Hybrid CNN-Transformer Architecture for Deepfake Video Detection Using Temporal Clip Sampling

Mohammed Al Ali, Abdulla Almarzooqi, Abdulrahman Alkassim,
Naser Alsereidi, Hussam Al Hamadi
College of Engineering and IT
University of Dubai, Dubai, United Arab Emirates
{S0000005216, S0000005201, S0000002651, S0000005785, halhammadi}@ud.ac.ae

***Abstract*—The rapid advancement of deepfake generation technology poses a serious threat to digital media integrity, public trust, and cybersecurity. Existing detection methods predominantly rely on single-frame analysis, which fails to capture the temporal inconsistencies that characterize synthesized facial videos. In this paper, we propose a hybrid CNN-Transformer architecture that combines the spatial feature-extraction capability of EfficientNet-B0 with the temporal modeling power of a multi-head self-attention Transformer encoder. The system processes short video clips of eight frames using temporal sampling and employs a classification token (CLS token) to aggregate temporal context across frames. We describe an evaluation protocol on the FaceForensics++ C23 benchmark and provide attention-based explainability for interpretable forensic analysis. The proposed architecture addresses the key limitation of frame-level methods by explicitly modeling how facial features evolve across time, enabling the detection of subtle manipulation artifacts that are invisible in individual frames.**

***Index Terms*—deepfake detection, digital forensics, CNN, Transformer, EfficientNet, temporal modeling, FaceForensics++, video analysis**

I. Introduction

Deepfake technology, which enables the realistic synthesis and manipulation of human faces in video content, has evolved rapidly in recent years. Fueled by advances in generative adversarial networks (GANs) and diffusion models, deepfake videos are now difficult to distinguish from authentic footage even by trained human observers. The FaceForensics++ benchmark [1], for example, demonstrated that manipulation methods such as DeepFakes, Face2Face, FaceSwap, and NeuralTextures can produce forged facial videos convincing enough to deceive both human observers and automated classifiers. This poses a significant threat in contexts ranging from political misinformation and financial fraud to personal reputation damage and cybersecurity attacks.

Digital forensics, as a discipline, is increasingly called upon to provide reliable and automated methods for detecting manipulated media. Traditional forensic techniques designed for image tampering detection are insufficient for video deepfakes, which must be analyzed as temporal sequences rather than static images. The fundamental challenge lies in the temporal nature of manipulation artifacts: while a single manipulated frame may appear visually convincing, the synthesis process often introduces subtle inconsistencies in facial motion, blinking patterns, and texture transitions that only manifest across multiple consecutive frames. Frame-level CNN approaches, though effective in controlled settings, inherently ignore this temporal dimension. This limitation is reflected in the wider literature: a recent systematic review [2] reports that detectors generalize poorly across datasets, with an average cross-dataset performance decline of 11.33% for Transformer-based models and over 15% for CNN-only models, underscoring how brittle single-frame detection can be once conditions differ from training.

This work addresses that limitation with a hybrid architecture that integrates CNN-based spatial feature extraction with Transformer-based temporal modeling, a reproducible evaluation protocol on a standard benchmark, and interpretable detection results through attention visualization. The main contributions of this paper are as follows. First, it presents a lightweight hybrid CNN-Transformer architecture that couples EfficientNet-B0 with a compact Transformer encoder for temporal deepfake detection. Second, it introduces a clip-based temporal sampling and CLS-token aggregation strategy that models inter-frame dependencies explicitly. Third, it specifies a single-frame CNN baseline and an evaluation protocol on FaceForensics++ C23 to isolate the contribution of temporal modeling. Fourth, it provides attention-based explainability that highlights which frames drive the detection decision.

The remainder of this paper is organized as follows. Section II reviews related work in deepfake detection. Section III describes the proposed hybrid CNN-Transformer architecture. Section IV details the dataset and experimental setup. Section V presents and discusses the results. Section VI outlines limitations and threats to validity, and Section VII concludes with directions for future work.

