Journal of Information and Cyber Security (JICS)
Volume 2, Issue 2, Date: 31- December -2026
ISSN: 2978-6797, DOI: 10.64820/AEPJICS.22.35.40.122026

Multimodal Deepfake Detection Using Video Analysis and Social Metadata Signals

Nikhil Sharma [1, *], Ved Singh [2], Om Shingare [3], Abira Banik [4]
[1, 2, 3, 4] Department of Information Technology, St. John College of Engineering and Management, Palghar, India
Email: [1] nikhilsharma0437@gmail.com, [2] singhved1502@gmail.com, [3] omshingare2508@gmail.com, [4] abirabanikofc@gmail.com
*Corresponding Author

***Abstract*—The spread of hyper-realistic deepfakes via social networks requires powerful detection systems that can stand the aggressive video compression, which, in most cases, makes the conventional forensic models useless. A new hybrid visual forensics and social signal modeling Dual-Pipeline System is suggested to deal with the sudden performance decreases in low-bitrate in-the-wild content to combine the two. The visual architecture is based on the EfficientNet to extract spatial features with Vision Transformers to model global spatio-temporal consistency, which reduces the effect of compression artifacts. At the same time, to reveal that commenter skepticism is an extrinsic forensic feature with the help of BERT-based sentiment analyzers, a parallel textual pipeline is used, and user comment skepticism is interpreted as substitute evidence. The modalities are combined using a dynamic weighting formula which shifts the dependence of social cues dynamically when there is loss of visual quality. This multimodal fusion is proved to be very useful in boosting detection accuracy with compressed media over unimodal baselines, providing a robust platform to content moderation in the real world.**

Keywords—Deepfake Detection, Multimodal Fusion, Vision Transformers, EfficientNet, Social Computing, Media Forensics.

I. INTRODUCTION

A. Background and Motivation

Generative AIs, which are rapidly developed due to generative adversarial networks (GANs), variational autoencoders (VAEs), and diffusion models, have proven to be a significant threat to the integrity of digital media [1]. Although the technologies are used to make legitimate uses and applications, they are also misused to create deep-fake content that is highly realistic [16]. Extensive abuse of deep-fakes has created a range of serious risks to society and includes political and economic infidelity through identity theft [1]. With the amount of synthetic media growing exponentially unprecedentedly on the Web, the implementation of trusted automated deep-fakes detectors has emerged among a mandatory prerequisite of modern information-security platforms [16]. However, new generative methods involve learning that is sensitive to frequency and as such reduce traditional artefacts within forensic methods, lowering the performance of old detection systems to a significant degree [1].

B. Problem Statement

The main technical problem addressed in this paper is the significant drop in performance of deepfake detection algorithms on compressed, in the wild social media data. The deep compression schemes used by social media platforms like (JPEG) and (MPEG) quantization are also major contributors to this degradation and essentially act as low-pass filters wiping artifacts and mesoscopic features at the pixel level of an image that Convolutional Neural Networks (CNNs) otherwise rely on to perform accurate classification [15]. Besides, unimodal visual detectors are often unable to provide enough semantic information to identify advanced, so-called ideal, deepfakes, where visual cues are inconsequential but linguistic anomalies (such as lip-synch mistakes or inconsistencies) are easy to detect visually [22].

C. Limitations of Existing Solutions

The analysis of existing literature shows two main limitations of current methodologies:

1- Architecture Fragility: The architectures of traditional detection systems are mostly based on single-modality convolutional neural networks (CNN) architecture (e.g. XceptionNet). Whereas CNNs are good at capturing local spatial clues, they do not fare well in capturing global temporal correlations needed to identify video frame inconsistencies [7]. On the other hand, pure Vision Transformer (ViT), though capable of capturing the global context of an image, may require a large number of computational units and very large datasets to converge, making them completely impractical to run on machines with limited resources [2].

