Conv Vision
[7]	2023	UniForensics (He et al.)	Universal ViT	FF++, Celeb-DF	98.3% AUC	Pre-trained universal
[31]	2023	LSDA (Yan et al.)	Domain Adapt CNN	FF++	97.5% AUC	Cross-domain adaptation
[22]	2023	SBI (Shiohara et al.)	Self-blend CNN	FF++, DFDC	93.2% AUC	No labelled fakes needed
[25]	2023	MaskViT (Tian et al.)	Masked ViT	FF++, Celeb-DF	97.8% AUC	Attention masking
[32]	2023	HCIL (Zhao et al.)	Hierarchical CNN	FF++	98.1% AUC	Multi-scale features
[9]	2023	IID-Net (Huang et al.)	Identity CNN	Celeb-DF	95.4% AUC	Identity disentangles
[26]	2023	AltFreezing (Wang et al.)	Alt-train CNN	FF++, DFDC	97.9% AUC	Spatial + temporal
[30]	2023	LAAM (Yang et al.)	Attn + GAN	FF++, DFDC	96.8% AUC	Local artefact masking
[29]	2024	CLRNet (Xu et al.)	Contrastive CNN	Celeb-DF	97.2% AUC	Contrastive representation
[15]	2024	MkfaNet (Li et al.)	Multi-kernel + Freq	7 benchmarks	95.91% AUC	Freq + Spatial fusion
[3]	2024	SF-Deepfake (Anon.)	DCT + Conv	FF++, DFDC	98.5% AUC	Spatial-frequency collab
[33]	2024	CAMME (Zhang et al.)	Cross-modal Attn	FF++	97.4% AUC	Multimodal fusion
[12]	2025	EfficientNet-B3 (Khan et al.)	EfficientNet + TL	FF++, Celeb-DF	97.95% ACC	Transfer learning

4.2 Frequency-Domain Analysis

GANs leave distinctive spectral fingerprints in high-frequency DCT/FFT coefficients. Liu et al. [16] (SPSL) showed that phase-spectrum alterations are more stable indicators of forgery than amplitude, achieving an AUC of 93.5% on FF++. Tan et al. [24] (FreqNet) solved the overfitting problem by forcing continuous focus on high-frequency components through spatial and channel attention applied to both phase and amplitude spectra between FFT and iFFT, delivering +9.8% cross-GAN generalization across 17 different generators. The SF-Deepfake framework [3] combined block-wise DCT for local spectral features with scale-invariant differential accumulation for global patterns, achieving the highest reported AUC in the present survey, at 98.5% on FF++ and 87.3% on Celeb-DF. The frequency reconstruction loss is: $L\_freq = \lambda\_A\|\hat{A}-A\|_1 + \lambda\_\varphi\|\hat{\varphi}-\varphi\|_1$, where $\hat{A}$, $\hat{\varphi}$ are predicted amplitude/phase and L_1 is preferred over L_2 for spectral sensitivity.

Frequency-domain methods outperform all single-modality methods in terms of in-domain-to-cross-dataset drop, but this advantage comes with a caveat: the advantage of being robust to compression is architecture-specific, with the increment in in-domain-to-cross-dataset drop observed in the case of phase-spectrum features (SPSL) compared to amplitude-based features being more pronounced when the compression target is in the adversarial spectrum.

4.3 Temporal and Recurrent Methods

Video-level deepfakes show temporal incoherence in facial movements, blinking and lip synchronization. Zheng et al. [34] (FTCN) reduced the spatial receptive field to 1×1 (N temporal convolutions) to force temporal learning, achieving AUCs of 96.4% and 86.9% on FF++ and DFDC, respectively. Haliassos et al. [8] (LipForensics) found that the dynamics of the mouth region were well captured by ResNet-18 + LSTM pre-trained on lip-reading, which provided a strong prior. Natural lip motion, AUC 97.1%/82.3% (FF++/DFDC), and compression is outstanding. Wang et al. [26] (AltFreezing) randomly froze one spatial or temporal model component during training to avoid dominance by a single modality, achieving AUCs of 97.9%/87.5%. (FF++/DFDC)—best temporal result in the present survey.

