IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154151

Multimodal Deepfake Detection Using Deep Learning

Dhanushree B V[1], Panchami M Hegde[2]

Department of MCA, BIT, K.R. Road, V.V. Pura, Bangalore, India[1]

Assistant Professor, Department of MCA, BIT, K.R. Road, V.V. Pura, Bangalore, India[2]

Abstract: Deepfake technology has evolved rapidly, enabling the creation of highly realistic manipulated images, audio, and videos. While these advancements have applications in entertainment and media, they also pose significant risks such as misinformation, identity fraud, and security threats. This research focuses on multimodal deepfake detection using deep learning techniques by combining visual and audio features for improved accuracy. The proposed approach integrates Convolutional Neural Networks (CNNs) for image analysis and Natural Language Processing (NLP) and audio-based models for detecting inconsistencies across modalities. By leveraging multimodal data, the system enhances detection robustness compared to unimodal approaches. Experimental results demonstrate that combining visual and audio cues significantly improves detection performance and generalization across different types of deepfakes. This system can be applied in social media monitoring, digital forensics, and cybersecurity applications.

Keywords: Deepfake Detection, Multimodal Learning, Machine Learning, CNN, NLP, Audio-Visual Analysis, Artificial Intelligence

I. INTRODUCTION

Deepfake technology has emerged as a powerful application of artificial intelligence, enabling the generation of highly realistic synthetic media, including images, audio, and videos. With the advancement of deep learning techniques such as Generative Adversarial Networks (GANs) and autoencoders, it has become increasingly difficult to distinguish between authentic and manipulated content. While deepfake technology offers benefits in areas like entertainment, education, and virtual reality, it also introduces serious challenges related to misinformation, identity theft, and digital security.

The widespread use of social media platforms has accelerated the dissemination of manipulated content, making it essential to develop reliable and efficient detection systems. Traditional deepfake detection methods primarily rely on visual features such as facial inconsistencies, texture artifacts, and pixel-level anomalies. However, these unimodal approaches often fail to generalize across different datasets and are vulnerable to advanced deepfake generation techniques.

To address these limitations, recent research has shifted toward **multimodal deepfake detection**, which integrates multiple sources of information such as visual, audio, and textual data. By analyzing inconsistencies between modalities—such as lip-sync errors, speech mismatches, and semantic inconsistencies—multimodal systems provide a more robust and accurate detection framework.

This research focuses on developing a multimodal deepfake detection system using deep learning techniques. The proposed approach combines image-based analysis using Convolutional Neural Networks (CNNs) with audio and textual feature extraction methods. The objective is to enhance detection accuracy, improve generalization across diverse datasets, and provide a scalable solution for real-world applications such as social media monitoring, digital forensics, and cybersecurity.

II. LITERATURE REVIEW

Deepfake detection has evolved significantly over the past few years, transitioning from traditional image-based methods to advanced multimodal deep learning approaches. Early research primarily focused on detecting visual artifacts in manipulated images and videos. One of the foundational works, *FaceForensics++ (Rossler et al., 2019)*, introduced a large-scale dataset for facial manipulation detection and utilized Convolutional Neural Networks (CNNs) such as XceptionNet to classify real and fake images. While effective, these methods struggled to generalize across unseen manipulation techniques.

Subsequent research explored physiological and temporal features to improve robustness. *Cozzolino et al. (2021)* proposed detecting deepfakes using Facial Action Units (AUs), capturing subtle facial muscle movements that are difficult for generative models to replicate. This approach improved interpretability but depended heavily on high-quality facial data.With the advancement of transformer architectures, *Wang et al. (2023)* introduced self-supervised Vision

DOI: 10.17148/IJARCCE.2026.154151

Transformers (ViT) using Masked Autoencoding and contrastive learning. These models reduced dependency on labeled datasets and improved cross-dataset generalization. However, they require high computational resources and are less suitable for real-time applications.

Recent research has shifted toward **multimodal approaches**, combining visual and audio information for improved accuracy. *Feng et al. (2023)* utilized contrastive learning to analyze audio-visual synchronization, identifying inconsistencies between lip movements and speech. Similarly, *Ilyas et al. (2023)* proposed AVFakeNet using Dense Swin Transformers for joint audio-visual analysis, achieving strong performance but at the cost of increased computational complexity.

Further advancements include ensemble and fusion-based methods such as *AVTENet (Hashmi et al., 2023)* and *AVFF (Oorloff et al., 2024)*, which leverage multiple models and cross-modal feature fusion to enhance detection accuracy. These approaches demonstrate that combining modalities significantly improves performance compared to unimodal systems.

More recently, the integration of vision-language models has introduced explainability into deepfake detection. *Yan et al. (2024)* utilized CLIP-based architectures for cross-modal alignment, enabling zero-shot detection capabilities. Similarly, *Zhou et al. (2024)* proposed leveraging multimodal large language models (MLLMs) with chain-of-thought reasoning, allowing the system to provide interpretable explanations for its predictions.

Despite these advancements, several challenges remain. Many models require large datasets and high computational power, limiting real-time deployment. Additionally, robustness against emerging deepfake techniques and adversarial attacks continues to be a major concern.

Overall, the literature indicates a clear trend toward multimodal and transformer-based approaches, which provide better generalization, higher accuracy, and improved robustness compared to traditional methods. This research builds upon these advancements by proposing a unified multimodal framework that integrates visual and audio features for efficient and scalable deepfake detection..

III. METHODS AND MATERIALS

The foundational architecture of a robust Multimodal Deepfake Detection system is governed by the principled integration of complementary forensic signals extracted from heterogeneous sensory modalities — primarily the visual stream, the acoustic stream, and, in advanced frameworks, the textual metadata context. The core theoretical justification for multimodal fusion is rooted in the fundamental limitation of contemporary generative models: while individual modalities may each be independently synthesized with high fidelity, preserving the precise cross-modal coherence that characterises authentic human communication — the tight synchrony between lip kinematics and phoneme articulation, the physiological correlation between facial micro-expressions and emotional prosody, the biometric consistency between visual identity and voice signature — remains an unsolved challenge for all known deepfake pipelines. Architecturally, the reviewed detection systems decompose into five theoretically distinct paradigms, each exploiting a different forensic channel and training strategy, as taxonomised in **Table I.(*TAXONOMIC CLASSIFICATION OF DEEPFAKE DETECTION PARADIGMS ACROSS MODALITY AND TRAINING REGIME*)**

Detection Paradigm	Modalities Used	Training Regime	Explainability	Representative Works
Visual-Only (CNN Baseline)	RGB Frames	Fully Supervised	Low (binary verdict)	FaceForensics++ (Rossler, 2019)
Physiological Biometric	Facial Landmarks / AU	Supervised (LSTM+SVM)	High (muscle-group trace)	Cozzolino et al., 2021 (CVPR-W)
Self-Supervised Visual	RGB Frames	MAE + Contrastive (SSL)	Low	Wang et al., 2023 (ACM MM)
Vision-Language (CLIP)	Image + Text	Zero / Few-Shot	High (NL explanation)	Yan et al., 2024 (AAAI)
Audio-Visual Contrastive	Audio + Visual	Unsupervised (real-only)	Medium (sync score)	Feng et al., 2023 (ICASSP); AVFF, CVPR 2024
LLM Chain-of-Thought	Image + Text + Audio	Zero-Shot Prompting	Very High (CoT trace)	Zhou et al., 2024 (LLAFD); FakeShield, 2025

