Detecting Low-Quality Deepfake Videos Using 3D Residual Vision Transformer

Amna Saga[1], Lili N. A[2], Fatimah Khalid[3], Nor Fazlida Mohd Sani[4],
Hussna E. M. Abdalla[5], Zulfahmi Syahputra[6], Rian Farta Wijaya[7]
Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang, Malaysia[1, 2, 3, 4]
Department of Computer and Communication System Engineering, Universiti Putra Malaysia, Serdang, Malaysia[5]
Universitas Pembangunan Panca Budi, Medan, Indonesia[6, 7]

***Abstract*—The rapid evolution of deep generative models has facilitated the creation of "Deepfakes", enabling the synthesis of hyper-realistic facial manipulations that threaten the trustworthiness of digital media. While forensic countermeasures have been developed to identify these forgeries, deepfake detection in real-world scenarios is severely hampered by video compression artifacts, which often obscure the subtle pixel-level traces exploited by conventional Convolutional Neural Networks (CNNs). This study introduces a robust detection framework designed specifically to withstand the aggressive compression inherent to social media dissemination. We present a hybrid 3D architecture that integrates the local spatiotemporal feature extraction capabilities of a 3D-ResNet-50 backbone with the global context modeling of a temporal Video Vision Transformer. Unlike frame-based or joint spatiotemporal attention approaches, the proposed model performs fully video-level reasoning and utilizes a factorized self-attention mechanism to decouple spatial and temporal modeling, thereby preserving stable temporal cues under compression while minimizing computational costs. Experimental results on the compressed protocols of the FaceForensics++ dataset as well as Celeb-DF-v2 and DFDC datasets, including cross-dataset generalization evaluation, validate the efficacy of this design, demonstrating that our method achieves superior detection accuracy and generalization compared to existing baselines, particularly on low-quality inputs.**

Keywords*—Deepfake detection; compressed deepfake videos; low-quality deepfakes; 3D convolutional neural networks; Video Vision Transformer

I. INTRODUCTION

Recent advances in deep generative modeling have enabled the creation of highly realistic synthetic audiovisual content, giving rise to what are commonly referred to as deepfakes [1]. A deepfake is a piece of media; typically a video, image, or audio clip, created or manipulated using deep neural networks to convincingly imitate real individuals or events. Although the term has become strongly associated with disinformation and online abuse, the underlying technologies offer several legitimate and beneficial applications [2]. These include privacy-preserving data generation for machine learning, synthetic training datasets for rare-event detection, improved dubbing in film production, educational reenactments, and assistive technologies for individuals with speech impairments.

Despite these positive uses, deepfakes pose significant societal and security risks. High-fidelity manipulations can be leveraged to spread political misinformation, facilitate identity or financial fraud, generate non-consensual explicit content, or erode public trust in digital media [3]. As generative models continue to improve in resolution, coherence, and computational efficiency, distinguishing authentic content from manipulated media becomes increasingly challenging. The pervasive dissemination of deepfake videos across social platforms exacerbates this issue, contributing to large-scale misinformation campaigns and undermining the reliability of digital communication ecosystems.

In response to this adversarial landscape, the forensic community has developed a variety of detection algorithms designed to identify the subtle, distinct trace anomalies left by generative models. However, a critical disparity exists between the performance of these detectors in controlled environments and their efficacy in real-world deployment [4]. While deepfake generation has become increasingly sophisticated, often outpacing conventional detection cues, the primary challenge in practical application lies in the transmission medium itself. To accommodate bandwidth constraints and storage limitations, online distribution platforms routinely subject video content to aggressive lossy compression standard H.264. This compression process introduces a complex, two-fold impediment to forensic analysis. First, the quantization inherent in lossy compression acts as a low-pass filter, effectively smoothing out the high-frequency noise patterns and subtle pixel-level artifacts that many detectors rely upon as fingerprints of manipulation. Second, the compression algorithms introduce their own structural noise, such as blocking artifacts and ringing, which can mask manipulation traces or mimic them, leading to increased false positives. Consequently, the robustness of existing state-of-the-art detectors degrades precipitously when applied to low-quality, highly compressed media [5]. Addressing this vulnerability created by compression, remains one of the most pressing open challenges in multimedia forensics.

This study presents a fully video-level approach for robust deepfake detection in highly compressed videos, a setting in which conventional frame-based detectors often fail due to the suppression of spatial and frequency-domain forensic cues. While most existing detection methods process videos on a frame-by-frame basis using 2D representations, contemporary deepfake generation pipelines operate sequentially across time, inherently introducing temporal patterns that persist even under aggressive compression. Motivated by this mismatch, we formulate compression-robust deepfake detection as an end-to-end spatiotemporal learning problem.

To this end, we propose a hybrid spatiotemporal architecture that integrates the local representational capacity of a 3D convolutional neural network (3D-CNN) with the global temporal modeling capability of a factorized self-attention video vision transformer. By consistently operating in the 3D domain, from feature extraction to long-range temporal reasoning, the proposed model jointly captures localized spatiotemporal artifacts and global temporal dependencies, enabling improved robustness under severe compression and in-the-wild video conditions.

The primary contributions of this work can be summarized as follows:

- We introduce a fully video-level detection framework that departs from the dominant frame-based paradigm by employing a pretrained 3D-CNN backbone to extract joint spatial–temporal representations directly from video clips, allowing the model to exploit motion and temporal coherence that remain informative in low-quality, compressed video streams.
- We employ a factorized attention mechanism that separately models spatial and temporal dependencies. Ablation experiments demonstrate that this decoupled formulation yields more robust performance than joint spatiotemporal attention under severe compression, suggesting that preserving distinct spatial and temporal representations is beneficial for low-quality video forensic analysis.
- We conduct extensive experiments including comparisons with state-of-the-art spatial and frequency-domain methods, and detailed ablation studies which demonstrate that the proposed fully 3D spatiotemporal approach achieves competitive performance under mild compression and consistently superior robustness under aggressive compression across multiple benchmarks.