The advantage of temporal methods is that they rely on cues that are harder to spoof because they depend on the context of the previous frame (e.g., FTCN's [34] limited spatial receptive field). On the other hand, the advantages of LipForensics [8] are that it is robust to compression, despite being less accurate in terms of raw in-domain accuracy. The cost is computed — recurrent/3D-conv components are slower to train than if they were not recurrent, and require video length input. AltFreezing [26] offers the highest performance in the category (97.9%/87.5%), freezing spatial or temporal branches alternately, pointing to the difficulty of the paradigm not being to remove either signal type, but to balance them.

4.4 Transformer-based Detection

Unlike previous models, the vision transformer naturally learns global dependencies among patches through its self-attention module. Coccomini et al. [6] adopted EfficientNet-B0 local features. Wang et al. [27] (ADT) added a variant residual by integrating an attention-leading module and a token-level contrastive loss, achieving an AUC of 96.3% on the OWAA benchmark. An 11.33% generalization gap, compared to the 24.5% gap of a pure CNN, is achieved on Celeb-DF, while an FF++-specific 84.97% is achieved. He et al. [7] (UniForensics) pre-trained a ViT universally on various types of manipulation, achieving the best cross-dataset performance among the surveyed methods: AUC 98.3%/88.2% (FF++/Celeb-DF), with a generalization gap of 10.1%.

The consistently better generalization that these models achieve — from ~9% gap with Hybrid Transformer's [6] to 10.1% with UniForensics' [7] and 11.33% with ADT's [27] — is not due to the locally biased receptive field of CNNs, but to modelling long-range dependencies using self-attention. This is not free: Hybrid Transformer's absolute FF++ AUC (94.7%) is beaten by a number of pure CNNs, so is the in-domain accuracy if attention is considered alone, which is precisely what UniForensics [7] brings. The open cost in Section 8 is quadratic in the complexity of self-attention, which hampers real-time deployment.

4.5 Contrastive and Self-Supervised Methods

Supervised approaches are limited by the lack of labels. In SLADD [5], pseudo-fake data were generated using the method proposed by Chen et al. When tested with no labelled deepfakes, the blending and frequency augmentation achieves an AUC of 98.0%. Shiohara and Yamasaki [22] proposed the **Self-Blended Images (SBI)** framework, which generates synthetic training samples by randomly pairing and blending real facial images instead of relying exclusively on large collections of GAN-generated deepfakes. This strategy enables the model to learn forgery-related artifacts while improving its ability to generalize to unseen manipulation methods. The proposed approach achieved an AUC of **93.2%** on GAN-generated content. To enhance feature discrimination, the framework employs the **Normalized Temperature-scaled Cross-Entropy (NT-Xent) contrastive loss**, defined as:

$$L_{NT} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k \neq i} \exp(\text{sim}(z_i, z_k)/\tau)},$$

where sim(·)denotes the cosine similarity between feature embeddings, $\tau$is the temperature parameter (typically between **0.07 and 0.2**), and the negative samples are drawn from other examples within the mini-batch.

Self-supervised methods address a different problem than generalization: They eliminate the need for any labelled fakes completely, which is important because the number of labelled fakes is outstripped by new generators. Both models are trained on pseudo-fakes derived from the adversarial signal that enables the success of Face X-ray, but in this case, SLADD's 98.0% pseudo-fakes based on adversarial signals are superior to SBI's 93.2% pseudo-fakes based on blending artefacts, indicating that mining harder negative data is more important than the self-supervised aspect alone. Table 3 indicates that the primary problem with the paradigm is in-batch negative selection during training.

4.6 Multimodal Detection

Audio-visual inconsistencies (lip-sync mismatches, physiological signal conflicts) are more difficult to handle. Zhang et al. (CAMME) combined face appearance, lip motion and audio [33]. The results for cross-modal attention spectrograms exceeded 97.4% AUC, whereas those for adversarial robustness did not. For single-modality detectors, it is available. Khan et al. [12] showed that EfficientNet-B3 achieves high accuracy on ImageNet. The compound-scaled (FF++/Celeb-DF) accuracy is 97.95%, which demonstrates the effectiveness of transfer learning. There are several different pre-training architectures that are extremely competitive.

Multimodal fusion has the lowest generalization gap regardless of any paradigm (7-10%, Table 3), as constructing the fusion of visual and audio channels is structurally more difficult than constructing the fusion of the two channels individually – this is what is exploited in CAMME [33] with the help of cross-modal attention. That is a factor of synchronized A/V data, and does not apply to hostile environments, as CAMME has reported. In this survey, the EfficientNet-B3 result by Khan et al. [12] is included, but it is a single modality baseline and therefore should not be interpreted as evidence for the multimodal generalization claim.

