International Research Journal of Multidisciplinary Scope (IRJMS), 2026;7(3):1462-1480 IRJMS

Original Article | ISSN (O): 2582-631X DOI: 10.47857/irjms.2026.v07i03.012628

A Hybrid Feature Fusion Approach for Cross-quality Deepfake Detection – F2CV

Diya Garg*, Rupali Gill

Chitkara University and Technology, Chitkara University, Punjab, India. *Corresponding Author's Email: garg.diya@chitkara.edu.in

Abstract

The rapid advancement of the deep generative models has enabled the creation of highly realistic deepfake videos capable of manipulating facial content with high visual fidelity. Detecting such manipulations becomes more challenging when videos are stored or transmitted in compressed formats, as compression alters pixel-level artifacts and may obscure forensic traces. This paper proposes a compression-aware multimodal deepfake detection framework that jointly exploits spatial, frequency and motion information from compressed videos. The framework first extracts frames and localizes facial regions using the Multitask Cascaded Convolutional Network (MTCNN) detector. Spatial features are obtained from RGB frames using ConvNeXt, frequency-domain features are extracted from Discrete Cosine Transform (DCT) representations using a Swin Transformer and temporal motion patterns are captured from motion vectors using a 3D CNN combined with a Temporal Convolutional Network (TCN). In addition, codec-level artifacts such as motion vectors and DCT coefficients are utilized to capture compression-induced inconsistencies associated with manipulated content. The extracted features are integrated using an attention-based fusion module that adaptively weights modality-specific representations before classification. By jointly modeling visual and codec-domain information, the proposed framework achieves an accuracy (ACC) level of 99.63% and Area Under Curve (AUC) of 99.83% on the FaceForensics++ (FF++) dataset, improving robustness to compression artifacts and enhances deepfake detection performance across different compression levels.

Keywords: Attention-based Fusion, Compressed Video Forensics, DCT Features, Deepfake Detection, Motion Vectors (Mvs), RGB Features.

Introduction

During the past decade, Internet traffic patterns have changed remarkably from text-based pages to multimedia-rich content. In addition, the rise of multimedia social applications like Snapchat, Instagram and TikTok is significantly influencing our lives. It not only influences the quality of an individual's life but also enables them to share their experiences in a better way. While multimedia content benefits the entertainment industry, artistic creativity and social interaction, it also poses risks to political security, individual privacy. The example of real and fake image has been shown in Figures 1A and 1B where the face of famous politician Volodymyr Zelensky (the president of Ukraine) has been altered. It is barely possible to notice anything unusual with the naked eye (1). At the same time, deepfake technology aggravates the major crisis of public trust which may further lead to the exposure of personal data, fraudulent telecommunications activities and harm to social justice. Nowadays, social media platforms rely mostly on compressed videos, as uncompressed videos consume a lot of storage whereas device memory is limited. When a user uploads a video on social media platforms like TikTok, Snapchat and Instagram, the platforms automatically compress it. If individuals send videos using platforms like WeChat, they often face file size restriction. Because of this restriction, they have to reduce the size of the video before uploading where it becomes challenging to detect such deepfake videos. While compressing, high frequency information is lost and quantization noise is added. Hence, when tested on compressed videos, high quality video detectors may suffer from significant performance degradation. The visual impairment of such deepfake video of FF++ dataset has been shown in Figures 2A-2F. The FF++ dataset consists of deepfake videos, which are represented in two different compressed formats, including C23 (less compressed) and C40 (more compressed). The existing methods tend to

(Received 29th April 2026; Accepted 15th July 2026; Published 24th July 2026)

suffer from poor performance results for deepfake detection even when working with the C40 version of FF++. Moreover, the currently available deepfake detection models are dependent on the publicly available deepfake datasets and they use convolutional neural networks-based models to detect deepfakes, but they are ineffective at identifying low-quality deepfake videos on a social network (2-5). This paper proposes a multimodal framework to identify a deepfake in a compressed video. This proposed research contributes the following:

To address this, we introduce a novel deepfake detector that simultaneously utilizes spatial texture artifacts (RGB), frequency inconsistencies (DCT) and temporal motion irregularities (MVs) in a single learning framework, which allows more thorough forensic representation than single-domain methods.

An attention-directed fusion structure is proposed to learn dynamic modality significance and cross-domain feature interactions, enabling the network to preferentially focus on the most discriminative forensic features under varying compression and manipulation levels.

This work presents one of the few deepfake detection frameworks explicitly designed for real-world compressed video scenarios by integrating DCT-domain representations with deep spatial and temporal encoders, demonstrating improved generalization and robustness across manipulation artifacts.

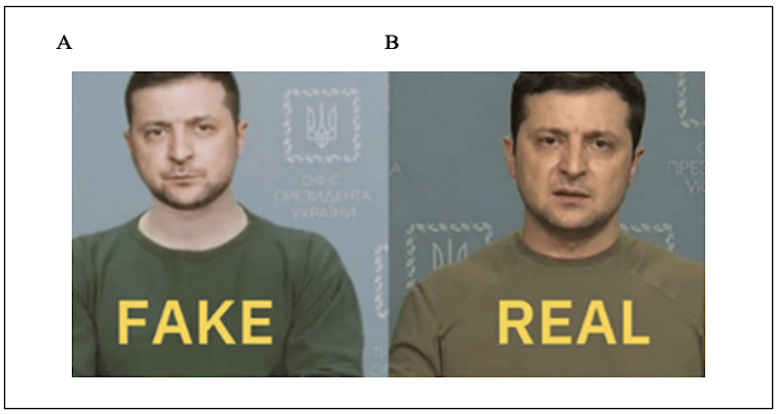

Figure 1: An Example of Executing DeepFake. (A) A Forged Picture, (B) Its Original Picture

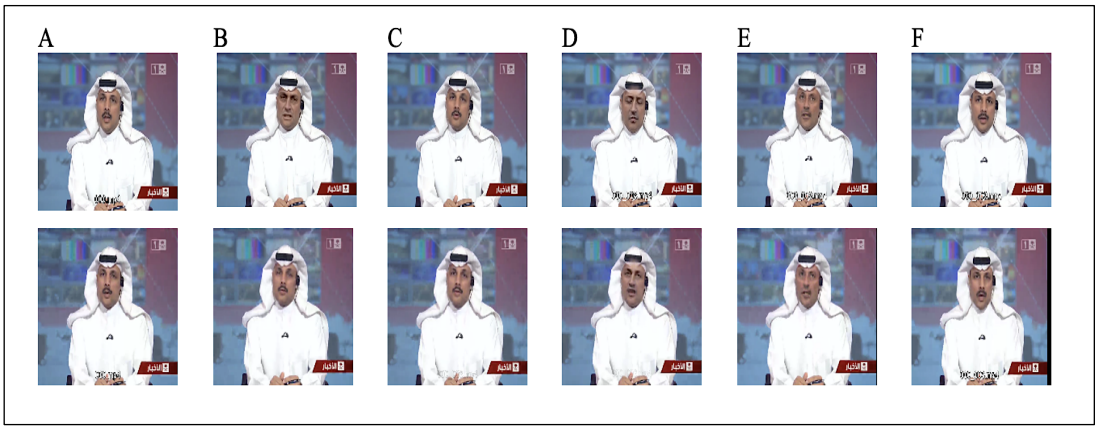

Figure 2: Real and Manipulated Frame Visual Comparison with Varying Compression Levels (C23 and C40) of FF++. (A) Real (B) Deepfake (C) Face2Face (D) FaceSwap (E) FaceShifter (F) NeuralTextures. The first row depicts C23 visuals and the second row shows C40 visuals

Related Work

In the existing research, academicians proposed various methods for detecting forgeries in videos. These methods detect object tampering (5, 6), object duplication (7), moving objects (8), frame interpolation (9), missing frames (10) and deinterlaced videos (11). As Deepfake videos create faces through artificial intelligence without leaving evidence of manipulation, traditional

methods cannot be applied to deepfake detection. To detect forged videos, several frame-based methods have been proposed (12, 13). However, these methods do not fully capture the characteristics of videos. Few studies combined both spatial and temporal information to enhance the precision of video classification (14-18).

However, some methods failed to detect forged videos where the blinking frequency was normal (14) and few approaches missed to detect forgeries where the geometric shapes and relative positions of the pupils in both eyes varied (15). Meanwhile, few methods stated in past study struggle to detect tampered faces in compressed Deepfake content (16-18). This was due to the fact that compressed Deepfake videos contain artifacts introduced during compression, which then mixed with the artifacts left by tampering. In order to fulfil this goal, the authors derived both temporal and frame level features of both the video datasets used and specifically focussed on i-frame rather than the whole frames (19). In previous study, accuracy of deepfake detection dropped significantly, focusing the need for most robust deepfake detection systems (20). To overcome the shortcomings of the given research, newer methods were developed which primarily focused on high frequency features like nose, eyes and mouth (21). Moreover, the past study emphasized the research that was basically concerned with both spatial and temporal space. CNN was employed to detect image level features in spatial domain. But for temporal domain, CNN have been used with sequential models like Long Short-Term Memory (LSTMs), Gated Recurrent Unit (GRUs) and transformers to detect temporal patterns in video frames (22, 23).

Deepfake detection of compressed videos is typically performed in either the pixel or the codec domain. As compared with pixel-domain methods that process all pixel values, codec-domain approaches use compressed video information directly, leads to lower computational complexity (24-27). However, recent advancements in deep learning architectures using spatial, temporal and facial landmarks have boosted the capabilities of detecting AI-generated images and videos (28-31).