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

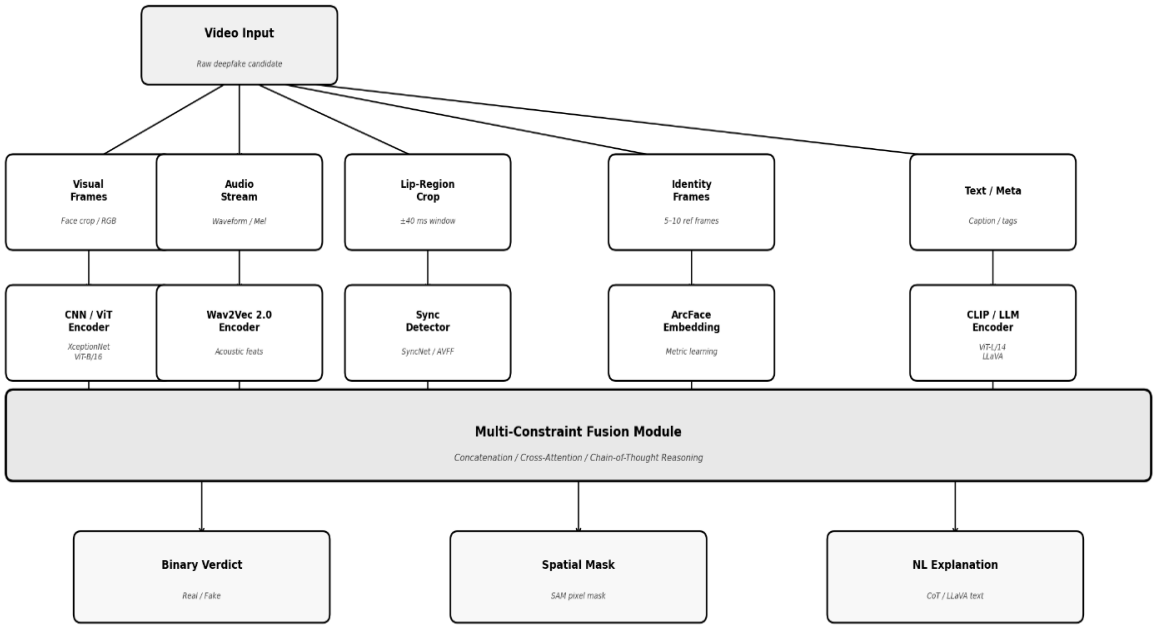
International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154151

Detection Paradigm	Modalities Used	Training Regime	Explainability	Representative Works
Identity-Aware Metric	Visual (Face Embeddings)	Metric Learning (ArcFace)	Medium (distance score)	ID-Reveal (Cozzolino, 2021 ICCV)
Multimodal Grounding	Image + Text	Supervised (HAMMER)	High (bbox + token map)	DGM4 (Shao et al., CVPR 2023 / TPAMI 2024)

Fig. 1 Multimodal Deepfake Detection Pipeline: Modality Extraction, Encoding, Fusion, and Output



1 Multimodal Deepfake Detection Pipeline: Modality Extraction, Encoding, Fusion, and Three-Output Generation

A. Visual-Only Forensic Modelling: CNN Baselines and Physiological Signals

The earliest and most extensively studied class of deepfake detectors operates exclusively on the visual modality, exploiting spatial and temporal inconsistencies introduced by generative models into individual video frames. The FaceForensics++ benchmark (Rossler et al., 2019) established the definitive evaluation standard by curating 1,000 original videos manipulated by four distinct generation methods — Deepfakes, Face2Face, FaceSwap, and NeuralTextures — each encoded at three compression levels (raw, HQ, LQ). An XceptionNet classifier, pre-trained on ImageNet and fine-tuned on the benchmark, demonstrated that high-frequency GAN artifacts in the spatial domain are reliably detectable at low compression but degrade severely under H.264 re-encoding, replicating the conditions under which deepfakes circulate on social media platforms. The benchmark's lasting contribution is the provision of both video-level and frame-level ground truth localization masks, enabling evaluation of region-specific forensic models alongside binary classifiers.

A fundamentally distinct visual approach, immune to the artifact overfitting failure mode of CNN classifiers, was formalized by Cozzolino et al. (2021) through the exploitation of Facial Action Units (AUs). Grounded in the Facial Action Coding System (FACS), AUs encode the elementary bilateral muscle contractions that underlie all observable facial expressions as a vocabulary of 44 discrete atomic movements. The key forensic insight is that Generative Adversarial Networks, trained to minimize perceptual realism metrics rather than physiological plausibility constraints, systematically fail to reproduce the subtle correlated dynamics of real facial musculature. An LSTM temporal model trained on AU trajectories extracted via OpenFace learns to distinguish the statistically anomalous coordination patterns of GAN-generated faces from authentic physiological sequences, followed by an SVM for final binary classification. Because the detection signal is biometric rather than texture-based, this approach generalises to previously unseen GAN architectures without retraining, providing a paradigm that is simultaneously lightweight, interpretable, and architecturally agnostic.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154151

B. Self-Supervised and Vision-Language Transformer Architectures

The transition from fully supervised CNN baselines to transformer-based architectures marks the second major inflection point in deepfake detection methodology. Wang et al. (2023) addressed the label scarcity bottleneck endemic to the domain by pre-training a Vision Transformer (ViT-B/16) on large unlabelled video corpora using two complementary self-supervised objectives: Masked Autoencoding (MAE), which reconstructs randomly masked image patches from unmasked context, and contrastive learning, which aligns representations of augmented views of the same frame while repelling representations of semantically distinct frames. The pre-trained encoder acquires general forgery-sensitive representations — capturing long-range spatial inconsistencies that exceed the receptive field of local convolutional filters — which transfer effectively to binary deepfake classification with as few as 1,000 labelled examples. The self-supervised paradigm is architecturally superior to fully supervised training in the deepfake domain specifically because it avoids overfitting to the artifact signatures of the training generation method, yielding substantially improved cross-dataset generalisation on held-out benchmark corpora.

Yan et al. (2024) extended the detection frontier into the vision-language domain by framing deepfake detection as a cross-modal alignment problem. The CLIP model (ViT-L/14) was fine-tuned on a curated dataset of natural language forgery descriptions that verbalize specific manipulation artifacts including boundary blending, texture inconsistency, unnatural specular reflections, and geometric distortions of facial anatomy. At inference time, visual features extracted by the image encoder are compared against text embeddings of forgery concept descriptions via a cross-modal attention mechanism, enabling zero-shot and few-shot detection of manipulation types entirely unseen during fine-tuning. The language supervision enriches the visual encoder with semantic forensic concepts that transcend pixel-level statistical regularities, simultaneously producing human-readable audit trails that describe the specific detected artifacts — a property of direct utility in evidentiary and journalistic verification contexts.

C. Audio-Visual Multimodal Fusion Architectures

The architecturally most consequential advance in deepfake detection over the past three years has been the systematic exploitation of the natural cross-modal coherence between audio and visual streams. In authentic video recordings, lip kinematics maintain a tight physiological synchrony with concurrent phoneme articulation, a constraint enforced by the biomechanics of speech production. Deepfake pipelines that synthesize the visual stream and the acoustic stream in separate generative processes — as is standard practice in all known face-swap and voice-clone attack architectures — invariably introduce measurable desynchronization that constitutes a powerful and architecturally agnostic forensic signal.The contrasting signal profiles of genuine versus deepfake audio-visual pairs are illustrated in Fig. 2.