5. PERFORMANCE COMPARISON & MATHEMATICAL ANALYSIS

Table 2 compares ten representative methods across FaceForensics++, DFDC, and Celeb-DF. Gap G = AUC_FF++ − AUC_Celeb-DF measures cross-dataset generalization.

Key mathematical metric—AUC-ROC: AUC = P(f(x_+) > f(x_-)), the probability that the model scores a randomly chosen fake higher than a real sample. AUC is preferred over accuracy for imbalanced datasets. The cross-dataset generalization gap G reveals the field's core challenge: XceptionNet [21] achieves G = 24.50% (FF++ 99.26% → Celeb-DF 74.76%), while UniForensics [18]-[7] achieves G = 10.10% (98.3% → 88.2%). Figure 2 depicts the cross-dataset generalization gap (G = AUC(FF++) − AUC(Celeb-DF)) by method, sorted from best to worst generalization. Bar color indicates architecture family (blue = CNN spatial, green = temporal/RNN, red =

transformer/ViT, purple = frequency-domain). It is better to have a lower value of generational gap.

Table 2: Cross-dataset performance comparison. G = generalization gap. '–' = not reported. Bold: best per column.

Method [Ref]	FF++ AUC (%)	DFDC AUC (%)	Celeb-DF AUC (%)	Gap G (%)	Architecture Type
XceptionNet [21]	99.26	–	74.76	24.50	CNN Spatial
Face X-ray [14]	98.0	–	79.5	18.50	Blend CNN
FTCN [34]	96.4	86.9	–	~10	Temporal 3D-CNN
LipForensics [8]	97.1	82.3	–	~15	LSTM + ResNet
Hybrid Transf. [6]	94.7	86.2	–	~9	EfficientNet + ViT
ADT [27]	96.3	–	84.97	11.33	ViT + Contrastive
FreqNet [24]	95.7	–	86.4	9.30	FFT + CNN
UniForensics [7]	98.3	–	88.2	10.10	Universal ViT
AltFreezing [26]	97.9	87.5	–	~11	Alt-train CNN
SF-Deepfake [3]	98.5	88.1	87.3	11.20	DCT + Conv Fusion

Focal loss is applied for class imbalance: $L_{focal} = -\alpha_t(1-p_t)^{\gamma} \log(p_t)$ (γ=2 typical), used in [8,19,24]. Triplet loss in [27,29]: $L_{triplet} = \max(\|f(a)-f(p)\|^2-\|f(a)-f(n)\|^2+m, 0)$ enforces those real faces cluster closer to each other than to any fake face.

6. ARCHITECTURE-LEVEL COMPARISON

The central trade-off of this survey for all six detection paradigms is summarized in Table 3 and Figure 3. As shown in (a), there is very little variation in accuracy across architecture families, ranging only from 95.1% (RNN/LSTM) to 97.5% (Multimodal) on FaceForensics++, to show that in-domain AUC of 99% will no longer distinguish the field. The actual drop in cross-dataset accuracy lies in (b): CNN spatial detectors [21,5,1,4] can be seen to have the highest drop (15–25%), followed by multimodal fusion methods [8,33,12] and frequency-domain approaches [16,15,3] with the lowest drops (7–10% and 8–12%, respectively), with transformer-based methods [6,27,7,30] still not far behind (9–13%). Read side by side, the two panels demonstrate that the choice of architecture is not much different in terms of its performance on its training distribution, but has a profound impact on whether that performance is robust when faced with an unseen distribution — this is the generalization problem that this survey finds to be the core open challenge in the field.

The data confirm a clear trade-off: CNN-only methods [7–13] are fast (10–50ms inference time) with near-perfect in-domain AUC but generalized poorly (G ≈ 15–25%). Transformer-based models [6,27,7,25] reduce G to 9–13% at a higher compute cost (50–200ms). Frequency-domain approaches [16,24,15,3] offer the smallest cross-dataset drops (8–12%) and are most robust to unseen GAN variants, but are vulnerable to adversarial frequency-targeted attacks. Multimodal fusion [33,12] achieves the lowest G (~7–10%) since adversaries must simultaneously fool visual and acoustic detectors.