Table 1: Comparative Analysis of Existing Deepfake Detection Method

Year	Feature	Dataset	Technique	AUC	ACC	Domain	References
2024	Spatial	FF++, Celeb-DF	CNN	-	FF++(C23) - 96%, FF++(C40) - 90%, Celeb- DF v2 - 85%	Pixel	(5)
2024	Pixel, Artifact	FF++	VGG16	-	92.26%	Pixel	(15)
2022	Temporal	FF++, Celeb-DF	CNN + Resnet	FF++ - 98%, Celeb -DF - 87%	FF++ - 94.64%, Celeb -DF – 80.74%	Pixel	(19)
2025	Spatial-Temporal	FF++, Celeb-DF v2, DFDC	CNN + LSTM	-	FF++(C23) - 94%, FF++(C40) - 90%	Pixel	(22)
2024	DCT	FF++	CNN	99.91%	98.91%	Codec	(24)
2025	Spatial-Temporal, Frequency	FF++	Conv3D + ResNet	99.8%	99%	Pixel	(25)
2018	Motion vector	CDNet2012	Statistical + Linear SVM	-	90%	Codec	(26)
2026	Spatial	DFD	ViT + TL	98.2%	98.11%	Pixel	(27)
2026	Spatial - Temporal	FF++, DFD	CNN + ViT	FF++ - 97.24%, DFD – 99.07 %	FF++ - 94.95%, DFD - 94.3%	Pixel	(28)
2025	Spatial-Temporal	DeepForensics, Celeb-DF v2	ViTs, + DenseNet	DeepForensics - 99%, CelebDF v2 - 97.4%	-	Pixel	(29)
2025	Frequency, Temporal	CelebDFv2	PhaseForensics	92.4%	-	Pixel	(30)
2025	Spatial-Temporal	FF++	CNN+LSTM	93%	96%	Pixel	(31)
2024	Spatial-Temporal	FF++, Celeb-DF v2	CNN + LSTM	FF++ - 94.12%, Celeb -DF – 86.62%	FF++ - 90.48%, Celeb -DF – 79.34%	Pixel	(32)

2023	Spatial	FF++	MesoNet, ResNet50	-	78.6%	Pixel	(33)
2023	MV, DCT, Macroblocks	KITTI, ImageNet VID	Survey	-	-	Codec	(34)
2022	Spatial	FF++, Celeb-DF v2, DFDC	ViT+Xception	-	FF++ - 98.5%, Celeb -DF - 99.2%, DFDC - 86.32%	Pixel	(35)
2023	Spatial-Temporal, MV	UCF-101, HMDB -51	Self-Attentive Octave ResNet (SOR-TC)	-	UCF-101 – 91.32%, HMDB - 51 – 60.02%	Codec	(36)
2019	Frequency	CDNet	DTCWT	-	Recall - .9287 and Precision - .9318	Pixel	(37)
2026	Spatial-Temporal, Frequency	Deep Voice + Lipreading	Cross-Attention Fusion with Adaptive Modality Gating	98.8%	96.7%	Pixel	(38)

Note: ACC - Accuracy, AUC - Area under curve, DCT – Discrete Cosine Transform, MV – Motion Vector, QP – Quantization Parameter, DFD – Deepfake Detection Dataset, FF++ – FaceForensics++, Celeb-DF – Celeb-Deepfake Dataset, DFDC – Deepfake Detection Challenge Dataset, CDNet – Change Detection Benchmark Dataset, CVLAB – Computer Vision Laboratory, ViT – Vision Transformer, TL – Transfer Learning, CNN – Convolutional Neural Network, LSTM – Long Short-Term Memory, VGG16 – Visual Geometry Group 16-layer Network, TEMSN – Temporal Enhanced Multi-Stream Network, DTCWT – Dual-Tree Complex Wavelet Transform, SVM – Support Vector Machine.

According to above literature, there is no doubt that the results that represented in Table 1, are very encouraging even under compression. however, the majority of these approaches are evaluated on selected datasets and do not have high generalization when used on unseen or cross-dataset cases (36-38). Additionally, their performance is likely to degrade, as the compression level increases and this is not reliable when used in the real world. This indicates that, regardless of accuracy, the robustness and generalization in other compressed environments is still a major challenge.

Methodology

Typical deepfake artifacts are generally present as inconsistencies in texture, frequency patterns and temporal motion. In video compression, the high-frequency components are dropped to reach lower bitrates in the process of transform coding and quantization. Consequently, the areas of the manipulated images become more difficult to distinguish from natural image content. There are three significant.

Challenges to Compression

Information Loss: Some fine-grained spatial details at facial boundaries, eyes, mouth areas and skin textures are lost. Quantization Effects: Quantization distorts DCT coefficients and diminishes discriminative frequency-domain signatures used by many deepfake detectors. Artifact Suppression: Smoothing of visual irregularities and blending artifacts that occur during face synthesis, to make forgery regions more realistic. To mitigate these effects, the suggested framework presents a new multimodal deepfake. As shown in Figure 3, the suggested framework presents a new multimodal detection approach that jointly utilizes visual, compression-domain and motion-based inconsistencies within a single architecture. Unlike current approaches that rely mostly on one modality or simple feature concatenation, our approach uses a hybrid feature extraction mechanism that retrieves complementary artifacts on the spatial, frequency and temporal domains (19, 38). Specifically, the combination of a DCT-based frequency analysis with channel attention allows effective modelling of subtle compression anomalies, whereas the motion branch uses temporal convolutional learning to model the frame-based inconsistencies that are caused by independently synthesized frames. Moreover, the attention-based fusion mechanism enables adaptive weighting of modality-specific features, which means that the model can boost the priority of the most informative cues on a per-sample input. This multi-domain feature learning and adaptive fusion combination offers a stronger and more generalized representation of deepfake detection, particularly across different compression levels and manipulation methods and thus differentiates

the proposed solution with the current state-of-the-art ones (19, 21, 25). The overall architecture of the suggested approach involves several stages, among which there are frame extraction, face detection, face extraction, multi-domain feature extraction, feature fusion and classification as shown in Figure 3. All these stages are used to detect inconsistencies added during the process of deepfake generation and compression of videos.

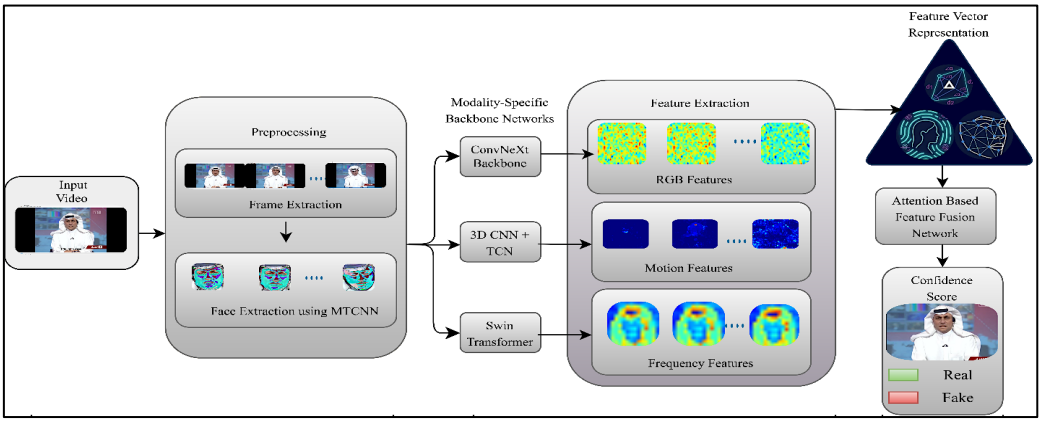

Figure 3: Hybrid Feature Fusion Framework Consisting of Preprocessing Stage (Frame Extraction, Face Detection and Face Cropping), Multi-modal Feature Extraction Using RGB, DCT and Motion-based Representations and Feature Fusion and Classification for Deepfake Detection. The video frames and feature visualizations shown in the preprocessing and feature extraction steps are taken from FF++ dataset and are displayed for illustration purposes only

Preprocessing

The preprocessing step focuses on converting unprocessed input videos into structured and standardized data formats that can be analysed using multimodal. This step is the initial predatory step to further extract spatial, frequency-domain and temporal features as it organizes the input into suitable formats for each modality. The detailed steps involved in this process are described in the following subsections that is frame extraction, face detection and face cropping which together make sure that the system focuses on the most informative parts of the video (22).

Frame Extraction

The first stage of preprocessing includes the retrieval of frames from the compressed videos. Frame extraction is used to transform the continuous video data into separate frame images that can be examined independently. It helps the system to recognize spatial appearance and temporal changes across frames, which are considered as significant indicators of manipulated content. These frames are then fed as input to the face detection module to detect facial regions in each frame (39, 40).

Face Detection

The deepfake manipulation primarily emphasizes on the facial parts; thus, the face detection is essential for effective analysis (39, 40). Once the frames are extracted, the facial areas in each frame are identified. To perform this task, a face detection algorithm is employed to each of the extracted frames to locate the bounding box proportionate to the facial region. The system only considers facial area and therefore, excludes other irrelevant background information that is not useful in detecting deepfakes.

Face Extraction

Face Extraction is the last step of preprocessing, where the identified facial regions are accurately cropped from each frame to isolate the area of interest. This step ensures that only the relevant facial information is stored while removing background noise and other irrelevant information (39, 40).