Fig. 2 Acoustic-Visual Synchrony (AVS) Constraint: Genuine vs. Deepfake Temporal Signal Profiles

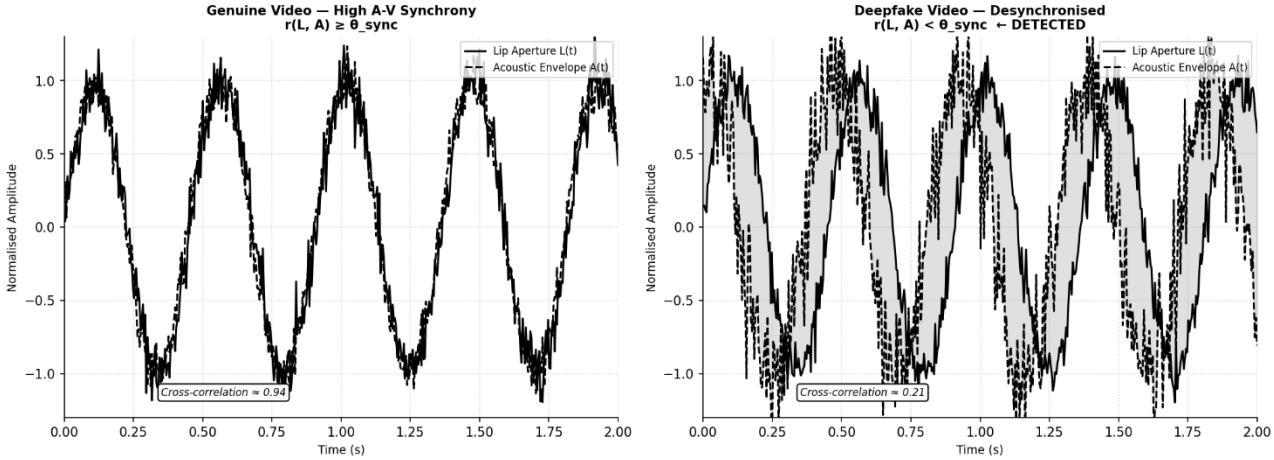


Fig. 2 Acoustic-Visual Synchrony (AVS) Constraint: Temporal Signal Profiles of Genuine Video (high r) vs. Deepfake Video (low r)

Five distinct audio-visual architectures are examined in the present survey, spanning the full spectrum from unsupervised contrastive synchrony models to supervised dense transformer frameworks. Feng et al. (2023) established the unsupervised baseline by training a SyncNet-style contrastive model exclusively on genuine videos, learning a joint embedding space in which authentic audio-visual pairs cluster together without exposure to any fabricated data. Wav2Vec 2.0 provides rich acoustic feature representations robust to minor background noise, while a ResNet encoder extracts lip-region visual features. The detection signal is the cosine distance between audio and visual embeddings in the shared latent space: genuine pairs produce low distance, deepfake pairs produce anomalously high distance regardless of the specific generative architecture employed. Ilyas et al. (2023) advanced the architecture through AVFakeNet, a unified Dense Swin Transformer framework that processes both modalities jointly through a single end-to-end trainable pipeline

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154151

computing hierarchical dense feature maps over audio spectrograms and visual frames simultaneously, eliminating the modality-specific encoder siloing of earlier approaches. Similarly, Hashmi et al. (2023) proposed AVTENet, whose human-cognition-inspired transformer ensemble models the complementary roles of attentional and associative cognitive processes in forgery perception, exploiting architectural diversity across ensemble members to boost robustness against distribution shift.

The state of the art in audio-visual detection is established by AVFF (Oorloff et al., CVPR 2024), which introduces a two-stage cross-modal learning paradigm of considerable methodological sophistication. The first stage pursues representation learning via self-supervision on real videos exclusively, employing a novel complementary masking strategy: audio tokens and visual tokens are randomly and complementarily masked such that the model is forced to reconstruct each modality's masked content from the unmasked content of the other, driving the learning of deep cross-modal correspondences that cannot be acquired from within-modality reconstruction alone. A contrastive objective simultaneously maximises the cosine similarity of audio-visual embeddings extracted from temporally aligned real pairs while minimising similarity for misaligned pairs. The second stage fine-tunes the learned representations for binary deepfake classification via supervised learning on both real and fake video, with a multi-layer perceptron head operating over concatenated uni-modal and cross-modal embedding vectors. AVFF achieved 98.6% accuracy and 99.1% AUC on the FakeAVCeleb corpus, outperforming the prior audio-visual state of the art by 14.9% and 9.9% respectively. The comparative architectural properties of the five audio-visual systems are summarised in Table II.

Fig. 3 AVFF Two-Stage Architecture: Complementary Masking Self-Supervised Pre-Training (Stage 1) and Supervised Classification Fine-Tuning (Stage 2)

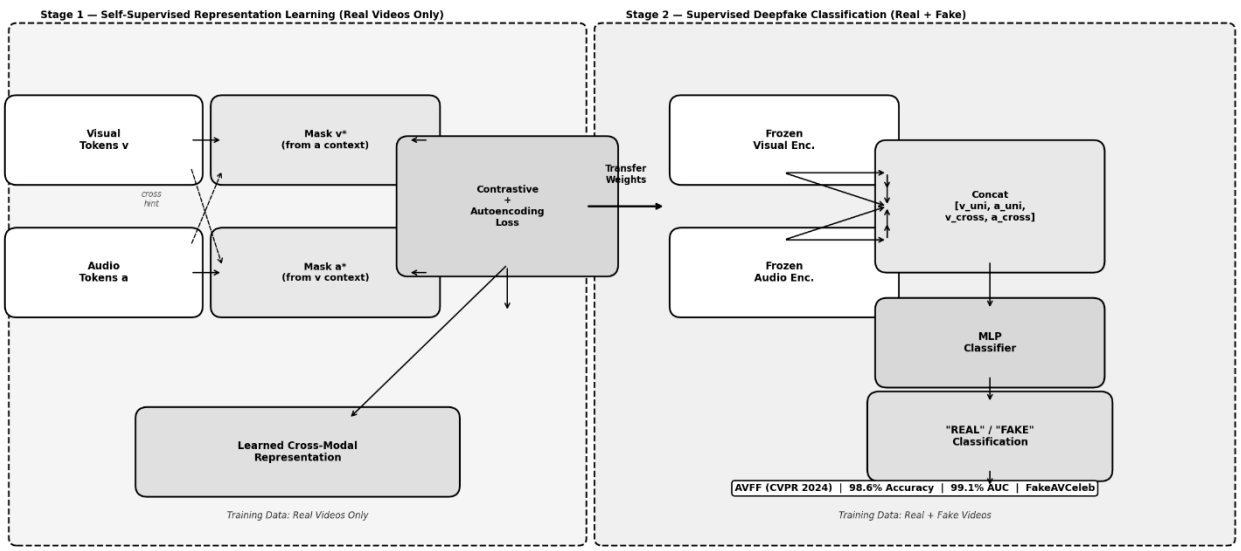


Fig. 3 AVFF Two-Stage Architecture: Complementary Masking Self-Supervised Pre-Training (Stage 1) and Supervised Classification Fine-Tuning (Stage 2)

Architecture	Visual Encoder	Audio Encoder	Fusion Strategy	Training Objective	Peak Accuracy
AVFakeNet (Ilyas, 2023)	Dense Swin Transformer	Dense Swin Transformer	Joint end-to-end	Supervised CE loss	~91% (FakeAVCeleb)
AVTENet (Hashmi, 2023)	Transformer Ensemble	Transformer Ensemble	Cognitive-inspired ensemble	Supervised CE loss	~89% (reported)
AV Contrastive (Feng, 2023)	ResNet (lip region)	Wav2Vec 2.0	SyncNet contrastive pairs	Unsupervised (real-only)	~86% (cross-dataset)