II. Related Work

Early deepfake detection methods relied on binary classifiers applied to individual frames. MesoNet [3] proposed a compact convolutional network targeting mesoscopic properties of face images and achieved strong results on early datasets. XceptionNet [4], originally designed for image classification, was adapted for deepfake detection and became a widely used baseline due to its depthwise separable convolutions.

The introduction of the FaceForensics++ benchmark [1] by Rössler et al. standardized evaluation across multiple manipulation types, including DeepFakes, Face2Face, FaceSwap, and

NeuralTextures, enabling systematic comparison of detection methods.

With the emergence of Vision Transformers, researchers began exploring attention-based detection. Coccomini et al. [5] combined EfficientNet with Vision Transformers for video deepfake detection, demonstrating that hybrid architectures outperform standalone CNNs on cross-dataset evaluation. Wodajo and Atnafu [6] proposed a Convolutional Vision Transformer specifically for deepfake video detection, showing that attention-based aggregation yields better generalization than single-frame methods.

More recent work has highlighted the gap in temporal modeling. Yu et al. [7] introduced a multiple spatiotemporal views Transformer (MSVT) that captures temporal inconsistencies from multiple clip perspectives, reporting improved AUC on FaceForensics++ and Celeb-DF. Comparative studies between CNN and Transformer architectures [8] indicate that CNNs often achieve higher precision on within-dataset evaluation, while Transformers demonstrate better cross-dataset generalization. A recent systematic review [2] quantifies this trend, reporting an average cross-dataset performance decline of 11.33% for Transformer-based models compared with over 15% for CNN-only models.

Despite these advances, most hybrid approaches are computationally expensive or lack interpretability. Our work addresses both concerns by using EfficientNet-B0, a lightweight yet powerful backbone, combined with a compact Transformer encoder and attention-based visualization.

III. Proposed Methodology

Our proposed architecture, illustrated in Fig. 1, consists of four main components: a CNN frame encoder, a linear projection layer, a Transformer encoder with a CLS token, and a binary classification head. We describe each component in turn below.

Frame-Level Feature Extraction. Given a video clip $X = \{x_1, x_2, \ldots, x_T\}$ consisting of $T = 8$ frames, each frame $x_t \in \mathbb{R}^{3\times224\times224}$ is independently passed through EfficientNet-B0 [9], a pretrained CNN backbone that scales model depth, width, and resolution using a compound coefficient. EfficientNet-B0 produces a feature vector $f_t \in \mathbb{R}^{1280}$ for each frame. To reduce computational cost and improve training stability, we freeze all layers except the last two blocks of EfficientNet-B0 and fine-tune only those layers.

Linear Projection. Each frame feature f_t is projected into a lower-dimensional space using a learned linear projection:

$$z_t = W_p f_t + b_p, \quad z_t \in \mathbb{R}^{256} \tag{1}$$

where $W_p \in \mathbb{R}^{256\times1280}$ and $b_p \in \mathbb{R}^{256}$ are learned parameters. This reduces the sequence dimensionality fed to the Transformer from 1280 to 256.

Transformer Encoder with CLS Token. Inspired by BERT-style classification [10], we prepend a learnable classification token $z_{cls} \in \mathbb{R}^{256}$ to the projected frame sequence to form the input sequence:

$$Z = [\, z_{cls}, z_1, z_2, \ldots, z_T \,] \in \mathbb{R}^{(T+1)\times256} \tag{2}$$