The remainder of this study is structured as follows: Section II introduce deepfake generation and reviews existing detection approaches. Section III details the proposed hybrid 3D-CNN and transformer-based architecture. Section IV describes the experimental setup and reports the results. Section V provides discussion, analysis, and insights. Section VI concludes the study.

II. Related Work

A. Deepfake Generation

Deepfake videos are commonly produced through deep generative models that learn statistical representations of facial structure, appearance, and motion. Early face-swapping systems employed autoencoder-based architectures in which a shared encoder mapped facial inputs into a latent representation, and identity-specific decoders reconstructed the manipulated output. These methods enabled basic identity replacement, but were often limited in realism.

Subsequent advancements in generative adversarial networks (GANs) [6] and variational autoencoders (VAEs) [7] marked a significant leap in synthesis quality. GAN-based models introduced adversarial training mechanisms that encouraged the generator to produce content indistinguishable from real data, resulting in improved texture fidelity, lighting realism, and temporal smoothness. More recently, diffusion models have demonstrated state-of-the-art performance in photorealistic image and video generation by iteratively denoising latent representations and capturing fine-grained visual structures.

These generative architectures underpin the two most prevalent and impactful categories of facial manipulation: Face Swap and Face Reenactment.

- Face Swap: This technique involves transferring the facial identity of a source individual onto a target person within a video. Initial approaches relied heavily on autoencoder architectures; however, recent advances predominantly utilize generative adversarial networks (GANs) to achieve higher-quality and more seamless facial replacements, surpassing the limitations of earlier methods.
- Face Reenactment: Unlike identity transfer, face reenactment manipulates the facial expressions, head poses, and movements (such as mouth and eye dynamics) of the target video to mimic those of a source video, while preserving the target's inherent identity. This is commonly realized by extracting facial landmarks or action units from the source and applying them to the target, followed by a GAN-based synthesis to produce photorealistic frames faithfully reflecting the altered expressions and movements.

The rapid evolution of these generative technologies has triggered an ongoing arms race between increasingly sophisticated forgery techniques and the development of advanced detection methodologies aimed at mitigating emerging threats posed by highly realistic digital manipulations.

B. Deepfake Detection Approaches

Despite continuous advancements, deepfake synthesis methods remain bound by inherent architectural limitations which, inevitably introduce subtle digital artifacts embedded within manipulated media. These artifacts can be visually identified within individual frames as spatial features, such as unnatural facial textures or lighting inconsistencies. Additionally, they may manifest as temporal features reflecting unnatural movements or inconsistent transitions between frames, including irregular expressions or blinking patterns. The spatial and temporal cues, while often imperceptible to the human visual system, provide critical forensic traces (i.e. fingerprints). The challenge of reliably detecting these traces has given rise to a diverse spectrum of detection methodologies and approaches, which broadly fall into three main categories: handcrafted-based feature, deep learning-based feature, and multi modal approaches.

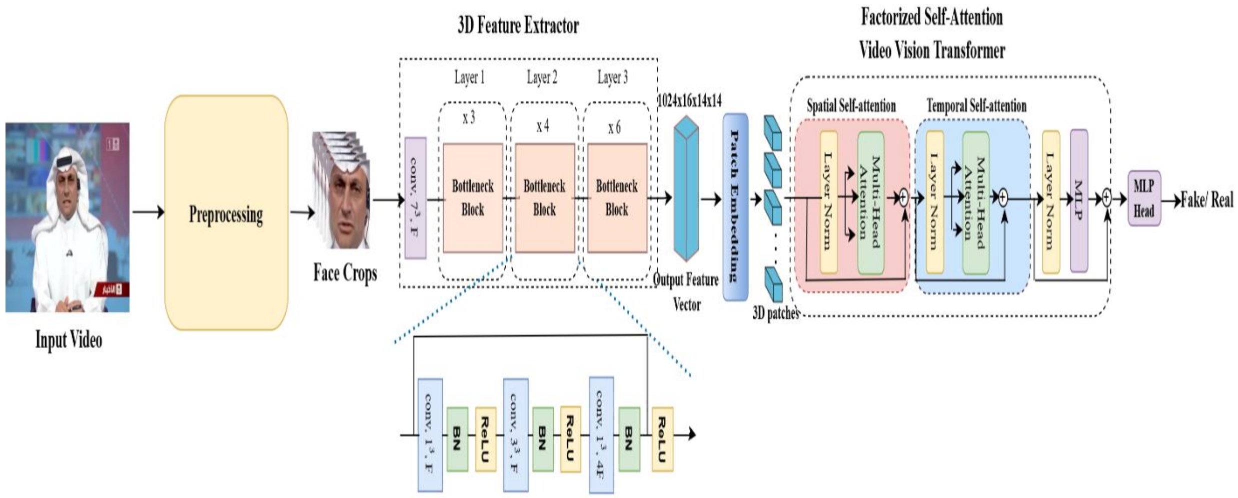


Fig. 1. Illustration of the proposed deepfake detection system.

1) Handcrafted-based features approach: This approach exploits traditional forensic cues such as inconsistencies in color blending, illumination, facial boundaries, and compression artifacts. For instance, inconsistent head poses indicate manipulation through unnatural geometric variations [8]. Face warping artifacts, introduced by affine transformations during deepfake synthesis, serve as discriminative signals [9]. Detection of blending boundary inconsistencies between the forged face and background is addressed by the face X-ray technique [10]. Another line of research leverages subtle physiological and behavioral inconsistencies inherent in deepfakes. Eye blinking patterns, typically involuntary and natural, are often absent or abnormal in synthesized videos, providing temporal cues for detection [11]. Heartbeat-induced photoplethysmographic (PPG) signals, reflecting blood flow and skin color variations, are also lacking in forgeries, effectively leveraged by Ciftci et al. [12]. Furthermore, phoneme-viseme mismatches, arising from a lack of synchronization between audio phonemes and lip movements (visemes), expose audiovisual inconsistencies unique to manipulated content [13].