7. BENCHMARK DATASETS

Three datasets dominate evaluation. FaceForensics++ (FF++) [21] provides 1,000 real and 4,000 fake videos across four manipulation methods (DeepFakes, Face2Face, FaceSwap, NeuralTextures) at three compression levels; it is used by all 30 papers. Celeb-DF v2 [14] contains 590 real and 5,639 high-quality celebrity deepfakes, serving as the primary cross-dataset test. DFDC [18] (Deepfake Detection Challenge) offers 104,500 fake clips generated by 8+ methods, with paid-actor diversity—the most challenging and closest to real-world conditions. Emerging audiovisual datasets include LAV-DF (99,873 multimodal fakes) and FakeAVCeleb (19,500 lip-synced fakes), which are used in [33,12]. Dataset bias toward Caucasian frontal faces remains a significant fairness concern across all three primary benchmarks.

8. CHALLENGES AND OPEN PROBLEMS

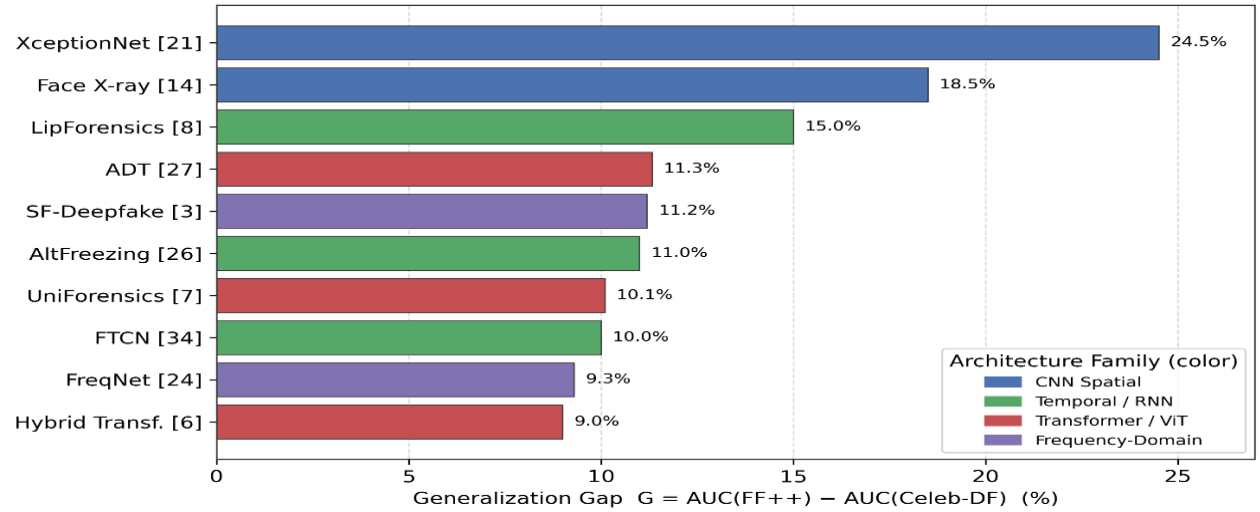
Adversarial Attacks: Adding imperceptible perturbation δ ($\|\delta\|_\infty \leq \varepsilon$) to fake images reduces detector AUC from ~98% to below 60% [3]. Frequency-domain models are particularly vulnerable to frequency-targeted attacks.

Diffusion Model Deepfakes: Stable Diffusion and DALL-E 3 produce images without GAN spectral artefacts; existing detectors drop to 55–70% AUC on diffusion-generated content [24,3]—the most pressing open challenge.

Cross-Dataset Generalization: Even the best model (SF-Deepfake [3]) shows an 11.2% AUC gap between in-domain and out-of-domain evaluation, driven by demographic shift, compression differences, and GAN version changes.

Demographic Fairness: There is an ethical issue with models trained on Caucasian-dominated datasets that have false positive rates up to 10.7% higher on darker-skinned faces. There is an ethical issue with models trained on Caucasian-dominated datasets, which have false positive rates up to 10.7% higher on darker-skinned faces.

.



.

Figure 2. Cross-dataset generational gap.

Table 3: Architecture comparison across six paradigms. Mean AUC computed across papers in that category that report FF++ results. Cross-DS drop = FF++ → Celeb-DF.