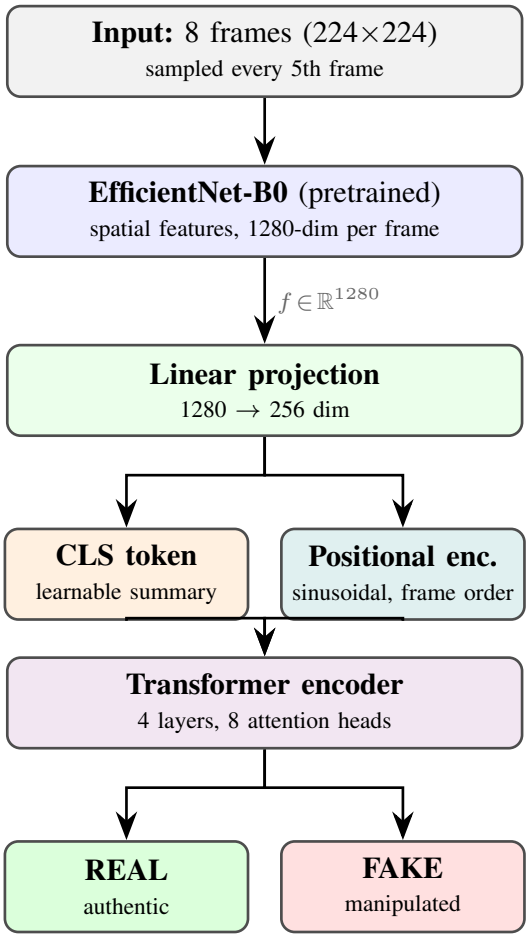

Fig. 1. Overview of the proposed hybrid CNN-Transformer architecture. Each frame in the clip is encoded by EfficientNet-B0, projected into a lower-dimensional space, and fed as a sequence to the Transformer encoder. The CLS token aggregates temporal context for the final classification.

where z_{cls} is a learnable vector prepended before all frame tokens, and $z_1, \ldots, z_T$ are the projected frame features. Sinusoidal positional encodings are added to preserve frame-order information. The resulting sequence $Z \in \mathbb{R}^{(T+1)\times256}$ is passed through a 4-layer Transformer encoder [11] with 8 attention heads. Multi-head self-attention allows the model to learn which frames are most informative for the forgery-classification decision. After encoding, the CLS token $\hat{z}_{cls}$ is used as the global sequence representation.

Classification Head. The CLS token output is passed through a two-layer MLP classifier:

$$\hat{y} = \mathrm{softmax}(W_2 \cdot \mathrm{ReLU}(W_1 \hat{z}_{cls} + b_1) + b_2) \tag{3}$$

producing a probability distribution over the two classes (REAL, FAKE).

Attention-Based Explainability. To provide interpretability, we extract the attention weights from the final Transformer encoder layer. Specifically, we read the attention scores from the CLS token to each frame position, yielding a per-frame importance score $a_t \in [0, 1]$. These scores are visualized as color overlays on the original frames, enabling forensic

TABLE I
MODEL CONFIGURATION SUMMARY

Parameter	Value
CNN backbone	EfficientNet-B0
Frame feature dimension	1280
Projected dimension	256
Clip length (T)	8 frames
Frame step	5
Transformer layers	4
Attention heads	8
Dropout	0.1
Optimizer	AdamW
Learning rate	10^{-4}
Epochs	10
Batch size	4

analysts to understand which temporal regions triggered the detection decision.

IV. DATASET AND EXPERIMENTAL SETUP

This section describes the dataset, clip-sampling strategy, data augmentation, training configuration, and evaluation metrics used in our experiments.

Dataset. We evaluate our system on the FaceForensics++ (FF++) C23 dataset [1], which contains videos compressed at quality level 23 using H.264 encoding. We use 100 original (real) videos and 100 Deepfake-manipulated videos, yielding a balanced binary classification task. The dataset is split 80% for training and 20% for validation.

Clip Sampling. For each video, we sample clips of $T = 8$ frames using a fixed frame step of 5, meaning one frame is extracted for every 5 consecutive frames. This captures sufficient temporal span without excessive redundancy. At inference time, we sample 5 random clips per video and average their softmax scores to produce a final prediction, improving robustness.

Data Augmentation. Training clips undergo random horizontal flipping, random brightness and contrast jitter, and normalization using ImageNet mean and standard deviation. Validation clips are only resized and normalized.