While handcrafted-based methods are effective against early deepfakes, it often struggle to generalize to advanced generation techniques.

2) Deep learning-based feature approach: Deep learning-based feature approaches primarily leverage convolutional neural networks (CNNs), vision transformers (ViTs) [14], and hybrid spatiotemporal architectures to automatically learn discriminative representations. These models capture subtle spatial distortions, temporal inconsistencies, and generator-specific fingerprints that are typically imperceptible to human observers, thereby enabling robust detection of deepfake content. Afchar et al. [15] proposed MesoNet, a lightweight CNN that targets mesoscopic spatial features for detecting forged faces. Capsule-forensics [16] employed Capsule networks, which are known by their ability to catch hierarchical relationships and preserve spatial features better than convolutional neural networks. F^{3}Net is a dual-branch network that captures both local and global frequency features, employing a CNN for classification in deepfake detection [17]. SurFake model [18], which identifies manipulations by analyzing the surface geometry of faces to capture anomalies in surface caused by the forging process. FreMask [19] is a method that enhances deepfake detection by applying random masks to frequency components during training. This frequency-domain augmentation improves model robustness and generalization. The work in [20] employs a learnable uniform visibility matrix to guide the model in emphasizing subtle, imperceptible artifacts that remain in videos even after compression. Additionally, plug-and-play modules [21] with adaptive notch filters specifically remove compression noise while multi-task learning optimizes detection across domains, boosting robustness.

Deep learning methods have become central in deepfake detection due to their ability to autonomously learn intricate spatial and temporal features that reveal subtle manipulations.

3) Multimodal approach: Multimodal deepfake detection approach integrates multiple sources of information, typically combining visual and audio data, to improve detection accuracy and robustness. By fusing complementary features such as facial expressions, lip movements, voice characteristics, and audio-visual synchronization, these methods aim to capture inconsistencies that are difficult to detect using single-modality models. Salvi et al. [22] proposed an integrated system combines CNN-based visual frame analysis with audio spectrogram processing, jointly exploiting visual and auditory cues to detect deepfakes. Gandhi et al. [23] introduced a framework that jointly analyzes facial characteristics and mel-spectrogram audio features using machine learning models, achieving high detection accuracy by fusing these modalities. Alsaeedi et al. [24] proposed a multimodal fusion mechanism that integrates pre-trained audio, visual, and textual emotion features, leveraging speech sentiment and facial expressions to enhance detection robustness and accuracy.

These multimodal methods demonstrate the power of combining multiple modalities to detect subtle and sophisticated deepfake manipulations.

III. Proposed Method

To address the problem of performance degradation in videos affected by compression, this study proposes an end-to-end learning framework that transcends conventional frame-level analysis. The model comprises two primary units: first, a spatiotemporal feature extractor utilizing 3D convolutional layers to process video segments of fixed spatial and temporal dimensions; second, a transformer-based classifier that leverages factorized attention mechanisms to effectively integrate spatial and temporal features. This design enables the model to robustly distinguish between authentic and manipulated videos, maintaining high detection accuracy even under varying compression levels and quality impairments. Detailed architectural specifications are provided in the following sections:

1) Feature extraction unit: To extract robust spatiotemporal representations from the input video sequences, we employ a 3D-CNN backbone derived from the 3D-ResNet-50 architecture [25]. This framework extends the standard ResNet-50 into the temporal domain, utilizing three-dimensional convolutional kernels to inherently model volumetric video data.

The baseline architecture consists of 50 layers organized into four stages of residual bottleneck blocks. As illustrated in Fig. 1, each bottleneck unit comprises a sequence of three convolutional layers: a 1×1 convolution for dimensionality reduction, a 3×3 convolution for joint spatial and temporal features extraction, and a final 1×1 convolution for restoring the dimensionality. These layers are interleaved with Batch Normalization (BN) and Rectified Linear Unit (ReLU) activation functions. A residual skip connection bridges the input and output of each block, preserving signal integrity during backpropagation and facilitating the convergence of deep architectures. To ensure robust initialization, the model is pretrained on the Kinetics-700 dataset [26], leveraging learned weights optimized for large-scale action recognition.

To tailor this architecture specifically for the detection of fine-grained manipulation artifacts, we introduce two critical modifications: 1) Temporal Resolution Preservation: We adjusted the stride of the initial max-pooling layer from the default *(2, 2, 2,)* to *(1, 2, 2)*. By removing the temporal downsampling at this early stage, we preserve the full temporal resolution of the input sequence, allowing the network to capture subtle high-frequency frame-to-frame inconsistencies. 2) Mid-Level Feature Extraction: We truncated the network by removing the final residual block (Layer 4) and the fully connected layer. Consequently, feature maps are extracted from the output of the third residual block (Layer 3). Unlike the high-level semantic abstractions found in Layer 4, Layer 3 features retain the fine-grained spatiotemporal details essential for identifying the minute artifacts introduced by neural rendering.

