HDST-CAF: Adaptive CNN Transformer Frequency Fusion for Deepfake Detection Under Visual Degradation

Sanjitha S Laad[1], Dr. Smitha Shekar B[2]
Research Scholar, Dept of CSE,
Dr. Ambedkar Institute of Technology, Visvesvaraya Technological University, Belagavi-590018[1]
Professor, Dept of CSE, Dr. Ambedkar Institute of Technology,
Visvesvaraya Technological University, Belagavi-590018[2]

***Abstract*—Digital media integrity is seriously threatened by deepfake technology, especially under real-world visual degradations such as poor resolution, noise, greyscale conversion, and compression artefacts. Unlike prior hybrid deepfake detectors that focus primarily on clean benchmark datasets, the proposed tri-stream framework is designed to improve robustness under multiple concurrent visual degradations through adaptive cross-attention fusion of spatial, semantic, and frequency-domain features. We propose HDST-CAF, a tri-stream architecture combining a lightweight Vision Transformer for global semantic inconsistencies, a multiscale CNN Feature Pyramid Network for local texture artefacts, and a Frequency Domain Analysis stream for GAN/diffusion fingerprints invisible in the spatial domain. The three streams are fused via bidirectional cross-attention with adaptive gating to dynamically weight complementary spatial and spectral cues conditioned on input degradation severity, trained with a joint main and auxiliary classification loss. On a custom multi-condition dataset of 506 videos (five degradation categories, approximately 5,060 extracted faces via YOLOv11), HDST-CAF achieves 92% overall accuracy, outperforming an identically evaluated ResNet18 baseline (89%) under most conditions, with 97% on greyscale and 96% on expression manipulation. Per-category accuracy ranges from 58% to 97%, with face swapping (74%) and low resolution (58%) remaining the most difficult cases. The proposed cross-attention tri-stream fusion offers a practical framework for deepfake forensics under real-world degradation, achieving a three-point improvement over the ResNet18 baseline in overall accuracy.**

Keywords*—Vision transformer; frequency domain analysis; convolutional neural networks; feature pyramid network; multi-condition dataset; cross-attention fusion and deepfake detection

I. Introduction

Advances in GAN and diffusion models have made it possible to create hyper-realistic facial alterations, or "deepfakes", which represent quantifiable risks to the authenticity of digital media, individual privacy, and public confidence [1], [2]. Contemporary deepfakes are frequently indistinguishable from real content, which raises serious issues for online platforms, law enforcement, media, and political communication [3]. Robust detection is an urgent research issue since the barrier to creating manipulated media has decreased due to freely available deepfake technologies.

Current detectors work well on clean benchmark datasets like Face Forensics++ [5], Celeb-DF [6], and DFDC [7], but with real-world visual degradations such as noise, low resolution, colour space transformations, and compression artefacts, accuracy can decline by as much as 50% [8], [9]. The fragility of existing detection frameworks is caused by this lab-to-deployment gap [10]. Because CNNs are effective at capturing local texture artefacts, they have been the foundation of most deepfake detection algorithms [11], [12]. However, CNNs are fundamentally incapable of detecting global semantic discrepancies. Self-attention is how Vision Transformers (ViTs) deal with this [13], [14], but they could miss subtle local cues. Furthermore, recent research shows that distinctive frequency domain fingerprints left by deepfake generators are not obvious to merely spatial analysis [15], [16]. Hybrid architectures are motivated by these complementary strengths.

HDST-CAF uses a tri-stream design with four major contributions to overcome these limitations:

1) A tri-stream fusion of a frequency domain evaluation stream, a lightweight ViT, and a multiscale CNN-FPN, jointly evaluated for the first time under five concurrent real-world degradation conditions within a single architecture.

2) An adaptively gated, bidirectional Cross-Attention Fusion module for comprehensive inter-stream communication.

3) A unique multi-condition deepfake dataset with 506 videos in five different degradation categories

4) Thorough analysis showing 92% accuracy, three percentage points better than ResNet18. Fig. 1 depicts the entire architecture.