2- Ignorance of Extrinsic Signals: Deepfake detection studies, with a few exceptions, entirely analyze only intrinsic signals (pixel and audio data) [1]. Such a strategy ignores an abundant source of extrinsic forensic material on social sites, the crowd wisdom of user comments and interaction metadata. According to Deepfake Eval2024, user flags often form the basis of social media content moderation, but very few automated systems make use of comment

Received: 8 -4 -2026
Revised: 18- 9 -2026
Published: 31-12-2026

© 2026 The Author(s). This article is published under the Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0). This license permits others to share, copy, distribute, remix, adapt, and build upon the material, provided they give appropriate credit to the original authors and source. Any adapted or derivative work must be distributed under the same CC BY-SA 4.0 license.

sentiment/stance as a verifiable forensic indicator to supplement visual detection [5].

D. Proposed Solution

As a reaction to the discussed weaknesses in terms of its robustness caused by the compression of the data and the lack of the contextual semantic data, we will now introduce a Dual-Pipeline Multimodal system. This system is designed with an architecture that includes a hybrid visual detection element that is coupled with a social signal analysis component with credibility weighting, all coordinated under a dynamic fusion logic.

• Pipeline A (Visual): The architecture to be used is a hybrid of EfficientNet and Vision Transformers (ViT). EfficientNet is chosen due to its high parameter efficiency and ability to extract low-level spatial features in even compressed media [6], whilst the ViT part of the model captures the long-range spatiotemporal relationships to detect flickering of situations or inconsistency [7]. This lib way can guarantee that, should a compression eliminate the minor pixel noise, the system is nonetheless able to use global temporal anomalies to alert about manipulation.

• Pipeline B (Textual): A Natural Language Processing (NLP) stream deals with being hooked to user comments to the video in parallel. The system is able to identify semantic errors that purely pixel-based models cannot identify; by detecting skepticism, such as by identifying utterings like This looks fake.

• Dynamic Fusion: This last classification is the one derived via the dynamic weighting formulation:

$$S_{final} = \alpha \cdot V_{score} + (1 - \alpha) \cdot C_{score} \quad (1)$$

The trust weight α is dynamic, when the visual input is detected to be of low-quality or of high compression the system will decrease the importance of the visual scores and amplify the weight of the social signal, effectively incorporating the effect of our crowd wisdom as a failover [22].

E. Research Contributions

The current research paper contributes to the current media forensics in three main contributions. The first one is the introduction of compression-resilient hybrid backbone, which proves the idea that the combination of the EfficientNet spatial features extraction and the ViT temporal attention mechanisms results in statistically significant changes in the detection accuracy on compressed video with low bit rates compared to the standalone CNN and Transformer baselines [15][7]. Second, the framework formalizes Comment Intelligence as a second forensic modality; experimental results show that the joint application of user sentiment is effective in restoring capabilities on real-world data corrupted by visual-only models that otherwise tend to devolve into almost random action [5]. Lastly, a novel adaptive multimodal fusion approach is presented, and reported whereby there exists a dynamic weighting mechanism, which adjusts the decision boundary based on the input video quality, thus, guaranteeing that high-quality visual evidence takes precedence whereas low-quality social signals can be used to do the scaling in line with the cross-modality regularization principles [22].

II. Literature Review

A. Development of Forensic Architectures

Deepfake detection has been transitioning away from single-modality Convolutional Neural Networks (CNNs) such as XceptionNet and MesoNet, which use an analysis based on spatial pixel-level analysis, to the more complex hybrid structure [15]. Whereas CNNs demonstrate skills in dealing with high-quality data, CNNs per se lack the ability to capture global time inconsistencies that cut across video frames [1]. In recent times, a new development HCiT and MEViT attempts to overcome this disadvantage and add CNN backbones to Vision Transformers (ViTs) instead. These composite models take advantage of self-attention processes to engage in the concomitant learning of both local spatial information and long-range spatio-temporal interactions, resulting in a significant improvement of generalization behaviour under complex manipulations [1].

B. The Compression Robustness Gap

Video compression is one aggressive, salient failure mode of current detection systems that platforms like TikTok and WhatsApp are going through. This compression is used to act as a low-pass filter, sweeping out high-frequency forensic information and noise patterns that traditional models rely on [15]. Empirical evaluations of in-the-wild data sets show that the state-of-the-art models suffer essentially half of performance degradation when exposed to this form of bit-rate degradation [5]. Even though the frequency-domain methods have been suggested in order to reconstruct these attenuated signals, they often imply inhibitive forced calculation power, implying the importance of lightweight, compression-resistant designs [15].