The video clip consists of t consecutive frames, denoted as $\{I_1, I_2, \ldots, I_t\}$ where each frame I_i is an RGB image of fixed spatial resolution; h and w represent the height and width of the image, respectively. The corresponding clip's label is $y \in \{0, 1\}$, where $y = 0$ indicates an authentic (Real) video, while $y=1$ indicates a manipulated (Fake) video. The input tensor for the network is represented by the following shape [see Eq. (1)]:

$$X \in R^{B \times C \times T \times H \times W} \tag{1}$$

where, B denotes the batch size, $C = 3$ is the number of input channels corresponding to the RGB image, T is the temporal dimension (number of frames), and $H = W$ represent the height and width of each frame. The network processes this input as a video volume producing a feature tensor of shape:

$$F=f(X) \in R^{B \times C' \times T' \times H' \times W'} \tag{2}$$

where, $B = 1,\ C' = 1024,\ T' = 16,\ H' = W' = 14$, are the batch size, channel, temporal, height, and width dimensions of the output feature map, respectively. This tensor F encapsulates spatiotemporal feature maps, encoding both spatial patterns per frame and temporal dynamics across frames, and serves as the input to the classification module.

2) Classification unit: The Vision Transformer (ViT) is a transformer-based architecture designed for image recognition tasks, which divides an input image into small patches and processes them as a sequence of tokens. The core component of ViT is the self-attention mechanism, which enables the model to adaptively weight the importance of different patches in the image by computing pairwise interactions. This mechanism allows the transformer to capture long-range dependencies and global context from the earliest layers, unlike convolutional neural networks that rely on local receptive fields.

Our model leverages a factorized self-attention transformer inspired by the Video Vision Transformer (ViViT) [27] architecture. Unlike conventional Vision Transformers (ViT), which operate on 2D patch embeddings extracted from individual frames, the proposed transformer processes 3D patch embeddings derived from the output of a 3D-ResNet backbone, enabling the modeling of volumetric spatiotemporal representations. As illustrated in Fig. 1, the factorized self-attention module decomposes self-attention into separate spatial and temporal components: spatial self-attention is first applied independently within each frame to capture local appearance relationships, followed by temporal self-attention across frames at each spatial location to model motion and temporal dynamics.

While similar factorized attention designs have been explored in prior video transformer architectures, their adoption has largely been motivated by computational efficiency. In contrast, our use of factorized spatial–temporal attention is driven by robustness considerations in compressed video. Under aggressive compression, spatial features are disproportionately degraded, whereas temporal structure often remains more stable. By explicitly decoupling spatial and temporal attention, the model avoids entangling corrupted spatial cues with temporal reasoning. As confirmed by our ablation studies, this design consistently outperforms joint spatiotemporal attention in highly compressed settings, indicating that factorized attention is better aligned with the characteristics of low-quality video forensic analysis.

The 3D-ResNet output features tensor F from Eq. (2) is divided into nonoverlapping 3D patches across temporal and spatial dimensions. Given a patch size (p_t, p_h, p_w), where p_t, p_h, p_w are the temporal, height, and width dimensions of each patch, respectively, the number of temporal patches N_T and spatial patches N_S is given by Eq. (3) and Eq. (4):

$$N_T = \frac{T'}{p_t} \tag{3}$$

$$N_S = \frac{H'}{p_h} \times \frac{W'}{p_w} \tag{4}$$

While the total number of patches is calculated by Eq. (5):

$$\mathrm{N} = \frac{\mathrm{T}'}{\mathrm{p_t}} \times \frac{\mathrm{H}'}{\mathrm{p_h}} \times \frac{\mathrm{W}'}{\mathrm{p_w}} \tag{5}$$

Each 3D patch x_i for (i = 1, ..., N) is then flattened into a vector and linearly projected into a fixed-dimensional embedding [see Eq. (6)]:

$$z_i = E(x_i) \tag{6}$$

where, $\mathrm{E} \in \mathrm{R}^{\mathrm{D} \times \left(\mathrm{C}' \cdot \mathrm{p_t} \cdot \mathrm{p_h} \cdot \mathrm{p_w}\right)}$ is a learnable matrix, B is the batch size, and D is the embedding dimension. Thus, linearly projected N patches form a sequence of discrete tokens as follows [Eq. (7)]:

$$Z_0 = [\mathrm{z}_1, \mathrm{z}_2, \ldots, \mathrm{z}_\mathrm{N}] \in \mathrm{R}^{\mathrm{B} \times \mathrm{N} \times \mathrm{D}} \tag{7}$$

An extra class token Z_{cls} and positional embedding P are prepended and added as follows [Eq. (8)]:

$$\mathrm{Z}' = [\mathrm{z_{cls}}; \mathrm{Z_0}] + \mathrm{P} \tag{8}$$

This patch embedding step converts the volumetric feature maps into a tokenized sequence suitable for the architecture of the video transformer.

The transformer encoder processes the tokens Z' with alternating spatial and temporal self-attention blocks:

The spatial attention is applied independently on tokens from each temporal slice, modeling spatial relationships within each frame as follows [Eq. (9)]:

$$Z_{spa} = Sp - attn(Z') \in \mathbb{R}^{B \times N_T \times N_S \times D} \tag{9}$$

Following the spatial attention, the temporal attention integrates the spatial encodings across all temporal slices by applying self-attention along the temporal dimension over the class tokens output from the spatial attention. The output of the temporal attention is given by Eq. (10):

$$Z_{tmp} = tmp - attn\left(Z_{spa}\right) \in \mathbb{R}^{B \times N_S \times N_T \times D} \tag{10}$$

Finally, the output class token from the temporal attention block is passed through a Multi-layer Perceptron (MLP) head and sigmoid activation function to estimate probabilities for the classes (Real vs. Fake) [see Eq. (11)]:

$$\hat{y} = \sigma(\mathrm{MLP}(\mathrm{Z_{cls}})) \tag{11}$$

IV. Experiments

A. Datasets

To evaluate the proposed model, experiments are conducted on three widely used public deepfake detection benchmarks: FaceForensics++ (FF++) [28], Celeb-DF-v2 [29], and the DeepFake Detection Challenge (DFDC) [30] dataset. These datasets collectively cover controlled and in-the-wild scenarios, multiple manipulation techniques, and varying levels of video quality and compression.