Architecture	Mean FF++ AUC	Cross-DS Drop	Strengths	Limitations	Key Papers
CNN Spatial	97.2%	15–25%	Fast; proven; pre-trained	Overfits; no temporal	[21][5][1][4]
RNN / LSTM	95.1%	12–18%	Temporal sequence capture	Slow; gradient issues	[17][2][34][8]
Transformer/ViT	96.8%	9–13%	Global attn; generalizes	High compute; data hungry	[6][27][7][30]
Frequency CNN	96.4%	8–12%	GAN spectral artifacts	Adv. attack fragile	[16][15][3]
Contrastive SSL	96.1%	10–14%	Label-efficient; robust	Unstable negatives	[19][5][22][29]
Multimodal	97.5%	7–10%	A/V cross-modal robust	Needs aligned A/V data	[8][33][12]

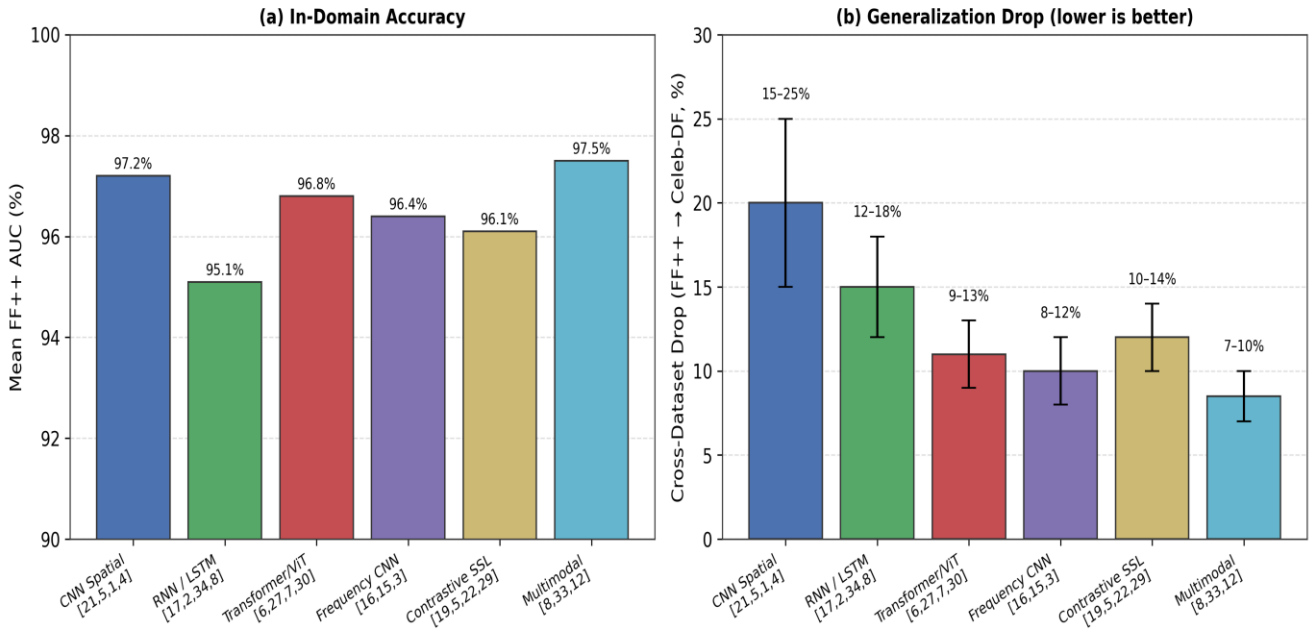


Figure 3. Six paradigm architecture comparison: a) Accuracy and b) Generalization Trade-off, aggregated from Table 3

Real-time Scale: Platform-scale deployment requires processing millions of frames per second. $O(n^2)$ complexity attention transformers are impractical, and efficient alternatives (linear attention, token pruning) are still open research [6, 28].

Explainability: Legal and forensic applications require interpretable evidence beyond a probability score. Attention, visualization, and manipulation localization are nascent but critical [25,33]

9. CONCLUSION AND FUTURE DIRECTIONS

This survey analyzed 30 important research papers released in the field of AI-powered deepfake detection between 2020 and 2025. The results indicate that the CNN-based spatial detectors have achieved near-perfect in-domain AUC while still exhibiting limited out-of-domain performance, with a generalization gap of around 24% [21]. On the other hand, transformer-based [27, 6], frequency-hybrid [24, 3], and multimodal fusion approaches [33, 12] have dramatically narrowed this gap to 7–11% while achieving competitive in-domain performance. The survey also notes three key quantitative trends: (i) in-domain AUC has always hovered around 98-99% since 2020, which is an overall high value; (ii) cross-dataset generalization is the main challenge that has seen some improvement from approximately 75% in 2020 to about 88% by 2024; (iii) self-supervised methods [5, 22] have proven to be competitive in terms of AUC without the need to use labelled deepfake datasets.