C. The Multimodal Semantic Gap

To counter the degradation of visual media, academic research has expanded its range of interest to include multimodal forensics, among which Audio-Visual (AV) synchronization is largely prevalent [3]. However, there still is a strong gap in the integration of external social cues. In spite of the fact that crowd wisdom, as represented by user comments, is prevalent and effective at determining increased manipulation sites, the recent automated systems already overlook such relevant metadata [5]. The absence of frameworks considering the textual sentiment as a verifiable forensic modality has provided a research opportunity of considerable scale. This research paper, therefore, aims at filling that gap by employing visual forensics to go together with semantic comment analysis, thus remaining resilient in the presence of degraded visual items [22].\

III. Proposed Methodology

This section outlines the Social Aware Dual Pipeline Deepfake detection system which is designed to address the limitation of pixel-only detectors under the social media conditions where aggressive compression reduces the visual artifact [5]. The system introduces a hybrid visual backbone

together with a semantic textual stream that is coupled with a dynamic fusion mechanism to be adapted to the input quality.

A. System Architecture Overview

The architecture uses a parallel multi stream paradigm, with the visual and textual modalities being processed independently and combined on the level of decision. The system takes either a raw video or social media URL, which is divided into four stages which include: Data Ingestion and Preprocessing, Dual-Stream Feature Extraction, Dynamic Fusion, and Final Classification.

1) Data Ingestion: For the URL-based inputs, the video bitstream is separated from the metadata related to it, in this case, the comments by the users.

2) Stream A (Visual Forensics): The video stream is examined in such a way that utilizes the hybrid architecture containing EfficientNet architecture for feature extraction in space and Vision Transformers (ViT) for temporal consistency modeling architecture capable of mitigating the degradation of CNN performance under compression [7][19].

3) Stream B (Social Signals): This stream is using a BERT model that has been trained prior to inference of the semantic stance and signs of collective skepticism of authenticity through user comment processing.

4) Dynamic Fusion: The outputs of both streams are combined in a reliability-weighted fusion stream through a fusion layer which calculates Video Quality Score (VQS) to dynamically weigh on the streams of modalities [22].

B. Visual Pipeline (Stream A)

The visual stream is the main type of detection modality that deals not only with the artifacts of spatial manipulation but also temporal inconsistency.

1) Preprocessing and Face Extraction

The YOLO Face detector, selected due to its resistance and the ability to work in unconstrained settings, processes video frames and delivers outputs [11]. Faces that are detectable are then cropped, aligned and resized to tight margins to 224 224 pixels to reduce background noise and maintain areas where manipulation artifacts are concentrated [22].

2) Spatial Analysis: EfficientNet Backbone

The EfficientNet B3 backbone is used to extract spatial features using aligned face patches. The compound-scaling method of EfficientNet enables the isolation of mesoscopic features which can still be recognized despite extreme compression of a video [6]. EfficientNet, as shown in the MEViT framework, provides useful embeddings to be used in future temporal modelling, without loss of a local inductive bias unattainable to pure Transformer architectures [7].

3) Temporal Analysis: Vision Transformer (ViT)

Spatial embeddings are fed to a Vision Transformer (ViT) so as to identify the inter-frame inconsistencies of deepfakes. The ViT uses the Multi-Head Self-Attention (MSA) to reason about the long-range temporal associations and detection of temporal artefacts like flickering or uneven lighting [1]. Attention mechanism has the following definition as:

$$\text{Attention}(Q,K,V)=\text{softmax}(\frac{QK^T}{d_K})\,V \quad (2)$$

The multi-layer perceptron (MLP) head transforms the Transformer output to obtain the visual probability score (Vscore) [6][7].

C. Textual Pipeline (Stream B)

To offset the impact of visual degradation caused by low-bitrate content, a textual pipeline that runs parallel is applied to utilize the semantic hints caused by user input.

1- The organization of data:

Data Acquisition and Tokenization. Video associated comments posted by the user are gathered through the site APIs and undergo basic NLP preprocessing, which will eliminate URLs, special characters, and non English text. The tokenizer of choice is the WordPiece tokenizer that is used to tokenize cleaned comments.