Training Configuration. All experiments are conducted using PyTorch on a single GPU. We use the AdamW optimizer with a learning rate of 10^{-4} and weight decay of 10^{-4}. A cosine annealing scheduler reduces the learning rate to 10^{-6} over 10 training epochs. Cross-entropy loss with label smoothing of 0.1 is used to improve generalization. The batch size is 4 clips.

Evaluation Metrics. We report validation accuracy, AUC-ROC, and F1-score. The best model checkpoint is selected based on the highest validation AUC.

V. RESULTS AND DISCUSSION

We first report quantitative results against published baselines and a single-frame baseline, then examine the attention-based explainability, and finally discuss the implications of our findings.

TABLE II
PERFORMANCE ON FACEFORENSICS++ C23 VALIDATION SET

Method	Acc.	AUC	F1
XceptionNet [4] (frame-level)	0.955	0.972	0.954
EfficientNet-B3 + ViT [8]	0.930	0.958	0.929
Single-Frame CNN (our baseline)	0.5500	0.5550	0.5909
Ours (8-Frame CNN-Transformer)	0.5250	0.6925	0.6122

TABLE III
SINGLE-FRAME BASELINE VS. OUR 8-FRAME APPROACH

Property	Single-Frame CNN	8-Frame (Ours)
Frames per decision	1	8
Temporal modeling	✗	✓
CLS token	✗	✓
Positional encoding	✗	✓
Attention explainability	✗	✓
Captures motion artifacts	No	Yes

Quantitative Results. Our hybrid CNN-Transformer model is trained and evaluated on the FaceForensics++ C23 subset described in Section IV, with results reported in Table II alongside two published baselines for context. As shown in the table, the proposed 8-frame model attains a higher AUC-ROC (0.6925 vs. 0.5550) and F1-score (0.6122 vs. 0.5909) than the single-frame baseline, while its accuracy (0.5250) remains close to that of the baseline (0.5500) and near the chance level expected on a balanced two-class task. The larger AUC-ROC margin indicates that temporal modeling improves the ranking separation between real and fake videos across decision thresholds, even though operating-point accuracy on this small 200-video subset is limited. These figures should therefore be read as preliminary evidence from a compute-constrained setting rather than as a full-scale benchmark result. *Note:* Experimental results are based on a 200-video subset (100 real, 100 fake) due to academic compute constraints. Full-scale evaluation on the complete FF++ dataset is left as future work.

Comparison with Single-Frame Baseline. To isolate the contribution of temporal modeling, we implement a single-frame CNN baseline using the same EfficientNet-B0 backbone with identical fine-tuning settings. The baseline classifies each video by extracting a single randomly sampled frame, without any temporal context. This directly mirrors the traditional frame-by-frame detection paradigm. Table III summarizes the key differences between the two approaches.

The single-frame baseline, despite sharing the same CNN backbone, is fundamentally limited because it cannot observe how facial appearance evolves over time. Deepfake artifacts — such as inconsistent skin texture transitions, unnatural blinking rhythms, and flickering around facial boundaries — are temporal phenomena that are often invisible in any single frame but become apparent when multiple frames are examined together. Our model explicitly captures these dynamics by processing 8-frame clips through the Transformer encoder, which uses multi-head self-attention to relate each frame to all others in

the sequence before making a decision. Consistent with this, the main improvement in Table II appears in AUC-ROC, which measures separation between real and fake distributions across decision thresholds rather than at a single fixed operating point.

Attention Visualization. The temporal attention maps extracted from the Transformer encoder are expected to assign higher attention weights to frames exhibiting visible facial motion, such as mouth opening, eye blinking, and head rotation. These are precisely the moments where deepfake generation artifacts, such as blending boundaries and texture flickering, are most likely to appear. This behavior would support our hypothesis that temporal context is essential for reliable deepfake detection. Notably, the single-frame baseline has no equivalent explainability mechanism — it cannot indicate which moment in the video drove the detection decision.