Multi-domain Feature Extraction

Most conventional deepfake detection methods rely on frequency domain representations or pixel-level visual artifacts retrieved after decoding the video. These methods are effective on high resolution videos (low compression) but their performance degrades when video is of low resolution (high compression), as in low resolution pixel level anomalies are distorted (41).

Also, as the computational cost of pixel domain is higher as compared to compression domain,

therefore the authors are considering compression domain. However, compression in videos is not a neutral process, due to introduction of some artifacts through codec i.e. motion vector, quantization and block wise transformation.

Importantly, creation of deepfake interferes with these artifacts, leaving perceptible traces at compression level. Hence for highly compressed videos, artifacts at codec level offers reliable and robust cue for deepfake detection. Thus, the feature extraction backbone of the proposed research is based on ConvNeXt, Swin Transformer and Temporal Convolutional Network (TCN) as they are complementary in capturing spatial, frequency-domain and temporal features of deepfake videos.

This research focuses on extraction of two codec domain features i.e. MVs for temporal inconsistencies and DCT features for frequency domain analysis. They are further combined with visual deep features to create a hybrid deepfake detection system.

Motion Feature Extraction

For motion feature extraction, a ResNet backbone is followed by a Temporal Convolutional Network (TCN) for modelling temporal inconsistencies. ResNet is used to extract strong frame-level motion representations and TCN is used to model temporal dependencies across video sequences via dilated causal convolutions. TCN is a recurrent architecture that allows for parallel computation, faster training and more stable gradient propagation over long temporal contexts compared to other recurrent architectures like LSTM and GRU. The features are suitable for detecting temporal artifacts caused by face manipulation methods. Optical flow is estimated between consecutive video frames to obtain motion information, instead of directly using motion vectors of compressed video bitstream, as illustrated with instances in Figures 4A- 4E. In particular, the Farneback optical flow algorithm, is used to calculate the motion of the pixels between adjacent frames.

Equation [1] shows a motion vector at spatial location L and time t1, where MV_l^{t1} represents vertical and horizontal movement of blocks.

$$MV_l^{t1} = (MVx_l^{t1}, MVy_l^{t1}) \quad [1]$$

GAN generated videos do not directly enforce smoothness in motion, resulting in unexpected changes in displacement magnitude. In deepfake videos, there is displacement in motion because of frames that are synthesized independently. To detect this, Temporal Motion Variation (TMV) measures temporary stability of motion strength across sequential frames which is represented in Equations [2, 3].

$$TMV = \frac{1}{F-1}\sum_{t1=2}^{F} |\mu(M^{t1}) - \mu(M^{t1-1}) \quad [2]$$

$$M^{t1} = \{M_l^{t1}\}_{l=1}^{N} \quad [3]$$

Where, F = Number of frames, N = Number of motion vectors in frame t1, μ = Spatial means.

The obtained motion data is further refined as shown in Equation [4], with the help of deep neural networks that produce intricate temporal patterns.

$$X^{t1} = f_{TCN}(f_{3D-CNN}(M^{t1-k}, \ldots, M^{t1})) \quad [4]$$

Where, M^{t1} = Motion vector set of frames t1, k = Temporal window size, f_{3D-CNN} = Spatial-temporal feature extractor, f_{TCN} = Temporal convolution network, X^{t1} = Final motion feature representation.

Compression/Frequency Domain Feature Extraction (DCT)

A Swin Transformer was used for frequency-domain analysis to handle the DCT-based representations. The deepfake artifacts in compressed videos can be spread out in several frequency bands and can be present in various spatial locations. The Swin Transformer uses shifted-window self-attention mechanism to model local and long-range dependencies in frequency representations, unlike the conventional convolutional network with fixed local receptive fields. This capability enables the network to detect fine discrepancies in the DCT coefficients, which can be introduced by manipulation and compression.

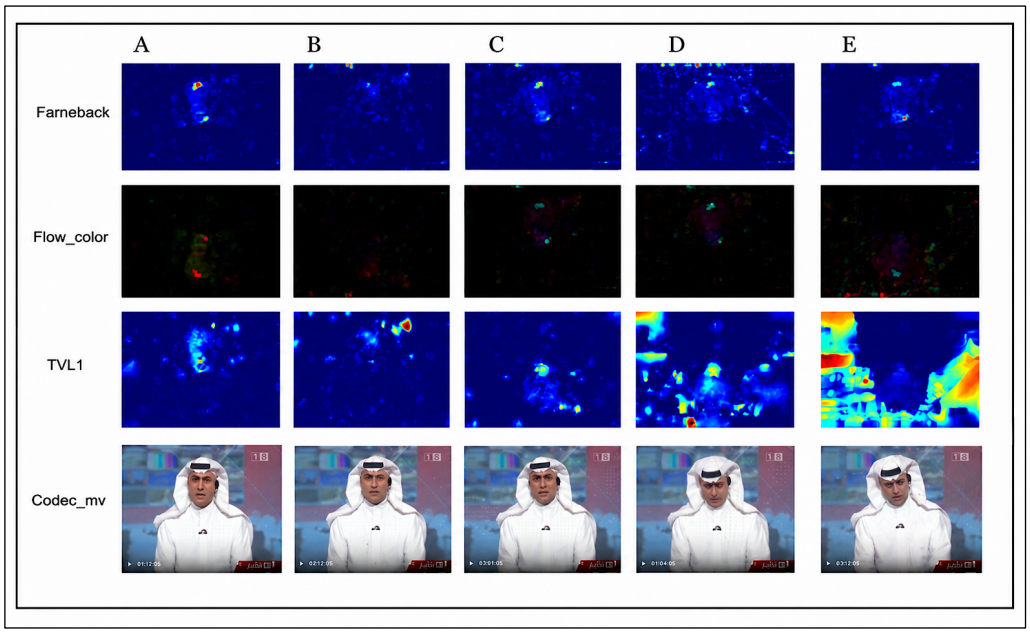

Figure 4: Visualization of Inconsistent Movements obtained from Different Stages of a Video Sequence. (A) Instance 1, (B) Instance 2, (C) Instance 3, (D) Instance 4, (E) Instance 5. The Rows Illustrate Farneback Optical Flow, Flow Color Representation, TVL1 Optical Flow and Codec-based Motion Information

In addition, the hierarchical structure of the Swin Transformer enables multi-scale feature learning, which is well suited for detecting anomalies in the frequency domain under different compression levels. The quantization process removes a part of the high-frequency content, but the deepfake generation process adds abnormal frequency distributions and inter-coefficient dependencies that cannot be completely removed. Thus, artifacts in the low- and mid-frequency DCT components are still informative for authenticating and distinguishing manipulated videos under high compression levels. As shown in Equation [5], the coefficients are extracted from compressed video frames to analyse spatial frequency patterns, particularly high-frequency components where deepfake inconsistencies are more prominent. Deepfake alterations manipulate pixel level correlations, as a result it produces abnormal frequency distributions after manipulation. For each residual block, DCT coefficients are grouped into frequency bands i.e. Low frequency components, Mid frequency components, High frequency components (34). The corresponding heatmap visualization of DCT feature extraction is shown with instances in Figures 5A-5E.

$$E_l^{t1} = \sum_{(u,v)\in\Omega} |D_l^{t1}(u,v)|^2 \qquad [5]$$

Where, $D_l^{t1}(u,v)$ = DCT coefficient where u, v presents how fast intensity changes and Ω = Selected frequency band. Moreover, to capture the temporal variation in the compression domain, the Temporal DCT Variation (TDV) is calculated as shown in Equations [6, 7].

$$TDV = \frac{1}{F-1}\sum_{t1=2}^{F} |\mu(E^{t1}) - \mu(E^{t1-1})| \qquad [6]$$

$$E^{t1} = \{E_l^{t1}\}_{l=1}^{N} \qquad [7]$$

The obtained DCT energy maps are then processed by Swin Transformer to learn hierarchical frequency representations represented in Equation [8].

$$F^{t1} = f_{Swin}(E^{t1}) \qquad [8]$$

Where, fSwin(.) denotes the Swin Transformer encoder and F^{t1} represents the extracted frequency-domain feature vector. This enables the model to learn both local and global dependencies in the frequency components, effectively modeling subtle artifacts and compression inconsistencies introduced by deepfake generation (34).

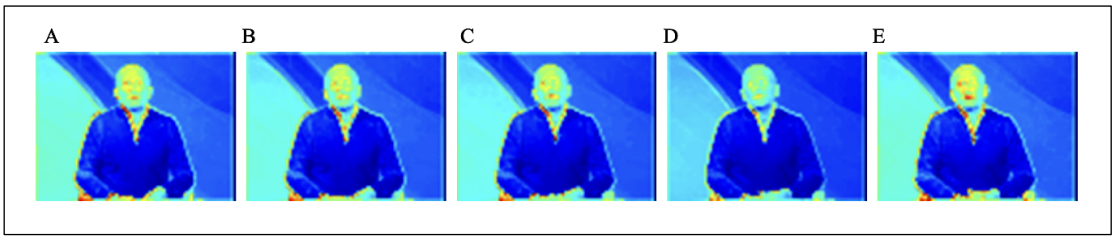

Figure 5: Heatmap Visualization of DCT Feature Extraction at Different Temporal Positions Within a Video Sequence. (A) Instance 1, (B) Instance 2, (C) Instance 3, (D) Instance 4, (E) Instance 5

Visual (RGB) Feature Extraction

In the visual domain, spatial features are learned from extracted frames to capture subtle artifacts that have been added by the deepfake generation processes. It uses a deep convolutional neural network ConvNeXt to provide high-level visual representations of RGB frames. ConvNeXt was chosen due to its fusion of the benefits of standard CNN architectures with those of recent developments. ConvNeXt has shown excellent performance in visual representation learning with low computational cost. It has a hierarchical feature extraction mechanism that is able to effectively extract both low-level texture information and high-level semantic facial features which are crucial for recognizing manipulated content. RGB features encode essential information related to facial region and texture inconsistencies which can be used to identify artifacts present in deepfake videos (42), as visualized with instances in Figures 6A-6E.