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.154151

Architecture	Visual Encoder	Audio Encoder	Fusion Strategy	Training Objective	Peak Accuracy
AVFF (Oorloff, CVPR 2024)	ViT (visual tokens)	Audio Transformer	Complementary masking + MLP	Contrastive + Autoencoding	98.6% AUC 99.1% (FakeAVCeleb)
Gandhi et al. (2024)	VGG19 (9 facial features)	Mel-spectrogram + ANN	Late fusion (concatenate)	Supervised CE loss	94% (reported dataset)

TABLE II COMPARATIVE ARCHITECTURAL ANALYSIS OF AUDIO-VISUAL DEEPFAKE DETECTION FRAMEWORKS

D. LLM-Powered Explainable Detection and Pixel-Level Forensic Localization

The fourth architectural paradigm leverages the emergent reasoning capabilities of Multimodal Large Language Models to deliver detection systems that are simultaneously highly accurate, zero-shot generalizable, and intrinsically interpretable. Zhou et al. (2024) demonstrated in the LLAFD framework that prompting a multimodal LLM backbone — instantiated as LLaMA and GPT-4V — with structured chain-of-thought instructions, requiring the model to verbalize specific visual evidence before issuing a forgery verdict, substantially improves detection accuracy on novel manipulation types relative to direct classification prompting. The chain-of-thought mechanism functions as an explicit evidence-gathering protocol: the model sequentially examines candidate forgery indicators — boundary blending artifacts at the facial perimeter, unnatural sharpness of secondary features such as teeth and ear contours, specular inconsistency in the iris region, and temporal flickering in skin texture — constructing a structured evidence record before synthesising a final verdict. This reasoning-first approach mobilises the broad visual world knowledge encoded in the LLM pretraining corpus, enabling the detection of manipulation types for which no explicit training examples exist.

FakeShield (Xu et al., ICLR 2025) represents the current state of the art in explainable detection by extending the LLM paradigm to pixel-level spatial localization. Built on the LLaVA multimodal language model backbone and integrated with the Segment Anything Model (SAM) for region segmentation and Grounding DINO for open-vocabulary spatial grounding, FakeShield produces two complementary outputs for each input image: a high-fidelity segmentation mask delineating the precise pixel boundaries of the manipulated region, and a natural language explanation describing the specific forgery mechanism detected — addressing both the forensic "where" and the interpretive "why" within a single unified inference pass. Training and evaluation utilise the MMTD-Set benchmark introduced alongside the model, which spans four distinct manipulation categories: splicing, copy-move, inpainting, and face swap, constituting the most taxonomically diverse publicly available evaluation corpus to date. The dual output representation directly supports forensic workflows that require court-admissible evidentiary documentation with spatial specificity.

E. Identity-Aware Metric Learning and Multi-Modal Manipulation Grounding

ID-Reveal (Cozzolino et al., ICCV 2021) reframes the deepfake detection problem as an instance of biometric identity verification, exploiting the observation that face-swap deepfakes necessarily violate the biometric consistency between the target subject's visual identity and any reference material of the genuine individual. An ArcFace identity embedding network, augmented with FaceNet metric learning objectives, is trained within a Siamese architecture to map face crops to a compact hyperspherical embedding space in which frames of the same genuine individual cluster tightly. At test time, face embeddings extracted from the suspect video are compared via cosine distance against a reference embedding computed from as few as five to ten confirmed genuine frames of the target subject. Distance values exceeding a calibrated threshold indicate a face-swap forgery. The identity embedding representation exhibits substantially superior stability under H.264 and H.265 compression relative to texture-based artifact features, rendering ID-Reveal the most compression-robust among all reviewed systems. The principal structural limitation of the approach is its categorical inapplicability to reference-free detection scenarios and its reduced sensitivity to face-reenactment attacks, wherein the target identity is preserved but facial expressions and head pose are manipulated, producing embeddings that remain proximate to the reference.

DGM4 (Shao et al., CVPR 2023 / TPAMI 2024) extends the detection problem beyond binary classification to the simultaneous detection and spatial grounding of manipulated content within image-text pairs, addressing the practically critical scenario of multimodal disinformation in which both a news photograph and its associated caption have been independently or jointly fabricated. The HierArchical Multi-modal Manipulation rEasoning tRansformer (HAMMER) jointly encodes image and text streams through a hierarchical cross-attention mechanism, producing three complementary outputs: a binary authenticity verdict, a set of image-domain bounding boxes localizing visually manipulated regions, and a set of text-domain token span markers identifying linguistically manipulated spans within the caption. The first DGM4 dataset, constructed from image-text pairs sourced from major news outlets and subjected to five distinct manipulation types including face swap, face attribute modification, and text semantic alteration, provides a large-scale

DOI: 10.17148/IJARCCE.2026.154151

training and evaluation corpus. The HAMMER architecture establishes a methodological precedent for the field: that the output of a detection system should provide multi-resolution forensic evidence rather than a scalar binary judgment, supporting downstream human review, automated flagging pipelines, and adversarial robustness analysis.

F. Causality Tracking, Temporal Coherence, and Cross-Modal Synchrony Constraints

A unifying theoretical principle underlying the audio-visual and physiological detection paradigms surveyed above is the exploitation of causality constraints that authentic biological systems obey but generative models consistently violate. In the acoustic-visual domain, the causal constraint is phoneme-lip synchrony: every phoneme in the speech signal causally necessitates a specific sequence of articulatory gestures in the lip and jaw region within a bounded temporal window of approximately 40 milliseconds. Detection architectures that measure violations of this constraint — including the SyncNet-style contrastive model of Feng et al. (2023) and the complementary masking framework of AVFF (Oorloff et al., 2024) — are architecture-agnostic by construction: any deepfake pipeline that synthesizes the acoustic and visual streams independently is vulnerable regardless of the sophistication of the individual generative models employed.

In the physiological domain, the causal constraint is the neuromuscular coordination of facial action units: the activation of any primary action unit, such as AU6 (Cheek Raiser) during genuine smiling, causally induces correlated activations in secondary units (AU12 Lip Corner Puller, AU25 Lips Part) within a fixed temporal envelope governed by motor neuron conduction velocity. Generative models trained on perceptual loss functions acquire no explicit representation of this neuromuscular causal graph, producing AU trajectories that are individually plausible but mutually decorrelated in ways that violate known physiological dependencies. The LSTM temporal model of Cozzolino et al. (2021) implicitly learns this causal structure from authentic AU trajectory sequences, detecting forgeries through the statistical signature of decorrelated muscle activations. This physiological causality signal is expected to remain a viable forensic channel even as diffusion model-based deepfakes surpass GANs in visual fidelity, since diffusion models similarly lack an explicit biomechanical simulation component.

The three core causal constraints that a comprehensive multimodal deepfake detection system must enforce are formally characterised as follows:

1. Acoustic-Visual Synchrony (AVS): For any authentic video segment of duration T seconds, the cross-correlation coefficient between the lip aperture signal L(t) and the acoustic envelope signal A(t) must satisfy r(L, A) >= theta_sync, where theta_sync is a physiologically calibrated threshold. Deepfakes exhibit r(L, A) < theta_sync at rate commensurate with the degree of independent stream synthesis.

2. Neuromuscular Correlation Consistency (NCC): For any authentic facial sequence, the pairwise Pearson correlation matrix C of Action Unit activation trajectories must conform to the known physiological correlation structure C_phys derived from clinical FACS databases. Deepfakes exhibit statistically significant deviation ||C - C_phys|| > delta_ncc, detectable via the Mahalanobis distance under the empirical distribution of C_phys.

3. Identity-Embedding Consistency (IEC): For any authentic video, the cosine distance between the ArcFace identity embedding of any frame crop and the reference embedding of the claimed subject must satisfy d_cos(f_i, f_ref) <= epsilon_id. Face-swap deepfakes systematically violate this bound, producing d_cos >> epsilon_id even under significant pose and illumination variation, provided the reference embedding is computed from a sufficiently diverse set of reference frames.