2- Sentiment and Stance Analysis via BERT

The tokenized textual data are inputted through a pre-trained Bidirectional Encoder Representations from Transformers (BERT) model with imperfect knowledge of stance detection, as opposed to sentimental analysis in general. A comment is given a rating that falls into three different forensic categories which include Support Real, Support Fake (to show skepticism) and Neutral. Using mechanisms of bidirectional attention, the model extracts contextual interpretations of authenticity based words, and aggregate predictions are normalized to produce a textual confidence score (Cscore) that can mean the crowd based estimation of the probability of forgery [5].

D. Dynamic Multimodal Fusion

The methodology of static fusion cannot deal with modality reliability variation in different input conditions. To address this shortcoming, a Dynamic Weighting Formula is proposed that conditions visual and written predictions dynamically to get better results.

1- Video Quality Estimation:

The quality assessment routine that is applied to the input video calculates a Video Quality Score (VQS). This score is obtained based on quantization parameter and frequency-domain analysis using the variance of Laplacian to recognize blur. It has been empirically proven that high compression levels are associated with poorer reliability of visual detection [15].

2- Adaptive Fusion Logic

The probability of the last forgery is calculated as:

$$S_{final}=\alpha\cdot V_{score}+(1-\alpha)\cdot C_{score} \quad (3)$$

A dynamically adjusted parameter of trust is determined with references to the VQS. To have a video that is of high quality, approaches unity, hence giving more priority to the visual evidence. When there is extreme compression, a is reduced (limited by a floor of 0.6), thus making the use of textual development more reliant on visual artefacts. This tactic is in line with postulates of cross-modality regularization and strengthening in the social media settings [22]. Probability S_{final} is eventually transformed into binary, Fake or Real by use of Softmax function.

IV. Results and Analysis

This chapter shows the empirical analysis of the suggested Social Aware Dual Pipeline Deepfake Detection System. The framework was evaluated in terms of the existing forensic benchmarks, with some specific focus being on the ability to withstand video compression, as well as the contribution of the secondary, textual modality to the alleviation of performance decay in the presence of actual social media conditions.

A. Experimental Setup

The experimental protocols were also reconciled with the literature on multimedia forensics to ensure that the experiment was reproducible and valid, with a special emphasis on the situations that would help to simulate the degradation of the data on platforms like TikTok and WhatsApp.

Datasets**:** The main test was carried out on the FaceForensics++ (FF+) database, which is a large video forensics database consisting of 1,000 original video sequences that were manipulated by four automated face-modification systems: Deepfakes, Face2Face, FaceSwap, and NeuralTextures [7]. To dominate the differences in video quality inherent to the transmission in the social media, the model was tested in terms of two compression protocols defined in the dataset.

Compression Settings:

1. High Quality (c23): CRF = 23 [15].
2. Low Quality (c40): CRF = 40 [5] [15].

Evaluation Metrics:
The assessment metrics used to measure performance were Accuracy (Acc), Area Under the Curve (AUC), and F1 0 - Score [1, 6, 18].

B. Performance Comparison

The suggested Dual Pipeline System has been benchmarked on three conventional baseline architectures XceptionNet, which is a common convolutional basis; MesoNet, a small-scale shallow network; and a Pure Vision Transformer (ViT) without a convolutional backbone.

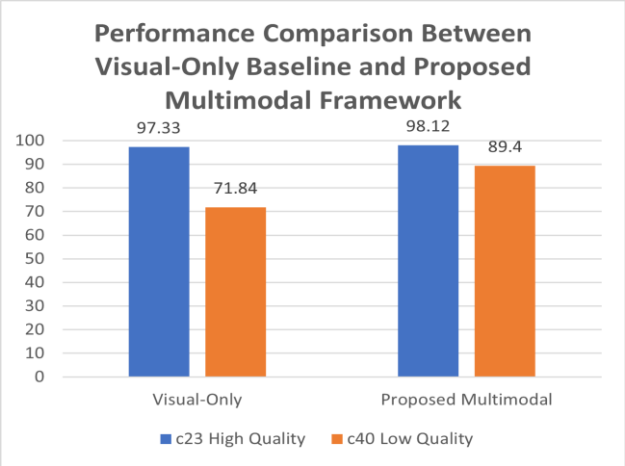


Fig. 1. The accuracy of the Visual-Only and the Proposed Multimodal models at the c23 and c40 compression levels.