The spatial feature representation expressed in Equation [9].

$$S^{t1} = f_{ConvNext}(f_{t1}) \qquad [9]$$

Where, S^{t1} is the spatial feature vector of frame t_1 and $f_{ConvNext}(.)$ is the ConvNeXt feature extraction function.

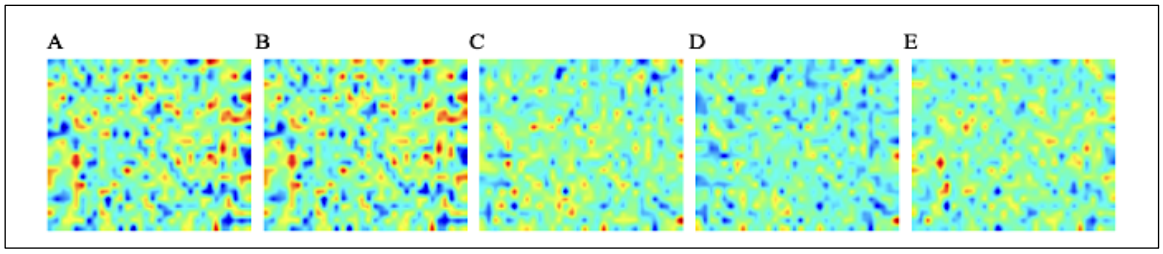

Figure 6: RGB Features Visualization Extracted from Representative Video Frames at Different Temporal Instances. (A) Instance 1, (B) Instance 2, (C) Instance 3, (D) Instance 4, (E) Instance 5

Multi-Domain Feature Fusion

Once the modality-specific representations of the RGB frames, DCT images and motion vectors are obtained the respective feature embeddings are combined with the help of an attention-based fusion mechanism. Unlike simple feature concatenation, which treats all modalities equally, the attention module learns adaptive importance weights for each modality to emphasize the most informative features for deepfake detection (25).

Let S^{t1}, F^{t1}, X^{t1} represent the feature representations of the ConvNeXt, Swin Transformer and 3D CNN - TCN networks, respectively. The attention module calculates weights i.e., ω_{rgb}, ω_{dct}, ω_{motion} on a modality basis, that are the relative contribution of each modality to the final representation. The obtained fused feature vector is then gained as shown in Equation [10].

$$F = \omega_{rgb}S^{t1} + \omega_{dct}F^{t1} + \omega_{motion}X^{t1} \qquad [10]$$

Where, F is the last multimodal feature embedding, which is applied to classification.

The attention fusion module allows the model to concentrate on the most selective cues across multiple domains by adaptively weighting spatial, frequency and temporal features. The attention weights ω_{rgb} , ω_{dct} and ω_{motion} are computed dynamically per input sample through a

lightweight gating network. Specifically, the three feature vectors S^{t1}, F^{t1}, X^{t1} are first projected into a shared embedding space of dimension d = 512 using separate fully connected layers. The projected embeddings are concatenated and passed through a two-layer Multi-layer Perceptron (MLP) with ReLU activation, producing a three-dimensional score vector.

A softmax normalization is then applied to obtain the final attention weights as given in Equation [11]:

$$[\omega_{rgb}, \omega_{dct}, \omega_{motion}] = \text{softmax}\ (\text{MLP}\ ((S^{t1};F^{t1}; X^{t1}))) \quad [11]$$

Where $(S^{t1};F^{t1}; X^{t1})$ denotes concatenation.

The softmax ensures that the weights sum to unity, allowing the model to selectively emphasize the most discriminative modality. Under heavy compression, the network tends to assign higher weight to the DCT and motion branches, as RGB-level artifacts become less reliable, whereas under lower compression the RGB branch receives comparatively higher weighting. This strategy increases resilience to compression artifacts and increases the ability to detect fine inconsistencies that were created by deepfake generation processes.

The produced fused feature vector is then fed to a classifier (Y) as expressed in Equation [12].

$$Y = \sigma(\text{Classifier}(F)) \quad [12]$$

Where, σ (.) denotes the sigmoid activation function.

Algorithm 1: Multidomain Feature Extraction and Attention Fusion on Video Deepfake Classification

Input: Compressed video V, Number of sampled frames F
Output: Final prediction y ∈ {Real, Fake}
Step 1: Frames Extraction

$$F = \{f_1, f_2, \ldots, f_F\}$$

Step 2: Visual Feature Extraction (RGB)
for $t_1 = 1 \rightarrow F$ **do**

$$S^{t_1} = f_{ConvNext}(f_{t_1})$$

end for
Step 3: Compression Domain Feature Extraction (DCT)
for $t_1 = 1 \rightarrow F$ **do**
//Compute DCT energy

$$E_l^{t_1} = \sum_{(u,v)\in\Omega} | D_l^{t_1}(u,v) |^2$$

// Construct energy set

$$E^{t_1} = \{E_l^{t_1}\}_{l=1}^{N}$$

// Extract frequency feature

$$F^{t_1} = f_{Swin}(E^{t_1})$$

end for
Step 4: Motion Feature Extraction
for $t_1 = 1 \rightarrow F$ **do**
//Extract motion vector

$$MV_l^{t_1} = (MVx_l^{t_1}, MVy_l^{t_1})$$

//Construct motion vector set

$$M^{t_1} = \{M_l^{t_1}\}_{l=1}^{N}$$

end for
//Compute temporal motion variation

$$TMV = \frac{1}{F-1}\sum_{t_1=2}^{F} | \mu(M^{t_1}) - \mu(M^{t_1-1}) |$$

//Extract motion feature

$$X^{t_1} = f_{TCN}(f_{3D-CNN}(M^{t_1-k}, \ldots, M^{t_1}))$$

Step 5: Attention-Based Multimodal Fusion
//Compute modality attention weights

$$\omega_{rgb}, \omega_{dct}, \omega_{motion}$$

//Fuse features

$$F = \omega_{rgb} S^{t1} + \omega_{dct} F^{t_1} + \omega_{motion} X^{t_1}$$

Step 6: Classification

$$y = \sigma(\text{Classifier}(F))$$

return y

Results

In this part, we assessed the effectiveness of the proposed technique that is compression aware multimodal deepfake detection on benchmark deepfake datasets. The multimodal deepfake detection framework suggested is tested with the FF++ dataset under a variety of compression (C23 and C40) to examine its robustness in the realistic conditions. Performance is evaluated based on extensive comparisons to state-of-the-art approaches based on various evaluation protocols, such as within-dataset evaluation, cross-quality evaluation across compression levels and cross-dataset analysis of generalization. These experiments prove the usefulness of the offered fusion-based attention framework to enhance the performance of deepfake detection. In the following section, we first describe the datasets used for training and evaluation, along with their characteristics.

Datasets

The study in this work is generated using multiple popular datasets that are produced by different synthesis techniques.

FF++(FaceForensics++): The FF++ dataset is one of the most widely utilized datasets in the field of deepfake detection. It was introduced to provide a standardized benchmark for evaluating deep learning models designed to identify forged facial videos.

The dataset contains both original (1000) and manipulated videos. The original videos are sourced from YouTube, while fake videos are generated using various deepfake generation techniques such as face swapping and face reenactment. Furthermore, the videos are provided in two compression levels, namely medium (c23) and high (c40), generated using the H.264 codec (43).

FSH (FaceShifter): FaceShifter is an improved deepfake generation technique and is included as part of the FF++ dataset (44).

DFD (Deepfake Detection Dataset): The Deepfake Detection dataset, developed by Google, is an open dataset aimed at facilitating advanced research in deepfake detection. It contains 360 original videos and 3068 manipulated videos (45).

DFDC (Deepfake Detection Challenge Dataset): The DFDC dataset was designed to support research in detecting deepfake content. It contains more than 1,100,000 video clips, including both real and fake samples. The dataset includes challenging scenarios such as low lighting conditions and extreme pose variations (46).

Celeb-DF: The Celeb-DF dataset consists of 590 original videos and 5659 manipulated videos. The original videos were collected from YouTube and manipulated videos were generated using an enhanced deepfake synthesis algorithm (47).

Deeper Forensics: This dataset is designed to evaluate deepfake detection methods under challenging real-world conditions. It includes real recordings from 100 actors, from which approximately 1.8 million frames are extracted, along with 11,000 manipulated videos (47).

Implementation and Hyper-parameters

The compression aware multimodal deepfake detection system is executed in the PyTorch deep learning platform and trained on a workstation with an AMD Ryzen 7000-series processor and NVIDIA GeForce RTX 4050 video card. The FaceForensics++ dataset is experimented with c23 and c40 compression setting. To conduct fair and reproducible performance measurement, the dataset is split in accordance with the standard FF++ evaluation protocol, i.e., about 70 percent was used to train the model, 15 percent to validate this model and 15 percent to test the performances. Video frames are extracted and facial regions are detected with MTCNN detector and then resized to 224×224.