1) FaceForensics++: FaceForensics++ (FF++) is a public dataset containing 1,000 authentic videos collected from YouTube and 4,000 manipulated videos generated using four manipulation techniques: Deepfakes (DF), Face2Face (F2F), FaceSwap (FS), and NeuralTextures (NT). The FaceForensics++ videos are compressed using the H.264 codec, with each manipulation technique provided in three compression variants: c0 represents raw or uncompressed videos, c23 represents high-quality videos with medium compression, and c40 represents low-quality videos under a high compression level. In this work, FF++ enables systematic evaluation of detection robustness across different manipulation types and compression regimes.

2) Celeb-DF-v2: This dataset is an extension of the original Celeb-DF dataset, which consists of 590 real videos for 59 celebrities collected from YouTube and 5639 fake videos generated from the real videos using an advanced face-swapping technique. Compared to earlier datasets, Celeb-DF-v2 exhibits significantly improved visual quality and reduced synthesis artifacts, making it one of the most challenging benchmarks for deepfake detection. The dataset is commonly used to assess generalization performance under high-fidelity manipulation conditions.

3) DeepFake Detection Challenge (DFDC): The DeepFake Detection Challenge (DFDC) dataset is a large-scale benchmark released as part of a competition organized by Meta (formerly Facebook). It contains over 128,000 videos generated from 3,426 paid actors spanning diverse ethnicities, genders, and age groups. Videos are captured under varying poses, lighting conditions, and recording environments to reflect real-world scenarios. Fake videos are generated using multiple manipulation techniques, including face swapping and expression manipulation.

In this work, DFDC is used exclusively to evaluate cross-dataset generalization. Models are trained on the FaceForensics++ dataset and tested on DFDC without fine-tuning, allowing assessment of robustness to unseen identities, manipulation pipelines, and real-world capture conditions.

B. Dataset Preprocessing

The auxiliary preprocessing unit takes the raw input MP4 video and transforms it into a standardized data format suitable for subsequent deepfake detection analysis. Initially, video frames are extracted sequentially using the OpenCV (cv2) library. For each extracted frame, the face detection and localization are performed via the Multi-task Cascaded Convolutional Networks (MTCNN) algorithm. The identified facial regions are then cropped and uniformly resized to 224×224 pixels to ensure network compatibility.

To enhance model robustness and generalization, a combination of both spatial augmentations (e.g., horizontal flipping, rotation, color jitter) and temporal augmentations (e.g., frame shuffling, dropouts) is applied. The face crops are then normalized using the mean and standard deviation values derived from the Kinetics dataset to standardize feature distribution. The final output of this unit is a tensor representing a face clip (a stack of T=16 cropped and processed frames) and a corresponding ground-truth label tensor (indicating Real or Fake) for supervised training.

C. Experimental Setup

All experiments were conducted on the Google Colaboratory (Colab) platform using Python 3.11.13 and PyTorch 2.6.0, running on an NVIDIA L4 GPU. The model dimension was set to 384, with 6 attention heads and an MLP dimension of 1536. Training utilized the Adam optimizer with a learning rate of 1×10−4 and a weight decay of 1×10−3 to mitigate overfitting. The Focal Loss function was employed with a focusing parameter γ=2 and balancing factor α=0.25. The model was trained for up to 30 epochs with a batch size of 32. The best checkpoint was selected based on performance. Evaluation metrics reported include:

- Accuracy (ACC): This metric measures the proportion of correct predictions (both true positives and true negatives) out of the total number of predictions made. If TP, TN, FP, FN denote true positives, true negatives, false positives, and false negatives, respectively. Accuracy is calculated as:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Accuracy reflects the overall correctness of the model's predictions but can be misleading on imbalanced datasets.

- Area Under the Curve (AUC): AUC is the area under the Receiver Operating Characteristic (ROC) curve, which plots the true positive rate (TP) against the false positive rate (FP) at varying classification thresholds. It quantifies the model's ability to distinguish between classes irrespective of any threshold. The AUC value ranges from 0 to 1, with 0.5 representing a random guess and 1.0 a perfect classifier. Higher AUC values indicate better model discrimination, especially useful for imbalanced classes.

D. Results

1) Model performance: In this experiment, we evaluated the effectiveness of the proposed model on detecting deepfake. We trained and tested the model on FF++ (raw), which composed of raw uncompressed videos, FF++ (high quality) consists of videos with medium compression (c23), FF++ (low quality), which includes videos with aggressive (c40) compression level, and the Celeb-DF-v2 datasets. The results of these experiments are presented in Table I.

2) Comparative study: To evaluate the effectiveness of the proposed model and establish a fair benchmark, a comparative study was conducted involving six state-of-the-art methods: MesoInception4 [15], F^{3}Net [17], SurFake [18], FreMask [19], Unseen Artifacts [20], and Multi-domain [21]. These methods were selected to represent a broad range of spatial-domain and frequency-domain detectors, which primarily rely on frame-level cues, enabling direct assessment of the benefit of explicit temporal modeling in comparison to non-temporal approaches. All models were trained and tested on four subsets of the FaceForensics++ (FF++) dataset, Deepfakes (DF), Face2Face (F2F), FaceSwap (FS), and NeuralTextures (NT), with consistent compression levels applied during both training and testing. Performance metrics for the baseline methods were cited from [21], ensuring consistency with previously established evaluation protocols. Table II presents the comparative performance results, while Fig. 2 further facilitates comparative analysis by visualizing the detection performance in terms of AUC across the four subsets of the FaceForensics++ dataset.