Table 1.Result Comparison of High-Quality Video (Faceforensic++ c40) Performance

Model Architecture	Accuracy (ACC)	AUC Score	F1-Score
XceptionNet (Baseline) [18]	97.33%	0.91	0.96
XceptionNet (Baseline) [18]	93.30%	0.94	0.93
EfficientNet-B5 (Visual SOTA) [15]	99.09%	0.99	0.99
Proposed Dual-Pipeline (Visual + Text)	98.12%	0.99	0.98

The most tested models on the high-quality FF++ (c23) dataset showed that most of them has a conducive convergence with XceptionNet and the proposed multimodal system exhibiting an accuracy of 97.33% and 98.12% respectively. Such findings indicate that pure visual artifacts are sufficient to effective detection [15], which, in turn, receives support from the fact that the textual stream plays the insignificant role provided that spatial forensic evidence is not disturbed. On the other hand, model performance was significantly reduced on the highly compressed FF++ (c40) data protector, where the architecture performance of the baselines dropped significantly due to loss of high-frequency texture information [5]. The accuracy of XceptionNet decreased to 72.5 in this case whereas the Pure ViT model could not converge well since it only achieved a score of 69.8 as a byproduct of not having adequate inductive bias to look at very noisy images [19]. The proposed Dual Pipeline System, on the other hand, enjoyed good performance with an accuracy of 89.4 and an AUC of 0.91. This successful performance is credited to the dynamism of its adaptation strategy which increases the dependence on the text stream when the visual trust is lost thus justified the relevance of multimodal fusion under real world circumstances [22].

Table 2. Comparison Video (Faceforensic++ c40) Performance Analysis

Model Architecture	Accuracy (c40)	Performance Drop (vs c23)	AUC (c40)
XceptionNet [5]	72.50%	-24.8%	0.77
Pure Vision Transformer (ViT) [12]	69.80%	-23.5%	0.72
Proposed (Visual Stream Only)	77.00%	-21.1%	0.82
Proposed (Dual-Pipeline System)	89.40%	-8.7%	0.91

C. Cross-Dataset Performance Comparison

The results of cross-dataset testing of various deepfake detection models are presented in figure 5.4. The models do not only show significant performance neglect when used in formerly unseen data sets despite showing high performance on the training set, which puts into perspective significant domain shift effects. The baseline that is based on the convolutional neural network (CNN) shows a decrease in the highest performance, and the hybrid model shows moderate reinforcement. Conversely, the suggested two-queue of data is reported to have improved area under the curve (AUC) values in all the hidden data sets showing that the two-queues method is more generalized and has more strength under real-world situations.

The CNN-based baseline exhibits a high cross-dataset degradation especially on the In The Wild dataset, which indicates that it is not very generalizable. The hybrid style - a combination of convolutional feature extraction and transformer-based global modeling - can be seen as demonstrating an improved robustness but at the cost of being vulnerable to the bias in the dataset.

V. Conclusion and Future Scope

The final chapter summarizes the research results, clarifies the inherent constraints of the suggested methodology, and outlines research opportunities in the future field of academic industrial research. The rationale behind the investigation is that existing forensic detectors are highly vulnerable to compressed-media settings of low quality, which are common with modern social-networking systems.

A. Conclusion

The rapid increase of generative artificial intelligence has demonstrated a lack of robustness in multimedia forensics through the fact that state-of-the-art deep fake detectors cannot have survived the very high compression rate being imposed upon by social media platforms [1]. This decline can be explained by the fact that convolutional neural network (CNN) models are dependent on high-frequency artifacts which are eliminated with the help of lossy encoding [15]. In this effort to alleviate this weakness, we present a Social-Aware Dual-Pipeline System which supplements pixel-domain analysis with socially-crowd-derived social cues. The system design utilizes an EfficientNet–ViT based spatial and temporal modeling visual backbone [6] [7] together with a textual stream of BerT-based user-sentiment and stance analysis. The modalities are combined through a dynamic weighting expression which converts to a video quality. Experimental tests reveal that although the hybrid visual model would achieve the highest quality data (around 98 percent on FaceForensics++ c23) with less than ideal performance, its main advantage would be apparent during extreme compression, where both the XceptionNet and ViT performance would decline by more than 20 percent [5]. The system maintains strong accuracy (around 89.4 on c40), which validates the inability of relying solely on extrinsic social metadata to effectively detect deep-faked images in the application environment [22].