Three complementary modalities are taken into consideration, namely, RGB, DCT and MV. The RGB branch is based on a ConvNeXt-Tiny backbone trained on ImageNet and the compression artifacts are approximated by a Swin Transformer applied over block-wise DCT representations. A 3D CNN

with TCN is used to capture temporal inconsistencies. The features generated by the three modalities are mapped into a shared embedding space and then the features are fused through an attention-based fusion process and a binary prediction is done with a fully connected classifier. The models are trained using AdamW optimizer and a learning rate of 1×10^{-4}. The batch size of 16 is considered in case of individual modality models and 8 in the case of the fusion network. The video level performance assessment is based on average standard measures such as accuracy, recall, F1-score, confusion matrix and AUC.

State of Art Comparison

A comparative study with the current state-of-the-art is performed in order to demonstrate the efficiency of the suggested multimodal deepfake detection framework.

Within-dataset Evaluation

In this section, a comparison between the proposed method and baselines is made on Deepfake, Face2Face, FaceSwap, NeuralTextures and FaceShifter datasets. The main aim of this study is to identify the existence of deep fake videos therefore; we test our approach on different compression qualities such as C23 and C40. As shown in Table 2, the proposed approach is better in terms of ACC and AUC scores compared to the baseline methods.

The results indicate that most of the methods perform well with high accuracy under C23 (high-quality compression) across various manipulation types, which means that deepfake detection is comparatively easier when the quality of video is preserved. Hybrid and advanced models are slightly better than traditional CNN-based approaches, yet overall differences are not very high. However, the quality of the results of almost every method is considerably lower under C40 (low-quality compression) because of loss of visual details and compression artifacts. Also, other methods struggle on manipulations such as Neural Textures, with poor generalization. In contrast, our approach demonstrates better robustness among all categories under compression and also improves generalization in low quality scenarios.

Table 2: The ACC (%) Scores with Respect to Our Method and Baseline Methods on Various C23 and C40 Deepfake Datasets

Methods	Deepfake		Face2Face		FaceSwap		Face Shifter		Neural Textures		References
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	
C23											
UA-Xcep.	94.64	-	94.64	-	95	-	-	-	-	-	(5)
UA-ResNet	93.57	-	93.57	-	94.29	-	-	-	-	-	(5)
ISD	92.3	98.0	-	-	86.7	94.0	-	-	-	-	(15)
CAHF	98.46	-	94.99	-	96.27	-	98.93	-	90.98	-	(25)
Cross ViT		86.75		86.95		82.51	-	-	-	79.62	(29)
AGD	99.02	99.95	98.58	99.64	99.2	99.90	98.8	99.86	91.8	97.13	(31)
Conv. ViT	-	84.33	-	92.36	-	82.57	-	-	-	74.05	(48)
Present study	99.33	99.96	99.17	99.98	99.63	99.89	99.5	99.96	96.13	99.92	
C40											
UA-Xcep.	95	-	73.21	-	91.79	-	-	-	-	-	(5)
UA-ResNet	93.21	-	77.5	-	92.14	-	-	-	-	-	(5)
ISD	89.6	96.0	-	-	79	88.0	-	-	-	-	(15)
FE-CNN	96.99	-	93.56	-	96.79	-	-	-	84.47	-	(22)
CAHF	97.87	-	93.40	-	95.53	-	98.85	-	85.90	-	(25)
AGD	95.17	98.73	86.87	93.84	90.19	96.47	89.89	95.77	70.37	77.36	(31)
MCW	69	77.1	65.4	73.1	64.1	70.2	-	-	59.1	62.5	(49)
Present study	97.80	99.10	89.20	94.90	96.8	98.98	93.40	96.96	89.39	92.99	

Note: ACC - Accuracy, AUC - Area under curve, UA-Xcep – Unseen Artifacts XceptionNet, UA-ResNet - Unseen Artifacts ResNet50, ISD – Identity Swap Attack Detection, CAHF - Compression-Aware Hybrid Framework, Cross ViT - Cross-Attention Multi-Scale Vision Transformer, AGD - AI-Generated Image Detection, Conv. ViT - Convolutional Vision Transformer, FE-CNN - Feature Extraction Convolutional Neural Network, CAHF - Compression-Aware Hybrid Framework, MCW - Multi-scale Central Weighted.

Cross-quality Datasets

As shown in Table 3, the cross-dataset findings (C23-C40 and C40-C23) provide that most of the existing techniques follow a significant decline in performance upon unseen compression levels, suggesting that they have restricted generalization ability. Models that have been trained on high quality data have problems with low quality data and vice versa particularly with the difficult types

of manipulation. Whereas there are some methods that perform moderately, their consistent across datasets is not reliable. In comparison, our approach has more robust and consistent results in both cross-quality conditions and it has more generalization and resistance to it has more generalization and resistance to compression variations. The ROC curves shown in Figures 7 and 8 under C23 (high-quality compression) and under C40 (low-quality compression). The proposed approach(ours) achieves strong performance with high AUC values, indicating effective deepfake detection across all manipulation types. However, under C40, the performance of several baseline methods drops significantly, especially for challenging cases like FaceSwap and NeuralTextures. In contrast, the proposed method (Ours) consistently maintains the highest AUC across all categories in both C23 and C40, demonstrating superior robustness and reliability even under heavy compression conditions.

Cross-dataset Evaluation

To evaluate the generalization capability of the proposed framework beyond the FF++ benchmark, cross-dataset experiments are conducted in which the model is trained on the FF++ dataset under C40 compression and tested on two independent datasets: Celeb-DF and the DFD. These datasets differ substantially from FF++ in terms of generation methods, visual quality and subject diversity, making them a rigorous test of the model's ability to generalize to unseen deepfake distributions. As shown in Table 4, the proposed method consistently outperforms the compared baseline approaches across both datasets in terms of ACC and AUC scores. On the Celeb-DF dataset, the proposed method achieves a notably higher AUC compared to baseline methods, demonstrating that the multimodal feature fusion strategy combining RGB, DCT and motion vector representations captures manipulation artifacts that are not specific to the FF++ manipulation pipeline. On the DFD dataset, which contains a larger proportion of manipulated videos generated using Google's synthesis pipeline, the proposed method again achieves competitive results, indicating that the codec-domain features incorporated in the framework retain discriminative power even when the generation method differs from the training data.

Table 3: The ACC (%) and AUC (%) Scores with Respect to Our Method and Baseline Methods on Cross-quality Dataset

Methods	Deepfake		Face2Face		FaceSwap		FaceShifter		NeuralTextures		References
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	
C23-C40											
UA-Xcep.	95.36	-	83.21	-	91.07	-	-	-	-	-	(5)
UA-ResNet	93.93	-	88.57	-	92.14	-	-	-	-	-	(5)
ISD	89.3	97.0	-	-	80.1	90	-	-	-	-	(15)
TSCN	88.78	-	79.95	-	77.66	-	-	-	77.78	-	(19)
CAHF	71.89	86.39	70.32	82.23	71.35	84.37	74.12	86.00	67.33	78.69	(25)
AGD	88.87	95.85	75.93	86.04	84.57	92.91	86.28	92.88	58.74	67.56	(31)
Present study	**95.60**	**99.34**	**90.99**	**88.89**	**89.74**	**94.30**	**82.45**	**88.80**	**75.76**	**90.12**	
C40-C23											
UA-Xcep.	94.28	-	91.07	-	91.07	-	-	-	-	-	(5)
UA-ResNet	95.00	-	91.78	-	93.57	-	-	-	-	-	(5)
ISD	85.6	94.0	-	-	73.7	84.0	-	-	-	-	(15)
TSCN	84.20	-	86.75	-	84.19	-	-	-	80.52	-	(19)
CAHF	89.08	95.34	77.29	85.71	82.11	89.53	80.46	88.24	69.40	77.67	(25)
AGD	95.65	99.31	90.43	95.71	93.02	98.02	89.55	97.33	66.89	69.62	(31)
Present study	**97.95**	**99.61**	**94.95**	**97.99**	**95.95**	**98.99**	**93.10**	**97.99**	**84.85**	**88.99**	

Note: ACC - Accuracy, AUC - Area under curve, UA-Xcep – Unseen Artifacts XceptionNet, UA-ResNet - Unseen Artifacts ResNet50, ISD – Identity Swap Attack Detection, TSCN - Temporality Two-stream Convolutional Network, CAHF - Compression-Aware Hybrid Framework, AGD - AI-Generated Image Detection.

Ablation Study

To validate the individual contribution of each modality within the proposed multimodal framework, an ablation study is conducted by evaluating three single-modality variants RGB only, MV only and DCT only against the full attention-based feature fusion model. Experiments are performed on the FaceForensics++ dataset under both C23 and C40 compression settings across all five manipulation types: Deepfake, Face2Face, FaceSwap, FaceShifter and NeuralTextures. As

shown in Tables 5 and 6, each individual modality contributes meaningful discriminative information, but none achieves the performance of the complete fusion model. Under C23 compression, the RGB branch alone performs competitively across most manipulation types, achieving 98.20% ACC and 98.80% AUC on Deepfake, reflecting that spatial texture artifacts remain well-preserved at lower compression levels.

Table 4: ACC (%) and AUC (%) Comparison on Cross-dataset Evaluation. The model is Trained on FF++ (C40) and Tested on Celeb-DF and DFD Datasets

Methods	Celeb-DF		DFD	
	ACC	AUC	ACC	AUC
F^3-Net	63.67	68.12	74.43	78.86
Xception	56.45	61.33	70.82	75.43
Capsule	55.81	59.48	66.29	71.12
CNN Detection	62.18	67.14	70.81	74.35
MesoInception4	53.81	62.29	63.12	69.95
RFM	61.83	67.78	67.41	72.04
Unseen Arti	63.18	70.23	72.36	77.19
FreMask	64.45	72.19	75.92	79.14
SurFake	60.11	67.41	70.93	76.52
Present study	**67.31**	**74.50**	**78.54**	**83.00**

Note: ACC - Accuracy, AUC - Area under curve, Celeb -DF – Celebrity Deepfake, DFD- Deepfake Detection.