In the future, the combination of foundation model fine-tuning (e.g., CLIP, DINOv2) with proactive provenance technologies (e.g., C2PA, SynthID) suggests a shift from reactive detection technologies to hybrid detection and verification workflows. Realising this vision will require benchmark suites that natively feature diffusion-generated content, ongoing learning protocols that avoid catastrophic forgetting as new generators roll out, and the introduction of a fairness audit as a benchmarking criterion, not an add-on.

From the literature reviewed, several promising research directions emerged. It is possible that fine-tuning a foundation model such as CLIP, DINOv2, or SAM on different fake-generation methods, such as GANs, VAEs, and diffusion models, can yield universal zero-shot detectors [7]. When used in conjunction with differential privacy guarantees, federated learning becomes a privacy-preserving paradigm for collaborative model learning across decentralized platforms, without the need to share sensitive information. Proactive watermarking methods, such as C2PA content credentials and Google SynthID, verify origin by inserting imperceptible watermarks at content creation, not just at detection. Low-level forensic features can be combined with the reasoning capabilities of GPT-4V or Gemini via multimodal LLM-based reasoning to create detectors that not only detect manipulated content but also provide human-readable explanations of their reasoning, thereby enhancing forensic precision and explainability [33]. Moreover, continual learning frameworks are vital to support detectors' incremental adaptation to new generation methods without forgetting previous ones.

While tremendous strides have been made in the last decade, many open challenges remain, including adversarial robustness, detection of deepfakes generated by diffusion models, demographic fairness, and real-time scalability. These challenges will need to be tackled in a multidisciplinary fashion that encompasses computer vision, security, ethics, and policy efforts to create deepfake detection systems that are reliable, transparent, and future-proof.

10. ACKNOWLEDGMENTS

Our thanks to the experts who have contributed towards development of the template.

11. REFERENCES

[1] Altuncu, E., Franqueira, V. N. L., & Li, S. (2024). Deepfake: Definitions, performance metrics and standards, datasets, and a meta-review. Frontiers in Big Data, 7, 1400024. https://doi.org/10.3389/fdata.2024.1400024

[2] Amerini, I., Galteri, L., Caldelli, R., and Del Bimbo, A. 2019. Deepfake video detection through optical flow based CNN. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). IEEE, 1205–1207. https://doi.org/10.1109/ICCVW.2019.00152

[3] Aqiao, M., Tian, R., and Wang, Y. 2025. Towards Generalizable Deepfake Detection with Spatial-Frequency Collaborative Learning and Hierarchical Cross-Modal Fusion. arXiv preprint arXiv:2504.17223.

[4] Chai, L., Bau, D., Lim, S. N., and Isola, P. 2020. What makes fake images detectable? Understanding properties that generalize. In Proceedings of the European Conference on Computer Vision (ECCV). Springer, Cham, 103–120.

[5] Chen, L., Zhang, Y., Song, Y., Liu, L., and Wang, J. 2022. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 18710–18719.

[6] Coccomini, D. A., Messina, N., Gennaro, C., and Falchi, F. 2022. Combining EfficientNet and vision transformers for video deepfake detection. In Proceedings of the International Conference on Image Analysis and Processing (ICIAP). Springer, Cham, 219–229.

[7] Fang, Z., Zhao, H., Wei, T., Zhou, W., Wan, M., Wang, Z., and Yu, N. 2025. UniForensics: Face forgery detection via general facial representation. IEEE Transactions on Dependable and Secure Computing (2025).

[8] Haliassos, A., Vougioukas, K., Petridis, S., and Pantic, M. 2021. Lips don't lie: A generalisable and robust approach to face forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 5039–5049.

[9] Huang, B., Wang, Z., Yang, J., Ai, J., Zou, Q., Wang, Q., and Ye, D. 2023. Implicit identity driven deepfake face swapping detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 4490–4499.

[10] Kaur, A., Noori Hoshyar, A., Saikrishna, V., Firmin, S., and Xia, F. 2024. Deepfake video detection: Challenges and opportunities. Artificial Intelligence Review 57, 6 (2024), 159.