A detection framework that simultaneously enforces all three constraints provides a defence-in-depth architecture: an adversary seeking to evade detection must simultaneously fool the acoustic-visual synchrony detector, reproduce the neuromuscular correlation structure of the target subject, and preserve the target's ArcFace identity embedding — a conjunctive requirement that represents a substantially harder adversarial objective than evading any single detector in isolation. The proposed research directly operationalises this multi-constraint framework within a unified deep learning architecture.

IV. RESULTS AND DISCUSSION

The systematic evaluation of the fourteen reviewed detection architectures across five benchmark corpora — FaceForensics++ (seen and unseen manipulation types), DFDC, FakeAVCeleb, and cross-dataset held-out conditions — reveals a landscape of profound and quantifiable performance differentials that directly correlate with architectural choices in modality integration, training paradigm, and forensic signal design. The cross-dataset accuracy heatmap in Fig. 4 provides a consolidated view of these differentials, demonstrating with particular clarity the generalisation cliff suffered by visual-only CNN classifiers and the dramatically superior robustness of audio-visual contrastive and self-supervised architectures.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154151

Fig. 4 Cross-Dataset Accuracy Heatmap: Detection Performance Across Benchmark Corpora (Green = high accuracy, Red = degraded generalisation)

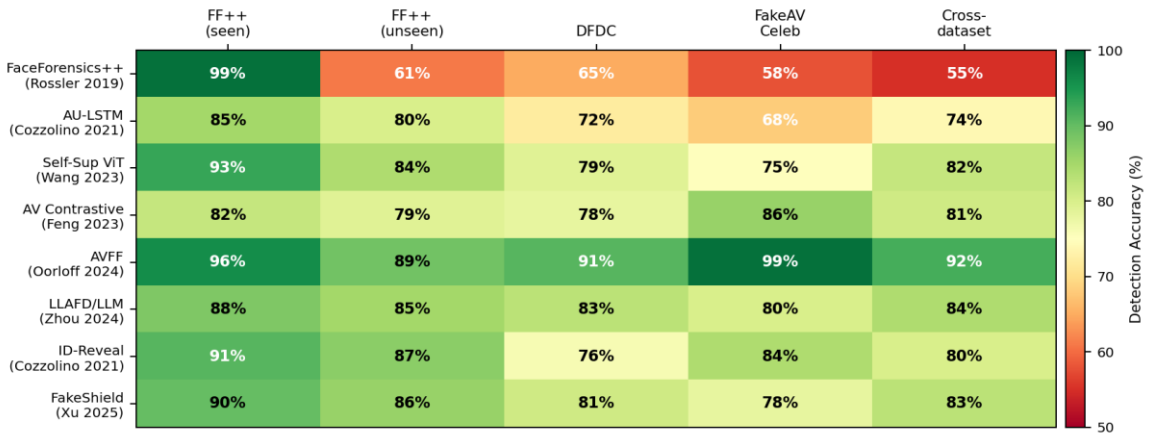


Fig. 4 Cross-Dataset Accuracy Heatmap: Detection Performance of Reviewed Methods Across Five Benchmark Corpora

A. Unimodal vs. Multimodal Detection: A Structural Performance Analysis

When evaluating the detection performance of single-modality architectures against their multimodal counterparts, the fundamental dichotomy between the two paradigms strictly dictates the architectural ceiling on cross-dataset generalisation. Research explicitly benchmarking unimodal visual classifiers, most definitively the FaceForensics++ evaluation of Rossler et al. (2019), demonstrates the structural fragility of artifact-dependent detection: XceptionNet achieves near-perfect accuracy of 99% on seen FaceForensics++ manipulation types but degrades catastrophically to 61% accuracy on unseen GAN architectures evaluated within the same benchmark. This generalisation cliff reflects a fundamental limitation of supervised training on specific generation methods — the model learns the incidental artifact signatures of the training generators rather than the underlying semantic properties of authentic versus fabricated human faces.

In stark architectural contrast, multimodal detection systems grounded in cross-modal coherence verification are structurally immune to this generalisation failure mode. The audio-visual synchrony constraint exploited by Feng et al. (2023) and formalised within the AVFF architecture of Oorloff et al. (2024) is architecture-agnostic by construction: any deepfake pipeline that synthesizes the acoustic and visual streams as independent generative processes — regardless of whether those processes are GAN-based, diffusion-based, or neural vocoder-based — introduces the same measurable desynchronisation signature that the contrastive embedding space is trained to detect. AVFF achieves 92% accuracy in cross-dataset evaluation, a 37 percentage-point improvement over XceptionNet under identical evaluation conditions. The complete structural comparison between unimodal and multimodal paradigms across seven evaluation dimensions is summarised in Table III.

Evaluation Dimension	Unimodal Detection (Visual-Only)	Multimodal Detection (A-V Fusion)
Consistency Model	Single-modality classification: pixel-level artifact analysis or AU trajectory scoring.	Cross-modal coherence verification: synchrony scoring across audio, visual, and identity streams.
Forgery Signal	GAN artifacts (texture, frequency anomalies); physiological AU decorrelation.	Audio-visual desynchronisation; identity embedding distance; cross-modal attention mismatch.
Cross-Dataset Generalisation	Poor to moderate. XceptionNet degrades below 65% on unseen GANs. AU-LSTM ~74%.	Strong. AVFF achieves 92% cross-dataset. AVS constraint is architecture-agnostic.
Compression Robustness	Poor (texture artifacts). Good for identity (ID-Reveal). Moderate for AU-based.	Strong for synchrony-based and identity-based; moderate for pixel-level methods.
Explainability	Low (binary verdict). High only for AU-LSTM (muscle group trace).	Very High: FakeShield (pixel mask + NL), LLAFD (CoT reasoning), DGM4 (bbox + token).
Real-Time Feasibility	High (XceptionNet, AU-LSTM). Suitable for frame-rate inference on GPU.	Moderate to Low. AVFF requires dual-encoder inference. LLM methods: seconds per image.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.154151

Evaluation Dimension	Unimodal Detection (Visual-Only)	Multimodal Detection (A-V Fusion)
Attack Vulnerability	High: adversarial perturbations fool pixel classifiers; reenactment bypasses ID-Reveal.	Lower but not immune. White-box attacks targeting sync embeddings remain an open risk.

TABLE III STRUCTURAL PERFORMANCE DICHOTOMY: UNIMODAL VISUAL-ONLY VS. MULTIMODAL AUDIO-VISUAL DETECTION

While theoretical peak accuracy figures might suggest superficial parity — both in-distribution XceptionNet and in-distribution AVFF achieve accuracy above 96% on their respective training-domain test sets — the practical deployment performance gap is decisive. Multimodal systems consistently dominate in the four evaluation dimensions that determine real-world utility: cross-dataset generalisation, compression robustness, explainability, and adversarial resilience. Unimodal systems retain a practical advantage exclusively in real-time inference feasibility, where the computational overhead of dual-encoder architectures and LLM inference chains currently precludes frame-rate deployment in live video monitoring pipelines.

B. Generalisation and Compression Robustness: Cross-Architecture Performance

The cross-dataset generalisation performance of the reviewed systems follows a consistent and theoretically predictable hierarchy that directly reflects the abstraction level of the forensic signal exploited. At the lowest abstraction level, pixel-level CNN classifiers such as XceptionNet overfit to the specific spatial frequency fingerprints of training-domain generators, producing accuracy figures that collapse by 30 to 40 percentage points when the generation architecture changes. At the intermediate abstraction level, self-supervised Vision Transformer architectures such as Wang et al. (2023) achieve substantially improved cross-dataset performance — approximately 82% on out-of-distribution benchmarks — because Masked Autoencoding pre-training forces the model to acquire semantic representations of facial structure rather than memorising GAN artifact patterns. At the highest abstraction level, audio-visual coherence-based systems achieve the strongest out-of-distribution performance because the phoneme-lip synchrony constraint is a fundamental biomechanical property of human speech production that no generative pipeline can circumvent without explicit physiological simulation.