TABLE I. PERFORMANCE OF THE PROPOSED MODEL ON FACEFORENSICS++ DATASET AT VARYING COMPRESSION LEVELS AND ON THE CELEB-DF-V2 DATASET

Dataset							
FF++ (Raw)		***FF++ (High quality)***		***FF++ (Low quality)***		***CDF-v2***	
ACC	***AUC***	***ACC***	***AUC***	***ACC***	***AUC***	***ACC***	***AUC***
97.05	99.88	93.00	98.40	90.38	92.50	94.98	98.78

TABLE II. COMPARATIVE PERFORMANCE OF THE PROPOSED MODEL AND STATE-OF-THE-ART METHODS ON FACEFORENSICS++ SUBSETS UNDER CONSISTENT COMPRESSION LEVELS IN TERMS OF ACCURACY

Method	**Dataset**							
	DF		**F2F**		**FS**		**NT**	
	C23	**C40**	**C23**	**C40**	**C23**	**C40**	**C23**	**C40**
Meso4 [15]	90.09	88.06	82.19	76.64	89.63	77.89	55.33	52.86
F^3-Net [17]	97.34	92.13	97.72	84.72	98.19	88.84	89.13	58.28
SurFake [18]	98.52	88.46	97.34	75.62	98.34	84.87	90.17	52.81
FreMask [19]	98.15	92.09	98.26	85.16	99.01	87.11	88.75	58.41
Unseen Artifacts [20]	98.14	94.19	98.12	87.12	97.32	88.14	89.23	64.81
Multi-domain [21]	**99.02**	**95.17**	98.58	86.87	**99.22**	90.19	91.80	70.37
Ours	96.79	94.29	**98.93**	**91.07**	97.50	**91.43**	**92.86**	**76.79**

TABLE III. CROSS-DATASET EVALUATION IN TERMS OF (AUC) MODELS ARE TRAINED ON FF++ AND EVALUATED ON DFDC DATASET

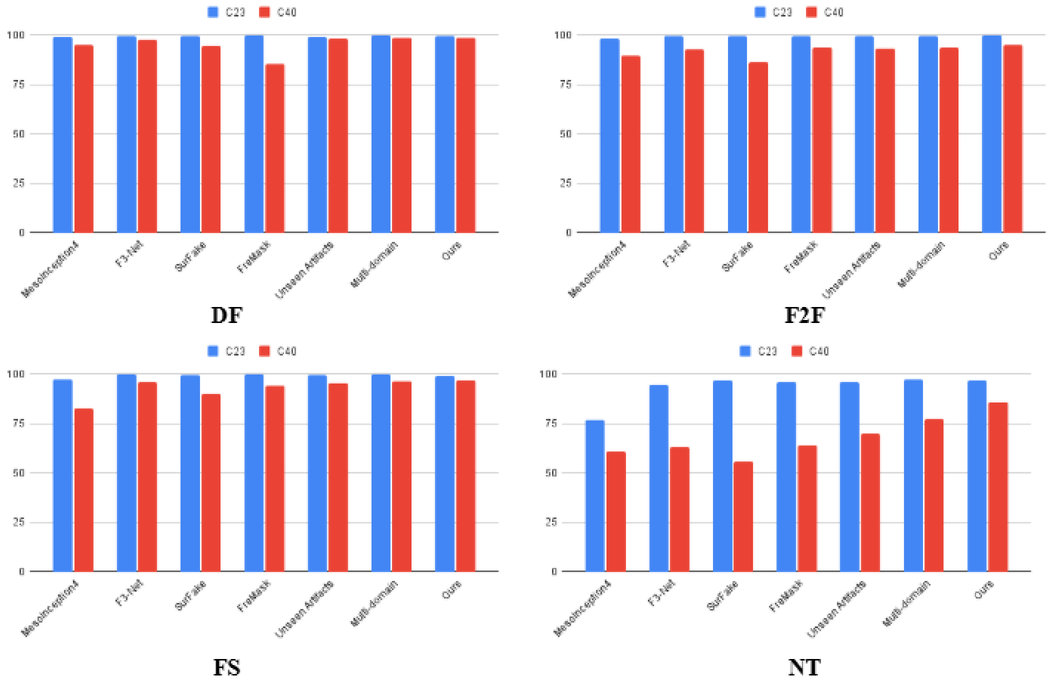
Model	Train dataset	Test dataset
		DFDC
MesoInception4	FF++	56.42
F^3-Net		63.72
SurFake		59.15
FreMask		68.23
Unseen Artifacts		64.81
Multi-domain		67.13
Ours		**71.33**

3) Cross-dataset evaluation: While our work objective is to detect deepfake in a compressed video setting, we additionally evaluated the generalization ability of our method to detect unseen deepfake. Table III summarizes the cross-dataset detection performance on the challenging DFDC test set.

4) Ablation study: To evaluate the impact of key components in the proposed model, several ablation studies were conducted. These studies included evaluating: 1) the choice between 2D and 3D convolutions for the backbone model, 2) the impact of the transformer self-attention design, and 3) the performance sensitivity to the patch size. The experiments were conducted using the Deepfakes (DF) and Face2Face (F2F) subsets under heavy compression, as they represent the two main types of Deepfakes. The results of these studies are presented in Table IV.

TABLE IV. AN ABLATION STUDY PERFORMANCE OF DIFFERENT CONFIGURATIONS ON DEEPFAKES (DF) AND FACE2FACE (F2F) DATASETS AT C40

Parameter	Setup	DF		F2F	
		ACC	AUC	ACC	AUC
Backbone	2D-ResNet50	89.64	95.17	75.36	84.57
	3D-ResNet50	93.93	97.83	82.50	91.16
Self-Attention Design	Joint Self-attention Encoder	92.50	97.80	84.29	92.81
	Factorized Self-attention Encoder	94.29	98.43	91.07	94.94
Patch size	(1, 7, 7)	92.50	96.21	86.43	93.41
	(1, 14, 14)	94.29	98.43	91.07	94.94
Hybrid Architecture	3D-Resnet50 + Factorized self-attention ViViT with patch size (1, 14, 14)	94.29	98.43	91.07	94.94



DF

F2F

FS

NT

Fig. 2. Illustration of performance degradation on the four subsets of FF++ dataset, as compression increased from medium (c23) to aggressive (c40).