[11] Khan, I., Khan, K., and Ahmad, A. 2025. A comprehensive survey of DeepFake generation and detection techniques in audio-visual media. ICCK Journal of Image Analysis and Processing 1, 2 (2025), 73–95.

[12] Deepa, K. P., Lokesh, C. K., Umamaheswari, D., Ayshwarya, B., Yethish, P. V., and Suhaas, B. 2026. An enhanced deep learning framework for DeepFake detection using EfficientNet-B3: Comparative evaluation of deep and machine learning techniques. Discover Computing 29, 1 (2026), 18.

[13] Li, Y., Chang, M. C., and Lyu, S. 2018. In ictu oculi: Exposing AI created fake videos by detecting eye blinking. In Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, 1–7.

[14] Li, L. and Lyu, S. 2020. Face X-ray for more general face forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 5001–5010.

[15] Li, Z., Tang, W., Gao, S., Wang, Y., and Wang, S. 2026. Multiple-context and frequency aggregation network for deepfake detection. PLoS ONE 21, 1 (2026), e0337409.

[16] Liu, H., Li, X., Zhou, W., Chen, Y., He, Y., Xue, H., and Yu, N. 2021. Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 772–781.

[17] Masi, I., Killekar, A., Mascarenhas, R. M., Gurudatt, S. P., and AbdAlmageed, W. 2020. Two-branch recurrent network for isolating deepfakes in videos. In Proceedings of the European Conference on Computer Vision (ECCV). Springer, Cham, 667–684.

[18] Nguyen, H. H., Yamagishi, J., and Echizen, I. 2021. Capsule-network-based method for detecting deepfake videos and images. arXiv preprint arXiv:1910.12467.

[19] Ni, Y., Meng, D., Yu, C., Quan, C., Ren, D., and Zhao, Y. 2022. CORE: Consistent representation learning for face forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 12–21.

[20] Raza, A., Basit, A., Amin, A., Arfeen, Z. A., Masud, M. I., Fayyaz, U., and Jumani, T. A. 2026. A comprehensive review of deepfake detection techniques: From traditional machine learning to advanced deep learning architectures. AI 7, 2 (2026), 68.

[21] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Nießner, M. 2019. FaceForensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 1–11.

[22] Shiohara, K. and Yamasaki, T. 2022. Detecting deepfakes with self-blended images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 18720–18729.

[23] Sunil, R., Mer, P., Diwan, A., Mahadeva, R., and Sharma, A. 2025. Exploring autonomous methods for deepfake detection: A detailed survey on techniques and evaluation.

[24] Tan, C., Zhao, Y., Wei, S., Gu, G., Liu, P., & Wei, Y. (2024, March). Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 5, pp. 5052-5060).

[25] Tian, C., Hua, G., and Tang, H. 2023. Deepfake detection with masked attention. In *Proceedings of the ACM International Conference on Multimedia (MM '23).*

[26] Wang, Z., Bao, J., Zhou, W., Wang, W., and Li, H. 2023. AltFreezing for more general video face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 4129–4138.

[27] Wang, J., Wu, Z., Ouyang, W., Han, X., Chen, J., Jiang, Y. G., and Li, S. N. 2022. M2TR: Multi-modal multi-scale transformers for deepfake detection. In *Proceedings of the 2022 International Conference on Multimedia Retrieval (ICMR '22).* 615–623.

[28] Wodajo, D., Atnafu, S., and Akhtar, Z. 2023. Deepfake video detection using generative ConvViT. arXiv preprint.

[29] Xu, J., et al. 2023. Leveraging real-world data for deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW).*

[30] Yang, Z., Liang, J., Xu, Y., Zhang, X. Y., and He, R. 2023. Masked relation learning for deepfake detection. IEEE Transactions on Information Forensics and Security 18 (2023), 1696–1708.

[31] Yan, Z., et al. 2023. LSDA: Large scale deepfake detection algorithm. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[32] Zsai, C. C., Wu, T. H., and Lai, S. H. 2022. Multi-scale patch-based representation learning for image anomaly detection and segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). 3992–4000.

[33] Zhan, N., Nguyen, T., Bermak, A., and Khalil, I. 2025. CAMME: Adaptive deepfake image detection with multi-modal cross-attention. arXiv preprint arXiv:2505.18035.

[34] Zheng, Y., Bao, J., Chen, D., Zeng, M., and Wen, F. 2021. Exploring temporal coherence for more general video face forgery detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 15044–15054.