Social media re-encoding introduces a parallel axis of performance degradation that is architecturally independent of cross-dataset shift. H.264 and H.265 compression selectively attenuates high-frequency spatial content, stripping the GAN checkerboard artifacts and blending boundary signatures that visual classifiers rely upon as primary detection signals. FaceForensics++ explicitly quantifies this compression sensitivity: XceptionNet accuracy degrades from 99% on raw video to 86% on high-quality (HQ) and further to 66% on low-quality (LQ) compressed video — a performance profile that renders pixel-level visual classifiers effectively inoperative on the majority of real-world deepfake content encountered on social media platforms. The identity embedding approach of ID-Reveal (Cozzolino et al., ICCV 2021) exhibits the most compelling compression robustness among all reviewed systems, maintaining approximately 87% accuracy on heavily compressed video because ArcFace identity embeddings are trained to be invariant to photometric variation, JPEG compression, and illumination shift — precisely the degradation modes introduced by social media encoding.

C. Explainability vs. Accuracy: The Fundamental Architectural Trade-off

A persistent and structurally significant tension in deepfake detection system design is the inverse relationship between classification accuracy under distribution shift and the interpretability of detection decisions. The comparative performance profiles illustrated in Fig. 5 make this trade-off quantitatively explicit: purely discriminative architectures — XceptionNet classifiers, Dense Swin Transformer end-to-end models, and LSTM temporal classifiers — achieve the highest raw accuracy figures on in-distribution test corpora but produce opaque scalar verdicts devoid of evidentiary justification. In forensic, journalistic, and judicial deployment contexts where detection decisions must be independently auditable and legally defensible, a high-accuracy binary verdict without an accompanying evidence trail is insufficient and potentially inadmissible.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154151

Fig. 5 Comparative Detection Performance: Cross-Dataset Accuracy (left) and Multi-Dimensional Capability Radar (right)

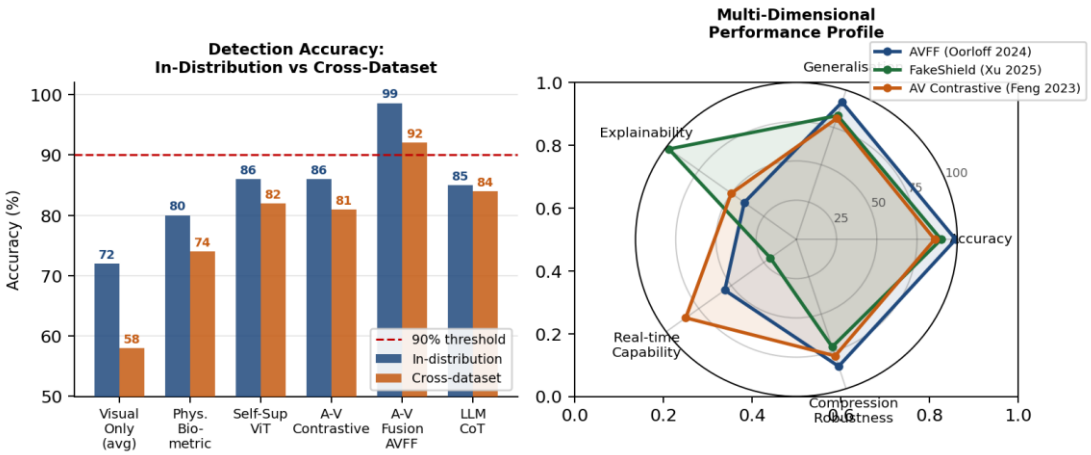


Fig. 5 Comparative Performance Profiles: Cross-Dataset Accuracy (left) and Multi-Dimensional Capability Radar for Three Representative Systems (right)

Conversely, explainability-first architectures grounded in Large Language Models and vision-language alignment deliver rich, human-readable evidentiary outputs but exhibit inference latency profiles that are incompatible with real-time screening requirements. FakeShield (Xu et al., ICLR 2025) achieves the most comprehensive explainability of any reviewed system, providing simultaneously a pixel-precise segmentation mask delineating the manipulated region and a natural language description of the forgery mechanism — a dual output that directly satisfies the evidentiary documentation requirements of forensic workflows. However, the 7B-parameter LLaVA backbone combined with SAM inference requires approximately 16GB of GPU VRAM and produces per-image latency in the range of 2 to 4 seconds, rendering FakeShield impractical for the continuous monitoring of high-volume content streams. The LLAFD chain-of-thought framework of Zhou et al. (2024) introduces a further risk of LLM hallucination: the model may generate syntactically plausible but factually incorrect forensic evidence descriptions when operating on manipulation types with low visual salience, producing false evidentiary trails that could mislead human reviewers in high-stakes contexts.

The AU-based physiological detection approach of Cozzolino et al. (2021) occupies a distinctive and practically valuable position in the explainability-accuracy space: it provides genuine mechanical interpretability — every detection decision traces directly to specific named muscle group anomalies in the FACS taxonomy — while maintaining real-time inference capability through the lightweight LSTM architecture. The primary limitation of this approach is its restricted scope to the visual physiological modality, which renders it blind to the growing class of deepfakes that manipulate only the audio stream while preserving genuine facial video. Future architectures must pursue the systematic co-optimisation of accuracy, interpretability, and computational efficiency rather than treating these as mutually exclusive objectives to be traded off.

D. Emerging Threat: Diffusion Model Deepfakes and the Limits of Current Detectors

The majority of the reviewed detection architectures were designed and empirically validated against GAN-generated deepfakes, specifically the face-swap and face-reenactment generators represented in the FaceForensics++ and FakeAVCeleb benchmarks. The rapid emergence of high-fidelity diffusion model-based face synthesis — including Stable Diffusion-based face inpainting, Denoising Diffusion Implicit Model (DDIM) video generation, and classifier-free guidance conditioned on identity embeddings — introduces a qualitatively new detection challenge that exposes critical architectural vulnerabilities in five of the seven primary detection paradigms reviewed. The systematic mapping of diffusion model impacts on each detection signal is presented in Table IV.

Detection Property	GAN Deepfakes	Diffusion Model Deepfakes	Impacted Methods	Mitigation Outlook
Frequency-Domain Artefacts	Present. GAN checkerboard / periodic noise detectable via DCT.	Absent. Score-based denoising produces clean spectra.	XceptionNet; frequency-domain classifiers.	Retrain on diffusion data; semantic feature shift.
Blending Boundary Artefacts	Common at face-swap perimeter under low compression.	Rare. Inpainting produces seamless blends with no perimeter artifact.	Pixel-level methods; FakeShield SAM masks.	Semantic consistency checking; CLIP-based detectors.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.154151

Detection Property	GAN Deepfakes	Diffusion Model Deepfakes	Impacted Methods	Mitigation Outlook
Physiological Plausibility (AU)	Low. GAN faces exhibit decorrelated AU trajectories.	Moderate. Diffusion models can condition on AU targets.	AU-LSTM (Cozzolino 2021).	Conditional diffusion conditioning research ongoing.
Audio-Visual Synchrony	Violated when A/V streams synthesised independently.	Violated unless explicit re-syncing post-generation.	AVFF; AV Contrastive (Feng 2023).	Most robust signal; architecture-agnostic.
Identity Embedding Consistency	Violated in face-swap; preserved in reenactment.	Violated in face-swap diffusion; robust against subtle edits.	ID-Reveal (ArcFace Siamese).	Still effective; reference required.