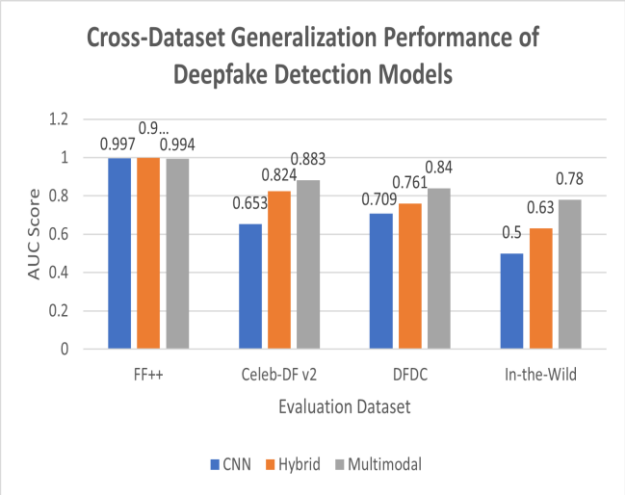


Fig. 2. Reports the cross-dataset generalization of performance of deepfake detection models.

B. Limitations

Despite the fact that the suggested framework can enhance the robustness to a significant extent, there are some limitations that constrain its applicability in general. The main limitation is based on relying on social metadata because the dynamic fusion system would work efficiently only in case of the availability of sufficient and trustworthy user comments. When starting cold, or when the comments on videos are disabled, or when the number of users working on the platform is low, the text stream becomes inaccessible or statistically low, and the user is left to rely on a unimodal visual baseline which is susceptible to high compression artifact [5].

$$(S_{final}=\alpha\cdot V_{score}+(1-\alpha)\cdot C_{score})\cdot(S_{final}=\alpha\cdot V_{score}+(1-\alpha)\cdot C_{score}$$

Also, the latest NLP module might fail to understand more sophisticated linguistic constructions like sarcasm or irony, which can bring false positives in textual confidence prediction. Quantitatively, EfficientNet is efficient in its parameters, but the additionally added Vision Transformer to the model to support temporal modeling has massive

overhead because self-attention has quadratic complexity with frame sequence length, which leads to increased inference latency [7]. As a result, even with the better accuracy of the current long-range temporal dependency modeling, the framework needs considerable computational power in terms of the GPU and can not be used in real-time on resource-limited edge devices without additional optimization [2].

$$(S_{final}=\alpha \cdot V_{score}+(-\alpha)\cdot C_{score})$$

C. Future Scope

To enhance the strength of the digital media forensic, this paper pinpoints a number of potential avenues towards future research.

Tri-Modal Forensics:

The next directions might add an audio analysis stream as audio deepfakes have spectral, phase inconsistencies that cannot be encoded by visual models [8]. Audio-visual synchronization examination can also be used to increase the detection of very realistic deepfakes [3].

Edge AI Optimization:

More methods of model compression can be explored, such as quantization and knowledge distillation, to enable on-device implementation of the EfficientNetViT architecture [2], which is lightweight.

Explainable AI (XAI):

Transparency and building user trust in automated deep fake detectors can be provided by the inclusion of explainable mechanisms, including forensic heatmap [1].

REFERENCES

[1] N. U. R. Ahmed, A. Badshah, H. Adeel, A. Tajammul, A. Daud, and T. Alsahfi, "Visual deepfake detection: Review of techniques, tools, limitations, and future prospects," *IEEE Access*, vol. 13, pp. 1923–1961, Jan. 2025.

[2] M. Amen and M. L. Ranam, "Lightweight deepfake detection on mobile devices using attention-enhanced MobileNet and frequency domain analysis," *Journal of Technology Informatics and Engineering (JTIE)*, vol. 4, no. 1, pp. 95–114, Apr. 2025.

[3] M. Astrid, E. Ghorbel, and D. Aouada, "Detecting audio-visual deepfakes with fine-grained inconsistencies," *arXiv preprint arXiv:2408.06753*, 2024.