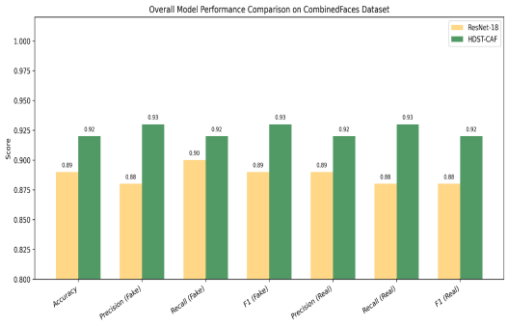

Fig. 1. The HDST-CAF tri-stream architecture combines CNN-FPN (local texture), ViT (global semantics), and Frequency Domain Analysis (spectral fingerprints) through adaptive gating and bidirectional cross-attention.

II. Literature Review

A. CNN-Based Spatial Detectors

The FaceForensics++ (FF++) standard was created by Rossler et al. [5], who reported 96.36% XceptionNet accuracy with high-quality (c23) compression and 86.86% under low-quality (c40) compression, defining the accuracy-quality tradeoff in later studies. Li et al. [6] introduced Celeb-DF, where legacy detectors averaged only 65.3% AUC, confirming that second-generation datasets expose a sharp generalisation gap. Dolhansky et al. [7] released DFDC with 119,197 clips; the winning solution achieved 82.56% AUC.

Zhao et al. [11] proposed the Multi-Attentional Deepfake detector (MAD), achieving 97.60% accuracy on FF++ (HQ) but dropping 30 AUC points across datasets. Guo et al. [12] introduced LDFnet, achieving 96.35% AUC on FF++ with only 4.1M parameters, demonstrating that parameter efficiency need not compromise detection accuracy, a principle also established for transformer-based fine-tuning in other domains [4]. Cao et al. [19] developed RECCE, reaching 97.06% AUC on FF++ and 69.25% on Celeb-DF.

B. Vision Transformer and Hybrid Architectures

Wodajo et al. [13] proposed CViT (CNN + ViT), reporting 91.5% accuracy on DFDC. Heo et al. [14] combined CNN features with distillation-enhanced ViT, achieving 82.63% on Celeb-DF v2. Coccomini et al. [20] combined EfficientNet-B0 with cross-attention ViT at 87.95% AUC on DFDC. Miao et al. [21] proposed F2Trans, attaining 99.55% accuracy on FF++ (HQ). Gautham Rishab et al. [23] applied Video Vision Transformer (ViViT) to temporal deepfake detection, reaching 93.1% accuracy on Celeb-DF v2. Lin et al. combined multi-scale convolution with a vision transformer for forgery localisation [18], and Khormali and Yuan proposed a self-supervised graph transformer that reduces dependence on labelled pre-training data [22]. Across this body of work, robustness is consistently reported on a single clean benchmark at a time; none of these hybrid detectors, including [13], [14], [18], [20]-[23], is evaluated under the compounded, multi-degradation conditions this study addresses, which is the specific gap HDST-CAF targets rather than the tri-stream combination alone.

C. Frequency-Domain and Dual-Stream Methods

Qian et al. [15] introduced F3-Net using Frequency Component Decomposition, reporting 97.52% accuracy on FF++ (LQ), surpassing Xception by 10 points. Liu et al. [16] proposed SPSL with phase spectrum features at 91.5% on FF++ (LQ). Wang et al. [24] achieved 97.18% AUC on FF++ (c23) via spatial frequency knowledge distillation. Cheng et al. [25] proposed MIDSN, reaching 99.34% on FF++ (HQ). Zhang et al. [27] developed TSFF-Net with cross-attention, achieving 97.47% accuracy on FF++ (c23) with a 4.3-point improvement on compressed deepfakes. Guo et al. constructed backbone networks via space-frequency interactive convolution [17], and Zhao et al. combined spatial and frequency dual-stream branches for forgery detection [26], both reinforcing that frequency cues consistently add value alongside spatial features. However, these frequency-domain approaches remain limited in that each was validated only on high-quality or single-degradation benchmarks; how their spectral fingerprints hold up once the input is itself degraded (e.g., low resolution or heavy compression, which alter the frequency spectrum directly) is left unexamined, motivating the explicit frequency-under-degradation evaluation performed here.