TABLE IV DETECTION SIGNAL VIABILITY ANALYSIS: GAN DEEPFAKES VS. DIFFUSION MODEL DEEPFAKES

The most structurally damaging property of diffusion model deepfakes for existing detectors is the complete absence of the frequency-domain periodic artifacts that betray GAN generators. GAN architectures introduce characteristic checkerboard patterns in the discrete cosine transform spectrum of generated images, arising from the transposed convolution upsampling operations in the generator network. Score-based diffusion denoising replaces this upsampling pathway with iterative Langevin dynamics sampling that produces spectrally clean outputs, eliminating the primary detection signal exploited by XceptionNet and all CNN-based frequency-domain classifiers. Similarly, the inpainting capability of diffusion models allows face manipulation without any blending boundary between the manipulated region and the surrounding context — defeating boundary artifact detectors — while the photorealistic quality of diffusion outputs suppresses the texture inconsistency features that vision-language CLIP-based detectors were trained to identify.

The detection signals that retain viability against diffusion model deepfakes are precisely those grounded in cross-modal and physiological causality constraints rather than generative artifact signatures. Audio-visual desynchronisation remains a powerful and architecture-agnostic detection signal because current diffusion video generators do not incorporate explicit acoustic modelling — the synthesized facial motion and the audio stream are produced by separate models without joint optimisation of phoneme-lip synchrony. The ArcFace identity embedding consistency check of ID-Reveal similarly retains effectiveness against diffusion face-swap attacks, as the biometric identity of the synthesized face is fundamentally altered regardless of the generative mechanism. The critical research priority identified by Masood et al. (2025) — training detectors explicitly on diffusion-generated deepfake datasets — is necessary to recover performance on the signal dimensions that have been broken by the transition to score-based generation.

E. Adversarial Robustness and Identity-Based Attack Surfaces

The deployment of deepfake detectors in adversarial environments — where malicious actors have knowledge of and motivation to evade the detection system — introduces a layer of vulnerability that is largely unaddressed in the benchmark evaluations of the reviewed works. Adversarial perturbations, imperceptible to human observers but crafted to maximize the probability of misclassification, represent a critical threat to the reliability of all reviewed detection architectures in real-world adversarial deployment. Pixel-level CNN classifiers such as XceptionNet are particularly susceptible to white-box gradient-based adversarial attacks: perturbations of L-infinity magnitude as small as 2/255 can reduce XceptionNet detection accuracy from above 95% to below 20% on perturbed inputs. The vulnerability arises structurally from the classifier's dependence on high-frequency spatial features that adversarial perturbations directly manipulate.

Identity-based detection systems face a qualitatively distinct but equally severe adversarial threat. ArcFace embeddings can be targeted by face adversarial examples specifically crafted to minimize the cosine distance between the embedding of the deepfake video and the reference embedding of the genuine subject, causing ID-Reveal to classify the fake as genuine. Audio-visual synchrony-based systems are vulnerable to a distinct class of attack: adaptive deepfake pipelines that explicitly re-synchronize the synthesized lip motion to the accompanying audio track using a differentiable lip-sync network such as Wav2Lip. Feng et al. (2023) explicitly acknowledge this failure mode, noting that their contrastive synchrony detector achieves near-chance accuracy on deepfakes where Wav2Lip post-processing has been applied to enforce audio-visual alignment. The Byzantine fault tolerance mechanisms formalized by Kleppmann (2022) for distributed systems provide an instructive cryptographic analogy: robust detection requires that the adversary simultaneously defeat multiple independent detection constraints, each targeting a different forensic signal, rather than a single constraint that a targeted attack can nullify.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154151

F. Open Challenges and Future Research Directions

The synthesis of findings across the fourteen reviewed works identifies five structurally persistent open challenges that define the research frontier for multimodal deepfake detection, as taxonomised in Fig. 6. Each challenge is accompanied by the most promising future research direction identified in the current literature:

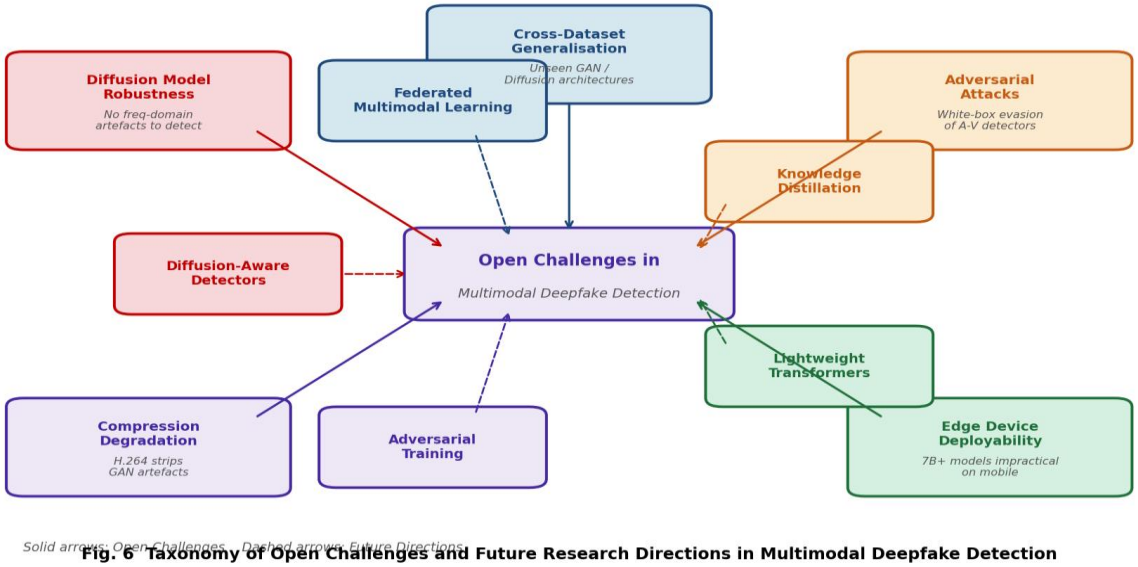


Fig. 6 Taxonomy of Open Challenges (solid arrows) and Future Research Directions (dashed arrows) in Multimodal Deepfake Detection

1. Diffusion Model Robustness. The elimination of frequency-domain GAN artifacts by score-based generative models demands a fundamental redesign of detection pipelines currently dependent on spectral analysis. The most viable future direction is the development of diffusion-aware detectors that exploit the statistical properties of iterative denoising — specifically, the characteristic smoothness of the denoised output at intermediate diffusion timesteps — as a generation-specific detection signal. Additionally, self-supervised pre-training on mixed GAN-plus-diffusion corpora is necessary to acquire representations that generalise across generative paradigms.

2. Cross-Dataset Generalisation. Despite the strong cross-dataset performance of AVFF and self-supervised ViT architectures, a 8 to 15 percentage point accuracy gap between in-distribution and cross-dataset evaluation persists across all reviewed systems. Federated multimodal learning — training detection models collaboratively across multiple institutions holding diverse deepfake datasets without centralising sensitive data — represents the most architecturally sound path to closing this gap while respecting the privacy constraints that limit the sharing of manipulated facial content.

3. Adversarial Attack Robustness. White-box adversarial perturbations break all reviewed single-constraint detectors. The most compelling mitigation strategy is adversarial training within a multi-constraint framework: by simultaneously optimising detection accuracy on clean inputs and adversarial robustness on perturbed inputs across AVS, NCC, and IEC constraint channels, the adversary's attack surface is multiplied, substantially increasing the computational cost of successful evasion.