However, the MV branch achieves notably lower accuracy on Deepfake (89.21% ACC) and NeuralTextures (86.75% ACC) under C23, indicating that motion-based cues alone are insufficient for subtle texture-level manipulations. The DCT branch shows intermediate performance, particularly benefiting FaceSwap detection (97.18% ACC under C23), where frequency-domain inconsistencies introduced during face replacement are more prominent. The benefit of multimodal fusion becomes more pronounced under C40 high-compression settings, where all single-modality variants experience a more significant performance drop. The MV branch is most affected, dropping to 76.20% ACC on Deepfake under C40, while the DCT branch degrades to 81.48% ACC on NeuralTextures. In contrast, the attention-based fusion model maintains substantially higher performance across all categories, achieving 97.80% ACC on Deepfake and 89.39% ACC on NeuralTextures under C40. This demonstrates that the attention-based fusion mechanism effectively compensates for the degradation of any single modality by dynamically re-weighting the contributions of the remaining informative branches.

Notably, FaceShifter shows an interesting pattern where the RGB branch (94.92% ACC) slightly outperforms the full fusion model (93.40% ACC) under C40, suggesting that for certain high-quality face-swapping methods, spatial features alone can be highly discriminative and the fusion of noisier compressed-domain features may introduce marginal interference. Overall, these results confirm that all three modalities are complementary and that the attention-based fusion is essential for achieving robust deepfake detection across varying compression levels and manipulation types.

Discussion

The proposed multimodal deepfake detection framework integrates RGB spatial features, DCT frequency-domain representations and motion vector-based temporal cues through an attention-based fusion mechanism. The experimental findings across within-dataset, cross-quality and cross-dataset evaluation settings collectively validate the effectiveness of this multi-domain approach. The within-dataset results on FF++ under C23 compression confirm that the proposed method achieves competitive performance across all manipulation types, including Deepfake, Face2Face, FaceSwap, FaceShifter and Neural-Textures. Under high-quality compression (C23), most baseline methods demonstrate reasonably strong performance, indicating that pixel-level artifacts remain sufficiently preserved to support detection. However, the transition to C40 reveals a stark divergence: most existing approaches experience a notable decline in both accuracy and AUC, particularly on challenging manipulation types such as NeuralTextures and FaceSwap (50, 51).

In contrast, the proposed framework maintains comparatively stable performance under C40, demonstrating the effectiveness of incorporating codec-level features specifically DCT coefficients and motion vectors that remain partially intact even under aggressive compression settings. The cross-quality evaluation (C23→C40 and C40→C23) further reinforces this observation. Methods trained on one compression quality and evaluated on another typically exhibit degraded generalization, reflecting an over-reliance on compression-specific visual artifacts. The proposed approach, by jointly modeling spatial, frequency and temporal domains, achieves greater consistency across unseen compression levels. This cross-quality robustness is particularly important for real-world deployment scenarios where video quality is highly variable depending on the platform, codec version and transmission channel.

The cross-dataset evaluation, conducted by training on FF++ (C40) and testing on Celeb-DF and DFD, further demonstrates the generalization capability of the proposed framework. While all methods experience a performance drop under the cross-dataset setting due to the distribution shift between manipulation artifacts in FF++ and those present in the target datasets, the proposed method consistently outperforms all baselines on both Celeb-DF and DFD in terms of ACC and AUC. This can be attributed to the complementary nature of the three modalities: even when spatial RGB artifacts shift across datasets, frequency-domain and motion-based cues provide additional discriminative signals that partially compensate for the distribution mismatch. Among the baseline methods, FreMask and F^3-Net achieve relatively stronger cross-dataset results owing to their frequency-aware feature extraction strategies, yet both still fall short of the detection accuracy achieved by the proposed framework.

The attention-based fusion module plays a central role across all evaluation settings. Unlike naive concatenation of features from different modalities, the adaptive weighting mechanism allows the model to dynamically prioritize the most discriminative modality depending on the input. For heavily compressed videos, the model can shift emphasis toward DCT and motion-based features when RGB-level artifacts are degraded and conversely, assign higher weight to spatial features when video quality is sufficient. The DCT branch, processed through a Swin Transformer, enables hierarchical modeling of both local and global frequency patterns. Deepfake generation methods alter pixel-level correlations in ways that manifest as abnormal frequency distributions, particularly in the high-frequency components.

By capturing these subtle spectral anomalies through block-wise DCT analysis, the framework detects traces that survive even after aggressive quantization. Similarly, the motion branch, combining a 3D CNN with a TCN, captures temporal inconsistencies arising from the independent frame synthesis inherent in GAN-based deepfake generation. The framework works on a temporal window of a fixed length (maximum 32 motion frames) instead of the entire video sequence, thus providing scalability for long duration videos.

Therefore, the computational complexity of the temporal modeling will not change significantly with respect to the length of the original video. This design is able to capture the discriminative temporal information required for deepfake detection, while simultaneously reducing the amount of computational overhead. Despite these strengths, two aspects warrant consideration for future improvement. First, the use of three parallel modality branches introduces additional computational overhead during training and inference, which presents an opportunity for future work to explore lightweight architecture designs or knowledge distillation strategies to reduce the inference cost without substantially sacrificing detection performance.

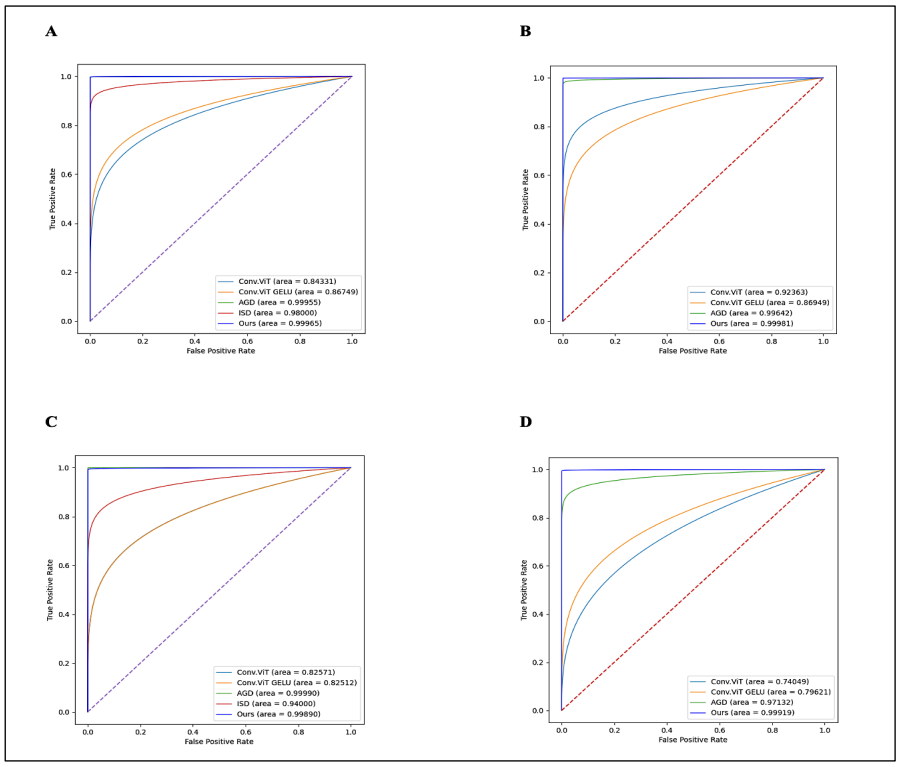

Figure 7: ROC Curves of C23 Compression Dataset. (A) DeepFake Dataset, (B) Face2Face, (C) FaceSwap, (D) Neural Texture

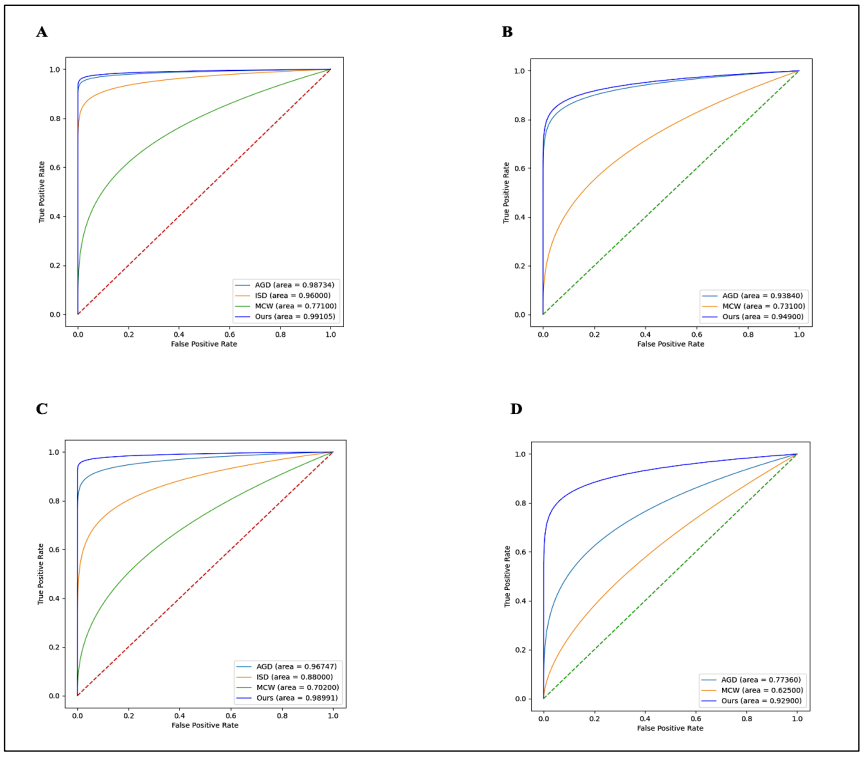

Figure 8: ROC Curves of C40 Compression Dataset. (A) DeepFake Dataset, (B) Face2Face, (C) FaceSwap, (D) Neural Texture