V. Discussion

From the extensive experimental work, we summarize our key findings as follows:

The comparative results in Table II illustrate the performance of the proposed method against representative spatial-domain and frequency-domain detectors under both medium (c23) and high (c40) compression settings across four manipulation types. Under c23, most methods achieve strong performance across datasets, indicating that moderate compression preserves sufficient manipulation artifacts for reliable detection. In this setting, the proposed method achieves competitive performance comparable to state-of-the-art approaches.

As compression increases to c40, a clear performance gap emerges between frame-based detectors and methods that explicitly leverage temporal information. While frequency-based and multi-domain approaches maintain reasonable performance on DeepFakes (DF) and FaceSwap (FS), their detection accuracy degrades substantially on Face2Face (F2F) and NeuralTextures (NT), which involve more complex motion and expression manipulations. In contrast, the proposed method consistently achieves the highest performance on F2F and NT under c40, with AUC improvements of up to +5.72% compared to the strongest baseline on NT.

This trend highlights the advantage of fully video-level spatiotemporal modeling under aggressive compression. When spatial and frequency-domain cues are suppressed by lossy encoding, temporal inconsistencies remain more reliable indicators of manipulation. By jointly leveraging a 3D convolutional backbone and factorized temporal attention, the proposed method is better suited to capture these cues, resulting in improved robustness in low-quality video scenarios.

Table III reports the cross-dataset detection performance on the DFDC dataset, where all models are trained on FaceForensics++ and evaluated on an unseen dataset without fine-tuning. Across all baselines, detection performance decreases substantially compared to within-dataset evaluation, reflecting the well-known domain gap between curated benchmarks and in-the-wild deepfake content.

Despite this challenge, the proposed method achieves the highest performance, with an AUC of 71.33%, outperforming all compared approaches. This improvement indicates stronger generalization to unseen identities, manipulation pipelines, and real-world capture conditions. The observed gains can be attributed to the model's fully video-level spatiotemporal modeling, which reduces reliance on dataset-specific spatial or frequency artifacts that do not transfer well across domains. By emphasizing temporal consistency cues, the proposed architecture demonstrates improved robustness in cross-dataset deepfake detection scenarios.

The ablation results in Table IV provide insight into the contribution of each architectural component under heavy compression. Replacing the 2D-ResNet50 backbone with a 3D-ResNet50 yields a substantial improvement across both manipulation types, increasing AUC from 95.17% to 97.83% on DeepFakes and from 84.57% to 91.16% on Face2Face. This confirms that explicitly modeling temporal information at the feature extraction stage is critical for robust deepfake detection in compressed videos, where frame-level spatial cues are often degraded.

The impact of the self-attention design is further evidenced by comparing joint spatiotemporal attention with the proposed factorized spatial–temporal attention. While joint attention improves performance over convolution-only baselines, factorized attention consistently achieves higher accuracy and AUC, particularly on Face2Face, where AUC increases from 92.81% to 94.94%. This result suggests that decoupling spatial and temporal attention enables more stable temporal reasoning by preventing compression-corrupted spatial features from dominating temporal modeling.

The effect of patch size is also evident, with larger spatiotemporal patches (1, 14, 14) outperforming smaller ones (1, 7, 7). Larger patches capture richer contextual information across both space and time, which is beneficial for detecting subtle temporal inconsistencies under compression. Finally, the complete hybrid architecture, combining a 3D backbone with factorized self-attention and an optimized patch size, achieves the best overall performance, validating the complementary role of each design choice.

VI. Conclusion

In this work, we proposed a 3D spatiotemporal deepfake detection framework that integrates a 3D convolutional backbone with a vision transformer to enable fully video-level reasoning. Unlike prior approaches that primarily rely on frame-based spatial or frequency-domain cues, the proposed model explicitly captures long-range temporal inconsistencies, resulting in improved robustness under aggressive video compression. The use of factorized spatial–temporal self-attention further enhances detection reliability by decoupling temporal modeling from compression-degraded spatial features.

Extensive experiments demonstrate that the proposed approach achieves competitive or superior performance across multiple manipulation types, particularly under low-quality compression and in cross-dataset generalization settings. With 62.7 million parameters and a model size of 361 MB, the proposed model is best suited for server-side or offline forensic analysis, where robustness and accuracy are prioritized over strict real-time constraints. Future work will focus on improving computational efficiency and further enhancing generalization to emerging deepfake generation techniques.

Acknowledgment

This research was conducted with the support of the Organization for Women in Science for the Developing World (OWSD), Swedish International Development Cooperation Agency (Sida), and Putra Graduate Initiative Grant (Putra-IPS).

References

[1] F.-A. Croitoru et al., "Deepfake Media Generation and Detection in the Generative AI Era: A Survey and Outlook," IEEE Trans Pattern Anal Mach Intell, vol. 50, no. NO. 1, Nov. 2024, Accessed: Dec. 22, 2025. [Online]. Available: https://arxiv.org/pdf/2411.19537

[2] N. Hynek, B. Gavurova, and M. Kubak, "Risks and benefits of artificial intelligence deepfakes: Systematic review and comparison of public attitudes in seven European Countries," Journal of Innovation &

Knowledge, vol. 10, no. 5, p. 100782, Sep. 2025, doi: 10.1016/J.JIK.2025.100782.