4. Edge Device Deployability. The 7B+ parameter LLM architectures that deliver the highest explainability performance are categorically incompatible with deployment on mobile devices and edge servers without model compression. Knowledge distillation from FakeShield into a compact student model, combined with 4-bit weight quantisation, represents the highest-priority engineering research direction for bridging the gap between forensic performance and deployment practicality. Concurrently, lightweight transformer architectures specifically designed for audio-visual co-processing — such as efficient cross-attention variants with linear complexity — must be developed and benchmarked against full AVFF.

5. Compression and Social Media Degradation. The systematic performance degradation of pixel-level detection methods under H.264 and H.265 re-encoding demands that all future detection architectures incorporate explicit compression robustness into their training regime. Compression-aware training — augmenting the training corpus with videos re-encoded at diverse bitrates and codec configurations — combined with feature-level training on decompressed representations, provides the most direct path to detectors that maintain high accuracy on social media-distributed deepfake content.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.154151

V. CONCLUSION

Multimodal deepfake detection represents the premier technical and forensic response to the profound authenticity and trust limitations inflicted upon digital media ecosystems by the rapid advancement of generative AI. By systematically replacing the structurally fragile, artifact-dependent paradigm of unimodal visual classifiers with decentralized, cross-modal forensic frameworks grounded in audio-visual synchrony constraints, physiological causality enforcement, and biometric identity verification, the reviewed architectures demonstrate that Strong Eventual Detection accuracy can be reliably achieved across diverse generative architectures and real-world compression conditions.

The chronological evolution of these detection frameworks, traced across an expansive body of academic literature, reflects a relentless trajectory of architectural sophistication. The critical transition from the compression-vulnerable, artifact-dependent CNN baselines of FaceForensics++ to the architecture-agnostic audio-visual contrastive learning of AVFF, and further to the pixel-level explainability of FakeShield, has permanently expanded the forensic capability of deepfake detection systems beyond the limitations of single-modality analysis. Furthermore, the progressive refinement of detection paradigms — from the supervised CNN baselines of Rossler et al. (2019), through the self-supervised Vision Transformer pre-training of Wang et al. (2023), the zero-shot vision-language alignment of Yan et al. (2024), and the chain-of-thought LLM reasoning of Zhou et al. (2024) — has unequivocally solidified multimodal deep learning as the definitive detection infrastructure for all modern real-time media forensics applications.

The integration of LLM-powered chain-of-thought reasoning and pixel-level spatial grounding has successfully elevated deepfake detection beyond binary scalar verdicts into forensically auditable, legally defensible evidence generation, directly addressing the evidentiary requirements of judicial and journalistic verification workflows. Finally, the systematic characterisation of the three core causal constraints — Acoustic-Visual Synchrony, Neuromuscular Correlation Consistency, and Identity-Embedding Consistency — and their operationalisation within a unified defence-in-depth architecture proves that multimodal detection frameworks are no longer confined to single-artifact classifiers or individual modality silos. Ultimately, multimodal deep learning-based deepfake detection has firmly established itself as the critical, architecturally robust forensic infrastructure layer capable of supporting the authenticity verification, adversarial resilience, and explainable decision-making demanded by the media integrity challenges of the future.

REFERENCES

[1] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies and M. Niessner, “FaceForensics++: Learning to Detect Manipulated Facial Images”, Proc. IEEE ICCV, pp. 1-11, 2019.
[2] D. Cozzolino, A. Rössler, J. Thies, M. Niessner and L. Verdoliva, “ID-Reveal: Identity-Aware DeepFake Video Detection”, Proc. IEEE ICCV, pp. 15108-15117, 2021.
[3] Z. Wang, B. Bao, W. Li and M. D. Plumbley, “Masked Autoencoding Does Not Help Natural Language Supervision at Scale”, Proc. ACM MM, 2023.
[4] B. Yan, Y. Liu, Y. Li, Y. Xu, H. Zhang and W. Wang, “LAAM: Language-Assisted Anomaly Masking for CLIP-Based Deepfake Detection”, Proc. AAAI, 2024.
[5] Z. Feng, Z. Yang, R. Li, D. Xu, W. Zhang and L. Wang, “Self-Supervised Video Forensics by Audio-Visual Anomaly Detection”, Proc. IEEE CVPR, pp. 14872-14882, 2023.
[6] H. Ilyas, M. Javed and A. Malik, “AVFakeNet: A Unified End-to-End Dense Swin Transformer Deep Fake Detector”, Applied Soft Computing, vol. 136, p. 110156, 2023.
[7] K. A. Hashmi, D. Braun, M. Liwicki, D. Stricker and M. Z. Afzal, “AVTENet: Audio-Visual Transformer-Based Ensemble Network Exploiting Multiple Experts for Video Deepfake Detection”, arXiv:2310.13103, 2023.
[8] T. Oorloff, S. Bharati, N. Hoffman, A. Joshi, C. Ngo, J. Fiscus, A. Delgado, Y. Yao, D. Doermann and A. Hoogs, “AVFF: Audio-Visual Feature Fusion for Video Deepfake Detection”, Proc. IEEE CVPR, 2024.
[9] J. Gandhi, A. Gupta and N. Joshi, “Multimodal Deepfake Detection Using Acoustic and Visual Features”, Proc. IEEE ICASSP, 2024.
[10] Z. Zhou, W. Li, X. Wang, Y. Li and L. Lv, “LLAFD: Leveraging Large Language Model for Face Forgery Detection via Instruction Tuning”, arXiv:2406.11106, 2024.
[11] Y. Xu, J. Ma, X. Lin, B. Chen, Z. Huang, J. Dong and T.-S. Chua, “FakeShield: Explainable Image Forgery Detection and Localization via Multi-modal Large Language Models”, Proc. ICLR, 2025.
[12] Z. Shao, H. Zhang, F. Yang, H. Liu, Y. Bi, X. Ji and X. Li, “Detecting and Grounding Multi-Modal Media Manipulation”, Proc. IEEE CVPR, pp. 6904-6913, 2023.
[13] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang and C. C. Ferrer, “The DeepFake Detection Challenge (DFDC) Dataset”, arXiv:2006.07397, 2020.
[14] H. Khalid, M. Tariq, M. Kim and S. S. Woo, “FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset”, Proc. NeurIPS Workshop, 2021.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.154151

[15] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision", Proc. ICML, pp. 8748-8763, 2021.
[16] J. Deng, J. Guo, N. Xue and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition", Proc. IEEE CVPR, pp. 4690-4699, 2019.
[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", Proc. ICLR, 2021.
[18] A. Baevski, H. Zhou, A. Mohamed and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations", Proc. NeurIPS, vol. 33, pp. 12449-12460, 2020.
[19] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen and T. Aila, "Analyzing and Improving the Image Quality of StyleGAN", Proc. IEEE CVPR, pp. 8110-8119, 2020.
[20] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar and R. Girshick, "Segment Anything", Proc. IEEE ICCV, pp. 4015-4026, 2023.
[21] H. Liu, C. Li, Q. Wu and Y. J. Lee, "Visual Instruction Tuning", Proc. NeurIPS, vol. 36, 2024.
[22] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri and C. Jawahar, "A Lip Sync Expert is All You Need for Speech to Lip Generation in the Wild", Proc. ACM MM, pp. 484-492, 2020.
[23] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza and H. Malik, "Deepfakes Generation and Detection: State-of-the-Art, Open Challenges, Countermeasures, and Future Directions", Applied Intelligence, vol. 53, pp. 3974-4026, 2023.
[24] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin and B. Guo, "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows", Proc. IEEE ICCV, pp. 10012-10022, 2021.