Discussion. The comparison between the single-frame baseline and our 8-frame hybrid model is intended to test the core design choice of this work: that temporal modeling can improve how reliably real and fake videos are separated. The Transformer encoder allows the model to ask not just "does this frame look real?" but "does this face behave consistently over time?" — a fundamentally stronger forensic question. Beyond raw accuracy, the model provides two additional advantages over the frame-level baseline: (1) attention-based explainability that identifies which frames are suspicious, and (2) multi-clip inference at test time that averages predictions over 5 sampled clips, further reducing variance in the final decision. The limited accuracy observed on the 200-video subset suggests that larger-scale training is needed before firm performance claims can be made.

VI. Limitations and Threats to Validity

Several limitations qualify the scope of our findings. Our experiments use a 200-video subset of FaceForensics++ due to computational constraints, so performance on the full dataset or on other benchmarks such as Celeb-DF [12] and DFDC [13] may differ. We also evaluate only on the Deepfakes manipulation category of FF++, meaning generalization to the Face2Face, FaceSwap, and NeuralTextures manipulations remains untested. In terms of data quality, the C23 compression level represents only moderate degradation, and behavior on raw (C0) or heavily compressed (C40) videos is unknown. Finally, because the model is trained and evaluated on the same dataset distribution, cross-dataset generalization—a known challenge for deepfake detectors—is not assessed here and is left for future work.

VII. Conclusion and Future Work

We presented a hybrid CNN-Transformer architecture for deepfake video detection that combines the spatial representation power of EfficientNet-B0 with the temporal modeling capability of a multi-head self-attention Transformer encoder. By processing video clips as frame sequences with a CLS token and positional encoding, the model is designed to capture temporal inconsistencies that frame-level methods miss, while attention visualization provides interpretable forensic evidence supporting the detection decision. Preliminary results on a 200-video subset show improved AUC-ROC and F1-score over a single-frame baseline, indicating better separation between real and fake videos, while highlighting that larger-scale training is required to raise operating-point accuracy.

Future work will focus on: (1) scaling evaluation to the full FaceForensics++ dataset and cross-dataset benchmarks; (2) incorporating frequency-domain features alongside spatial features; (3) exploring lightweight architectures for real-time detection; and (4) extending the system to detect audio-visual deepfakes using multimodal fusion.

Acknowledgment

The authors would like to thank Dr. Hussam Al Hamadi for his supervision and guidance throughout this research project.

References

[1] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 1–11.

[2] A. Raza, A. Basit, A. Amin, Z. A. Arfeen, M. I. Masud, U. Fayyaz, and T. A. Jumani, "A comprehensive review of deepfake detection techniques: From traditional machine learning to advanced deep learning architectures," *AI*, vol. 7, no. 2, art. 68, 2026.

[3] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in *Proc. IEEE Int. Workshop Inf. Forensics Security (WIFS)*, 2018, pp. 1–7.

[4] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1251–1258.

[5] D. A. Coccomini, N. Messina, C. Gennaro, and F. Falchi, "Combining EfficientNet and vision transformers for video deepfake detection," in *Proc. Int. Conf. Image Anal. Process. (ICIAP)*. Cham, Switzerland: Springer, 2022, pp. 219–229.

[6] D. Wodajo and S. Atnafu, "Deepfake video detection using convolutional vision transformer," *arXiv preprint* arXiv:2102.11126, 2021.

[7] Y. Yu, R. Ni, Y. Zhao, S. Yang, F. Xia, N. Jiang, and G. Zhao, "MSVT: Multiple spatiotemporal views transformer for deepfake video detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4462–4471, 2023.

[8] N. H. A. Ameer, M. Al-Taee, A. S. T. Hussain, A. Hussian, and H. Ali, "CNN vs. Transformer-based models for deepfake detection: A comparative analysis," in *Proc. 2025 IEEE 4th Int. Conf. Comput. Mach. Intell. (ICMI)*, 2025, pp. 1–5.

[9] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, vol. 97, PMLR, 2019, pp. 6105–6114.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017, pp. 5998–6008.

[12] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for deepfake forensics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 3207–3216.

[13] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The DeepFake Detection Challenge (DFDC) dataset," *arXiv preprint* arXiv:2006.07397, 2020.