[3] S. Singh and A. Dhumane, "Unmasking digital deceptions: An integrative review of deepfake detection, multimedia forensics, and cybersecurity challenges," MethodsX, vol. 15, p. 103632, Dec. 2025, doi: 10.1016/J.MEX.2025.103632.

[4] M. Alrashoud, "Deepfake video detection methods, approaches, and challenges," Alexandria Engineering Journal, vol. 125, pp. 265–277, Jun. 2025, doi: 10.1016/J.AEJ.2025.04.007.

[5] Y. Lu and T. Ebrahimi, "Assessment framework for deepfake detection in real-world situations," EURASIP Journal on Image and Video Processing 2024 2024:1, vol. 2024, no. 1, pp. 6-, Feb. 2024, doi: 10.1186/S13640-024-00621-8.

[6] I. J. Goodfellow et al., "Generative Adversarial Nets," Adv Neural Inf Process Syst, pp. 2672–2680, 2014, [Online]. Available: http://www.github.com/goodfeli/adversarial

[7] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings, 2014., Dec. 2013. [Online]. Available: http://arxiv.org/abs/1312.6114

[8] X. Yang, Y. Li, and S. Lyu, "Exposing Deep Fakes Using Inconsistent Head Poses," ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, vol. 2019-May, pp. 8261–8265, May 2019, doi: 10.1109/ICASSP.2019.8683164.

[9] Li Y and Lyu S, "Exposing DeepFake Videos By Detecting Face Warping Artifacts," IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 1–6, Jun. 2019.

[10] L. Li et al., "Face X-Ray for More General Face Forgery Detection," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5000–5009, Jun. 2020, doi: 10.1109/CVPR42600.2020.00505.

[11] Y. Li, M. C. Chang, and S. Lyu, "In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking," International Workshop on Information Forensics and Security, Jul. 2018, doi: 10.1109/WIFS.2018.8630787.

[12] U. A. Ciftci, I. Demir, and L. Yin, "FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals," IEEE Trans Pattern Anal Mach Intell, pp. 1–1, Jul. 2020, doi: 10.1109/TPAMI.2020.3009287.

[13] S. Agarwal, H. Farid, O. Fried, and M. Agrawala, "Detecting deep-fake videos from phoneme-viseme mismatches," IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, vol. 2020-June, pp. 2814–2822, Jun. 2020, doi: 10.1109/CVPRW50498.2020.00338.

[14] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," International Conference on Learning Representations, 2020.

[15] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," 10th IEEE International Workshop on Information Forensics and Security, WIFS 2018, Jan. 2019, doi: 10.1109/WIFS.2018.8630761.

[16] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos," ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, vol. 2019-May, pp. 2307–2311, May 2019, doi: 10.1109/ICASSP.2019.8682602.

[17] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in Frequency: Face Forgery Detection by Mining Frequency-Aware Clues," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12357 LNCS, pp. 86–103, 2020, doi: 10.1007/978-3-030-58610-2_6.

[18] A. Ciamarra, R. Caldelli, F. Becattini, L. Seidenari, and A. Del Bimbo, "Deepfake detection by exploiting surface anomalies: the SurFake approach."

[19] C. T. Doloriel and N. M. Cheung, "FREQUENCY MASKING FOR UNIVERSAL DEEPFAKE DETECTION," ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, pp. 13466–13470, 2024, doi: 10.1109/ICASSP48485.2024.10446290.

[20] S. Chhabra, K. Thakral, S. Mittal, M. Vatsa, and R. Singh, "Low-Quality Deepfake Detection via Unseen Artifacts," IEEE Transactions on Artificial Intelligence, vol. 5, no. 4, pp. 1573–1585, Apr. 2024, doi: 10.1109/TAI.2023.3299894.

[21] Y. Wang, Q. Sun, D. Rong, and R. Geng, "Multi-domain awareness for compressed deepfake videos detection over social networks guided by common mechanisms between artifacts," Computer Vision and Image Understanding, vol. 247, p. 104072, Oct. 2024, doi: 10.1016/J.CVIU.2024.104072.

[22] D. Salvi et al., "A Robust Approach to Multimodal Deepfake Detection," J Imaging, vol. 9, no. 6, p. 122, Jun. 2023, doi: 10.3390/JIMAGING9060122.

[23] K. Gandhi, P. Kulkarni, T. Shah, P. Chaudhari, M. Narvekar, and K. Ghag, "A Multimodal Framework for Deepfake Detection," Journal of Electrical Systems, Oct. 2024, doi: 10.53555/jes.v20i10s.6126.

[24] A. Alsaeedi, A. AlMansour, and A. Jamal, "Audio-Visual Multimodal Deepfake Detection Leveraging Emotional Recognition," International Journal of Advanced Computer Science and Applications, vol. 16, no. 6, pp. 213–226, Jun. 2025, doi: 10.14569/IJACSA.2025.0160622.

[25] K. Hara, H. Kataoka, and Y. Satoh, "Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 6546–6555, Nov. 2017, doi: 10.1109/CVPR.2018.00685.

[26] J. Carreira, E. Noland, C. Hillier, and A. Zisserman, "A Short Note on the Kinetics-700 Human Action Dataset," Oct. 2022, [Online]. Available: http://arxiv.org/abs/1907.06987

[27] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "ViViT: A Video Vision Transformer," Proceedings of the IEEE International Conference on Computer Vision, pp. 6816–6826, Mar. 2021, doi: 10.1109/ICCV48922.2021.00676.

[28] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Oct. 2019, pp. 1–11. doi: 10.1109/ICCV.2019.00009.

[29] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3204–3213, 2020, doi: 10.1109/CVPR42600.2020.00327.

[30] B. Dolhansky et al., "The DeepFake Detection Challenge (DFDC) Dataset.," arXiv: Computer Vision and Pattern Recognition, 2020.