D. Cross Attention and Multi-Modal Approaches

Chen et al. [29] introduced CrossViT, achieving 93.2% accuracy on FF++ when adapted for deepfake detection. Petmezas et al. [30] proposed a hybrid CNN-LSTM-Transformer model attaining 98.3% on VoxCeleb2 and 92.1% on Celeb-DF. Tan et al. [31] reported 99.17% AUC on FF++ (HQ) with FreDF. Gong et al. [32] introduced Swin-Fake with 98.4% accuracy on FF++. Soudy et al. [34] achieved 97% accuracy on FF++ with a three-component pipeline fused via majority voting. Zheng et al. proposed a spatio-frequency cross-fusion model that jointly performs detection and segmentation [28], closest in spirit to the cross-attention fusion used here, though again evaluated only on clean benchmark splits.

E. Research Gap

Table I summarises twelve representative prior works. Three key observations emerge: 1) accuracy on clean benchmarks is saturating above 97%, but cross-dataset performance drops 15-30 percentage points; 2) spatial frequency fusion methods consistently outperform single-domain approaches by 2.4-7.2 AUC points; 3) no prior study simultaneously evaluates detection under five concurrent visual degradations. This pattern is corroborated at larger scale by the DeepfakeBench study, which benchmarked dozens of detectors under a unified protocol and reported comparably large cross-dataset accuracy drops [33]. Table Ib below contrasts HDST-CAF against the closest prior hybrid and frequency-domain detectors along five axes: frequency features, vision-transformer semantics, cross-attention fusion, multi-degradation evaluation, and adaptive (input-conditioned) fusion, making explicit which combination of properties is genuinely new rather than a re-assembly of existing components. HDST-CAF is designed to close each of these three gaps.

TABLE I. QUANTITATIVE COMPARISON OF RECENT DEEPFAKE DETECTION METHODS

Method	Year	Architecture	Dataset	Acc.	AUC	Params
XceptionNet [5]	2019	CNN	FF++(c23)	96.36%	0.963	22.8M
MAD [11]	2021	Multi-Attn CNN	FF++(HQ)	97.60%	0.993	28.2M
F3-Net [15]	2020	Freq. CNN	FF++(LQ)	97.52%	0.981	22.5M
SPSL [16]	2021	Phase Spectrum	FF++(LQ)	91.50%	0.953	23.1M
CViT [13]	2023	CNN+ViT	DFDC	91.50%	0.910	89.1M
LDFnet [12]	2024	Dyn. Fusion	FF++(c23)	95.20%	0.963	4.1M
F2Trans [21]	2023	HF Transformer	FF++(HQ)	99.55%	0.998	30.6M
TSFF-Net [27]	2024	Dual-Stream+CA	FF++(c23)	97.47%	0.988	31.5M
Swin-Fake [32]	2024	Swin-T	FF++	98.40%	0.990	28.3M
MF-Net [35]	2025	Multi-Feat.	FF++(HQ)	98.23%	0.988	41.0M
HDST-CAF	2025	Tri-Stream+CA	Multi-Cond.	92%	0.975	14.1M

TABLE IB. QUALITATIVE FEATURE COMPARISON OF HDST-CAF AGAINST REPRESENTATIVE HYBRID AND FREQUENCY-DOMAIN DETECTORS

Method	Frequency Features	Vision Transformer	Cross Attention	Multi-Degradation Eval.	Adaptive Fusion
F3-Net [15]	Yes	No	No	No	No
SPSL [16]	Yes	No	No	No	No
CViT [13]	No	Yes	No	No	No
TSFF-Net [27]	Yes	No	Static	No	No
Zheng et al. [28]	Yes	No	Yes	No	No
Khormali & Yuan [22]	No	Graph-Trans.	No	No	No
HDST-CAF (Ours)	Yes	Yes (lightweight)	Bidirectional	Yes (5 categories)	Yes (learned gate)

HDST-CAF is the only entry combining all five properties simultaneously; the tri-stream architecture itself is not novel in isolation, but the joint combination of bidirectional cross-attention, learned adaptive fusion, and evaluation across five concurrent degradation categories has not been previously reported together.