Table 5: Ablation Study: ACC (%) and AUC (%) of Individual Modalities vs. Attention-based Fusion on C23 Dataset

C23	RGB		MV		DCT		Attention based Feature Fusion	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
Deepfake	98.20	98.80	89.21	92.78	93.42	95.57	99.33	99.96
Face2Face	96.48	98.1	97.55	98.85	94.85	97.20	99.17	99.98
FaceSwap	96.72	98.39	95.86	97.95	97.18	98.77	99.63	99.89
FaceShifter	97.86	99.05	94.73	97.14	95.60	97.81	99.5	99.96
NeuralTextures	97.35	98.60	86.75	91.90	91.20	94.80	96.13	99.92

Note: ACC - Accuracy, AUC - Area under curve, RGB – Red, Green and blue features, MV – Motion-based Features, DCT – Discrete Cosine Transform features.

Table 6: Ablation Study: ACC (%) and AUC (%) of Individual Modalities vs. Attention-based Fusion on C40 Dataset

C40	RGB		MV		DCT		Attention based Feature Fusion	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
Deepfake	95.00	96.17	76.2	86.89	84.69	89.89	97.80	99.1
Face2Face	88.40	92.15	78.60	84.80	82.70	87.90	89.20	94.90
FaceSwap	94.45	96.44	90.62	94.50	91.15	94.64	96.8	98.80
FaceShifter	94.92	97.31	81.42	90.04	92.50	96.27	93.4	96.96
NeuralTextures	87.29	91.63	77.86	83.74	81.48	85.92	89.39	92.99

Note: ACC - Accuracy, AUC - Area under curve, RGB – Red, Green and blue features, MV – Motion-based Features, DCT – Discrete Cosine Transform features.

Second, while the proposed framework demonstrates strong cross-dataset generalization, the rapid evolution of deepfake generation methods including diffusion model-based synthesis means that extending the framework to handle completely unseen manipulation techniques remains a promising and important direction for future investigation.

Conclusion

This paper proposes a compression-aware deepfake detection method based on the multimodal fusion of spatial, frequency and temporal features extracted from compressed videos. Unlike existing single-modality methods, our method exploits the inherent artifacts caused by video compression and extracts frequency and temporal-based features from compressed videos. In particular, the method first computes the RGB visual features using ConvNeXt. Then, frequency-domain features are extracted using block-wise DCTs of the video frames and encoded using the Swin Transformer. Finally, temporal motion features are extracted by leveraging the motion vector information using the TCN, which is used to encode the temporal inconsistency in videos. To combine the modality-specific features, an adaptive fusion layer is employed that uses attention mechanisms to weigh the importance of each modality in the input samples.

The experimental evaluation involves testing the proposed approach under various settings, including within-dataset evaluation using FF++ dataset, generalization across different compression quality levels and cross-dataset generalization using the Celeb-DF and DFD datasets. In all scenarios, the proposed method outperforms the baselines with respect to accuracy and AUC. Interestingly, significant gains in performance are observed in the high-compression case and cross-quality scenarios, where most existing methods experience significant performance degradation. Moreover, the experiments show promising results in cross-dataset generalization, which indicates that the multimodal representation is robust to changes in generation pipelines and subject attributes. This suggests that the adaptive fusion layer can leverage the most important modality under diverse compression and manipulation settings.

In future this research can be extended to include designing lightweight models or adopting techniques such as knowledge distillation to enable efficient deepfake detection on mobile devices. Moreover, introducing transformers to encode long-range dependencies between video frames can help improve the detection

performance for temporal inconsistencies. In future, other fusion strategies like feature concatenation and weighted averaging will be explored and compared to further assess the effectiveness of the proposed attention-based fusion mechanism. Additionally, applying self-supervised learning to learn multimodal representations of videos can generalize the proposed framework to emerging manipulation methods.

Abbreviations

DCT: Discrete Cosine Transform, HLSR: High-Level semantic Reduction, LSTM: Long Short-Term Memory, MTCNN: Multitask Cascaded Convolutional Network, MV's: Motion vectors, TCN: Temporal Convolutional Network, TDV: Temporal DCT Variation, TMV: Temporal Motion Variation.

Acknowledgement

The authors express their sincere gratitude to Chitkara University for providing the facilities and technical support necessary for this work.

Author Contributions

Both the authors contributed equally to this manuscript. All authors have read and agreed to the published version of the manuscript.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability

The dataset used in this study is publicly available and can be accessed by researchers for academic and research purposes. The compiled data are available from the corresponding author on a reasonable request.

Declaration of Artificial Intelligence (AI) Assistance

No artificial intelligence technologies were used for this manuscript preparation.

Ethics Approval

Not applicable.

Funding

None.

References

1. Liao X, Wang Y, Wang T, Hu J, Wu X. FAMM: Facial muscle motions for detecting compressed deepfake videos over social networks. IEEE Transactions on Circuits and Systems for Video Technology. 2023;33(12):7236-51. https://doi.org/10.1109/TCSVT.2023.3278310
2. Wang L, Meng X, Li D, Zhang X, Ji S, Guo S. DEEPFAKER: A unified evaluation platform for facial deepfake and detection models. ACM Transactions on Privacy and Security. 2024;27(1):1-34. https://doi.org/10.1145/3634914
3. Wang T, Cheng H, Chow KP, Nie L. Deep convolutional pooling transformer for deepfake detection. ACM transactions on multimedia computing, communications and applications. 2023;19(6):1-20. https://doi.org/10.1145/3588574
4. Kohli A, Gupta A. Detecting deepfake, faceswap and face2face facial forgeries using frequency CNN. Multimedia Tools and Applications. 2021;80(12):18461-78. https://doi.org/10.1007/s11042-020-10420-8
5. Chhabra S, Thakral K, Mittal S, Vatsa M, Singh R. Low-quality deepfake detection via unseen artifacts. IEEE Transactions on Artificial Intelligence. 2023;5(4):1573-85. https://doi.org/10.1109/TAI.2023.3299894
6. Aloraini M, Sharifzadeh M, Schonfeld D. Sequential and patch analyses for object removal video forgery detection and localization. IEEE Transactions on Circuits and Systems for Video Technology. 2020;31(3):917-30. https://doi.org/10.1109/TCSVT.2020.2993004
7. D'Amiano L, Cozzolino D, Poggi G, Verdoliva L. A patchmatch-based dense-field algorithm for video copy–move detection and localization. IEEE Transactions on Circuits and Systems for Video Technology. 2018;29(3):669-82. https://doi.org/10.1109/TCSVT.2018.2804768
8. Chen S, Tan S, Li B, Huang J. Automatic detection of object-based forgery in advanced video. IEEE Transactions on Circuits and Systems for Video Technology. 2015;26(11):2138-51. https://doi.org/10.1109/TCSVT.2015.2473436
9. Ding X, Zhu N, Li L, Li Y, Yang G. Robust localization of interpolated frames by motion-compensated frame interpolation based on an artifact indicated map and tchebichef moments. IEEE Transactions on Circuits and Systems for Video Technology. 2018;29(7):1893-906. https://doi.org/10.1109/TCSVT.2018.2852799
10. Feng C, Xu Z, Jia S, Zhang W, Xu Y. Motion-adaptive frame deletion detection for digital video forensics. IEEE Transactions on Circuits and Systems for video Technology. 2016;27(12):2543-54. https://doi.org/10.1109/TCSVT.2016.2593612
11. Wang W, Farid H. Exposing digital forgeries in interlaced and deinterlaced video. IEEE Transactions on Information Forensics and Security. 2007;2(3):438-49. https://doi.org/10.1109/TIFS.2007.902661