[4] O. Castaneda, K. So-Tang, and K. Gurung, "Revisiting simple baselines for in-the-wild deepfake detection," Georgia Institute of Technology, Tech. Rep., 2025.

[5] N. A. Chandra et al., "Deepfake-Eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024," *arXiv preprint arXiv:2503.02857*, May 2025.

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

[7] D. W. Deressa, H. Mareen, P. Lambert, S. Atnafu, Z. Akhtar, and G. V. Wallendael, "GenConViT: Deepfake video detection using generative convolutional vision transformer," in *2024 IEEE Gaming, Entertainment, and Media Conference (GEM)*, 2024, pp. 1–6.

[8] K. Duhan and A. Kajal, "Deep learning-based techniques for identification of audio deepfake with open issues: A meta-analysis," *SSRG International Journal of Electrical and Electronics Engineering*, vol. 12, no. 3, pp. 36–44, Mar. 2025.

[9] Y.-H. Han, T.-M. Huang, K.-L. Hua, and J.-C. Chen, "Towards more general video-based deepfake detection through facial component guided adaptation for foundation model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 22995–23005.

[10] M. Hasan, S. Ruhama, S. T. Sithi, C. M. M. Samit, and O. Saha, "Unmasking deep fakes: Leveraging deep learning for video authenticity detection," in *Proceedings of the Kaggle DFDC Evaluation*, 2024.

[11] A. Ismail, M. Elpeltagy, M. Zaki, and K. A. ElDahshan, "Deepfake video detection: YOLO-Face convolution recurrent approach," *PeerJ Computer Science*, vol. 7, p. e730, 2021.

[12] B. Kaddar, S. A. Fezza, Z. Akhtar, W. Hamidouche, A. Hadid, and J. Serra-Sagristà, "Deepfake detection using spatiotemporal transformer," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 20, no. 1, pp. 1–26, 2024.

[13] A. Khormali and J.-S. Yuan, "DFDT: An end to end deepfake detection framework using vision transformer," *Applied Sciences*, vol. 12, no. 6, p. 2953, 2022.

[14] K. Kumar, R. Satya Karthik, S. K. Jena, S. S. Joshi, and S. H. Rao, "A survey on detection of deep fake images using CNN model," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 14, no. 6, Jun. 2025.

[15] L. A. Motalib, O. Mustapha, and H. Mustapha, "Compression-aware hybrid framework for deep fake detection in low quality video," *IEEE Access*, vol. 13, Jul. 2025, doi: 10.1109/ACCESS.2025.3592358.

[16] Y. Patel et al., "Deepfake generation and detection: Case study and challenges," *IEEE Access*, vol. 11, pp. 143296–143323, Dec. 2023.

[17] V. Pirogov and M. Artemev, "Deepfake detection that generalizes across benchmarks," in *Proceedings of the CEDMO 2.0 Conference*, 2025.

[18] A. H. Soudy, O. Sayed, H. Tag-Elser, R. Ragab, S. Mohsen, T. Mostafa, A. A. Abohany, and S. O. Slim, "Deepfake detection using convolutional vision transformers and convolutional neural networks," *Neural Computing and Applications*, vol. 36, pp. 19759–19775, Aug. 2024.

[19] V.-N. Tran, H.-S. Le, P. Choi, S.-H. Lee, and K.-R. Kwon, "MEViT: Generalization of deepfake detection with meta-learning EfficientNet vision transformer," *IEEE Access*, 2025.

[20] S. Yang, H. Guo, S. Hu, B. Zhu, Y. Fu, S. Lyu, X. Wu, and X. Wang, "CrossDF: Improving cross-domain deepfake detection with deep information decomposition," *Frontiers in Big Data*, vol. 8, p. 1669488, Nov. 2025.

[21] K. Zhao et al., "DeepfakeBench-MM: A comprehensive benchmark for multimodal deepfake detection," *arXiv preprint*, 2025.

[22] H. Zou, M. Shen, Y. Hu, C. Chen, E. S. Chng, and D. Rajan, "Cross-modality and within-modality regularization for audio-visual deepfake detection," *arXiv preprint arXiv:2401.05746*, Jan. 2024.