12. Alnaim NM, Almutairi ZM, Alsuwat MS, Alalawi HH, Alshobaili A, Alenezi FS. DFFMD: a deepfake face mask dataset for infectious disease era with deepfake detection algorithms. IEEE Access. 2023;11:16711-16722. https://doi.org/10.1109/ACCESS.2023.3246661
13. Joshi D, Kashyap A, Arora P. Enhanced MesoNet-based deepfake detection using deep learning: A robust framework for multimedia forensics. Journal of Forensic Sciences. 2026;71(2):785-98. https://doi.org/10.1111/1556-4029.70275
14. Cheng H, Guo Y, Wang T, Li Q, Chang X, Nie L. Voice-face homogeneity tells deepfake. ACM Transactions on Multimedia Computing, Communications and Applications. 2023;20(3):1-22. https://doi.org/10.1145/3625231
15. Agarwal A, Ratha N. Detection of identity swapping attacks in low-resolution image settings. Journal of Information Security and Applications. 2025;89:103911. https://doi.org/10.1016/j.jisa.2024.103911
16. Elhassan A, Al-Fawareh M, Jafar MT, Ababneh M, Jafar ST. DFT-MF: Enhanced deepfake detection using mouth movement and transfer learning. SoftwareX. 2022;19:101115. https://doi.org/10.1016/j.softx.2022.101115
17. Raza SA, Habib U, Usman M, Cheema AA, Khan MS. MMGANGUARD: a robust approach for detecting fake images generated by GANS using multi-model techniques. IEEE Access. 2024;12:104153-104164. https://doi.org/10.1109/ACCESS.2024.3393842
18. Masud U, Sadiq M, Masood S, Ahmad M, Abd El-Latif AA. LW-DeepFakeNet: a lightweight time distributed CNN-LSTM network for real-time DeepFake video detection. Signal, Image and Video Processing. 2023;17(8):4029-37. https://doi.org/10.1007/s11760-023-02633-9
19. Hu J, Liao X, Wang W, Qin Z. Detecting compressed deepfake videos in social networks using frame-temporality two-stream convolutional network. IEEE Transactions on Circuits and Systems for Video Technology.2021;32(3):1089-102. https://doi.org/10.1109/TCSVT.2021.3074259
20. Ding X, Pang S, Guo W. Noise-aware progressive multi-scale deepfake detection. Multimedia Tools and Applications. 2024;83(36):83677-93. https://doi.org/10.1007/s11042-024-18836-2
21. Suratkar S, Kazi F. Deep fake video detection using transfer learning approach. Arabian Journal for Science and Engineering. 2023;48(8):9727-37. https://doi.org/10.1007/s13369-022-07321-3
22. Zhao C, Wang C, Song Z, Hu G, Wang L, Miao D. Multi-definition Deepfake detection via semantics reduction and cross-domain training. Pattern Recognition. 2025;163:111469. https://doi.org/10.1109/ICME52920.2022.9859736
23. Passos LA, Jodas D, Costa KA, Souza Júnior LA, Rodrigues D, Del Ser J, Camacho D, Papa JP. A review of deep learning-based approaches for deepfake content detection. Expert Systems. 2024;41(8):e13570. https://doi.org/10.1111/exsy.13570
24. Uddin M, Fu Z, Zhang X, Arnob AB. Spatial and frequency feature fusion using multi-scale cross attention for enhancing deepfake face detection. Multimedia Systems. 2025;31(4):270. https://doi.org/10.1007/s00530-025-01833-2
25. Motalib LA, Mustapha O, Mustapha H. Compression-Aware Hybrid Framework for deep fake Detection in Low-Quality Video. IEEE Access 2025. https://doi.org/10.1109/ACCESS.2025.3592358
26. Zhao L, He Z, Cao W, Zhao D. Real-time moving object segmentation and classification from HEVC compressed surveillance video. IEEE Transactions on Circuits and Systems for Video Technology. 2016;28(6):1346-57. https://doi.org/10.1109/TCSVT.2016.2645616
27. Khubaib M, Abbasi W, Saeed F, Aljohani A. Deepfake video detection using Vision Transformers: exploiting self-attention for visual manipulation detection.The Visual Computer. 2026;42(1):116. https://doi.org/10.1007/s00371-025-04319-4
28. Raghavajosyula V, Pawar D. Hybrid deep learning architecture for comprehensive deepfake video facial forgery detection and segmentation. Multimedia Tools and Applications. 2026;85(1):2. https://doi.org/10.1007/s11042-026-21320-8
29. Siddiqui F, Yang J, Xiao S, Fahad M. Enhanced deepfake detection with DenseNet and Cross-ViT. Expert Systems with Applications. 2025;267:126150. https://doi.org/10.1016/j.eswa.2024.126150
30. Prashnani E, Goebel M, Manjunath BS. Generalizable deepfake detection with phase-based motion analysis. IEEE Transactions on Image Processing. 2024;34:100-112. https://doi.org/10.1109/TIP.2024.3441821
31. Sohail S, Sajjad SM, Zafar A, Iqbal Z, Muhammad Z, Kazim M. Deepfake Detection Using Deep Learning: A Unified Forensic Approach to Detect AI-Generated Images and Videos with Fusion of Eye, Nose and Mouth Landmarks. 2025. doi: 10.20944/preprints202502. 0552.v1
32. Wang Y, Sun Q, Rong D, Geng R. Multi-domain awareness for compressed deepfake videos detection over social networks guided by common mechanisms between artifacts. Computer Vision and Image Understanding. 2024;247:104072. https://doi.org/10.1016/j.cviu.2024.104072
33. Vidya K, Ramesh P, Viknesh H, Devanand S. Compressed deepfake detection using spatio-temporal approach with model pruning. Procedia Computer Science. 2023;230:436-444. https://doi.org/10.1016/j.procs.2023.12.099
34. Zhai D, Zhang X, Li X, Xing X, Zhou Y, Ma C. Object detection methods on compressed domain videos: An overview, comparative analysis and new directions. Measurement. 2023;207:112371. https://doi.org/10.1016/j.measurement.2022.112371
35. Ganguly S, Ganguly A, Mohiuddin S, Malakar S, Sarkar R. ViXNet: Vision Transformer with Xception Network for deepfakes based video and image forgery detection. Expert Systems with Applications. 2022;210:118423. https://doi.org/10.1016/j.eswa.2022.118423
36. Zhang J, Wang X, Wan Y, Wang L, Wang J, Yu PS. SOR-TC: Self-attentive octave ResNet with temporal consistency for compressed video action recognition. Neurocomputing. 2023;533:191-205.

https://doi.org/10.1016/j.neucom.2023.02.045
37. Sengar SS, Mukhopadhyay S. Moving object detection using statistical background subtraction in wavelet compressed domain. Multimedia tools and applications. 2020;79(9):5919-40. https://doi.org/10.1007/s11042-019-08506-z
38. Waseem SS, Shabbir N, Hassan SR, Lee K. Sensors-Driven Multimodal Deepfake Detection: A Cross-Attention Fusion Approach with Adaptive Modality Gating. Sensors. 2026;26(12):3695. https://doi.org/10.3390/s26123695
39. Kingra S, Aggarwal N, Kaur N. Assessing deepfake detection methods: a comparative evaluation on novel large-scale Asian deepfake dataset. International Journal of Data Science and Analytics. 2025;20(5):4611-30. https://doi.org/10.1007/s41060-025-00741-y
40. Javed M, Zhang Z, Dahri FH, Kumar T. Enhancing multimodal deepfake detection with local–global feature integration and diffusion models. Signal, Image and Video Processing. 2025;19(5):400. https://doi.org/10.1007/s11760-025-03970-7
41. Jugade Y, Syed Z, Dcosta C, Panicker A, Vora D. Deepfake detection via spatial–temporal deep networks: leveraging CNNs and LSTMs for enhanced accuracy. Discover Applied Sciences. 2026;8(2):179. https://doi.org/10.1007/s42452-025-08014-w
42. Dong F, Zou X, Wang J, Liu X. Contrastive learning-based general Deepfake detection with multi-scale RGB frequency clues. Journal of King Saud University-Computer and Information Sciences. 2023;35(4):90-9. https://doi.org/10.1016/j.jksuci.2023.03.005
43. Wang T, Liao X, Chow KP, Lin X, Wang Y. Deepfake detection: A comprehensive survey from the reliability perspective. ACM Computing Surveys. 2024;57(3):1-35. https://doi.org/10.1145/3699710
44. Jiang J, Li B, Wei B, Li G, Liu C, Huang W, Li M, Yu M. FakeFilter: A cross-distribution Deepfake detection system with domain adaptation. Journal of Computer Security. 2021;29(4):403-21. https://doi.org/10.3233/JCS-200124
45. Kaur A, Noori Hoshyar A, Saikrishna V, Firmin S, Xia F. Deepfake video detection: challenges and opportunities. Artificial Intelligence Review. 2024;57(6):159. https://doi.org/10.1007/s10462-024-10810-6
46. Lamichhane B, Thapa K, Yang SH. Detection of image level forgery with various constraints using DFDC full and sample datasets. Sensors. 2022;22(23):9121. https://doi.org/10.3390/s22239121
47. Zhang T. Deepfake generation and detection, a survey. Multimedia Tools and Applications. 2022;81(5):6259-76. https://doi.org/10.1007/s11042-021-11733-y
48. Soudy AH, Sayed O, Tag-Elser H, Ragab R, Mohsen S, Mostafa T, Abohany AA, Slim SO. Deepfake detection using convolutional vision transformers and convolutional neural networks. Neural Computing and Applications. 2024;36(31):19759-75. https://doi.org/10.1007/s00521-024-10181-7
49. Guan L, Liu F, Zhang R, Liu J, Tang Y. Mcw: a generalizable deepfake detection method for few-shot learning. Sensors. 2023;23(21):8763. https://doi.org/10.3390/s23218763
50. Chen J, Che X, Xu K, Hu X, Li Q. Cross-modal deepfake detection: integrating textual and frequency domains. Journal of Cloud Computing. 2026;15:78. https://doi.org/10.1186/s13677-026-00898-2
51. Lai CY, Jian CY, Chuang PC, Lee CM, Hsu CC, Hsu CT, Lin CW. UMCL: Unimodal-generated Multimodal Contrastive Learning for Cross-compression-rate Deepfake Detection. International Journal of Computer Vision. 2026;134(1):40. https://doi.org/10.1007/s11263-025-02606-0

How to Cite: Garg D, Gill R. A Hybrid Feature Fusion Approach for Cross-quality Deepfake Detection – F2CV. Int Res J Multidiscip Scope. 2026;7(3):1462-1480. DOI: 10.47857/irjms.2026.v07i03.012628