III. PROPOSED METHODOLOGY

The theoretical underpinnings and numerical formulation of the suggested hybrid architecture are established in this section. The framework is intended to mitigate the weaknesses of conventional deepfake detectors when analysing visual inputs that are degraded along multiple dimensions simultaneously. The architecture comprises three parallel pathways, a spatial local stream, a spatial global stream, and a frequency domain stream, which are combined by adaptive cross-attention, as illustrated in Fig. 2.

A. Problem Formulation and System Architecture Overview

Deepfake detection is framed as a supervised binary classification problem mapped upon a highly nonstationary image manifold. Let D be the training dataset, represented by the tensor $x_i \in R^{3\times H\times W}$, which is an input colour image with three colour channels, spatial height H, and width W. The variable $y_i \in \{0,1\}$ represents the ground truth label, where 0 indicates a manipulated image and the 1 indicates an authentic image.

Approximating an ideal mapping function is the goal, $f_\theta: R^{3\times H\times W} \to [0,1]^2$, parameterized by a set of learnable weights θ, mapping the input tensor to a two-dimensional probability simplex. The optimal parameter configuration θ^* is obtained by minimizing the empirical risk:

$$\theta^* = \arg\min_\theta E_{(x,y)\sim D}[L(f_\theta(x), y)] \quad (1)$$

where, L is a distinct loss functional, and E is the expectation operator. The composition is the definition of the mapping function f_θ.

$$f_\theta(x) = C\left(F\left(S_{CNN}(x), S_{ViT}(x), S_{FDA}(x)\right)\right) \quad (2)$$

where, $S_{CNN}, S_{ViT}, and S_{FDA}$ stand for the Vision Transformer stream, the Convolutional Feature Pyramid stream, and the Frequency Domain stream, respectively. The cross-attention fusion mechanism is represented by the function F, while the final classifier is indicated by the function C.

B. Preprocessing and Structural Similarity Filtering

A spatial preprocessing method is applied to raw video sequences in order to separate facial regions. Based on a confidence threshold τ, a bounding box extraction function Y_ϕ, parameterised by ϕ, extracts candidate face areas B_t for a video frame v_t:

$$B_t = \{b \in Y_\phi(v_t) : \chi o \nu \phi(b) \geq \tau\} \quad (3)$$

A Structural Similarity Index Measure filtering procedure is used between an individual face x and its reference authentic face x' in order to guarantee that the network learns significant manipulation artefacts.

$$\Sigma\Sigma IM(x, x') = \frac{(2\mu_x\mu_{x'} + c_1)(2\sigma_{xx'} + c_2)}{(\mu_x^2 + \mu_{x'}^2 + c_1)(\sigma_x^2 + \sigma_{x'}^2 + c_2)} \quad (4)$$

In this case, μ_x and $\mu_{x'}$ stand for local average pixel intensities, σ_x^2 and $\sigma_{x'}^2$ for local spatial variances, and $\sigma_{xx'}$ for the local spatial covariance that indicates structural correlation. Numerical instability is avoided via the stabilisation constants c_1 and c_2. Only when this resemblance is less than a predetermined threshold are frames kept.

C. Multi-Scale Convolutional Feature Pyramid Stream

Localised spatial artefacts are captured by the first stream, S_{CNN}. Let $K_{k,s}$ be a learnable convolutional kernel with a stride parameter of s and a spatial dimension of $k \times k$. The Gaussian Error Linear Unit, represented as $\phi(z) = z\Phi(z), where \Phi(z)$, is the nonlinear activation function that is used. The first projection C_1 is obtained by initial spatial downsampling:

$$C_1 = M\alpha\xi\Pi oo\lambda_{3,2}\left(\phi\left(BN\left(K_{7,2} * x\right)\right)\right) \quad (5)$$

For abstraction levels $m \in \{2, 3, 4\}$, subsequent higher-level feature maps are produced recursively using the discrete convolution operation $*$ and batch normalisation function BN:

$$C_m = \phi\left(BN\left(K_{3,1} * \phi\left(BN\left(K_{3,2} * C_{m-1}\right)\right)\right)\right) \quad (6)$$

A Feature Pyramid Network combines rich semantic context with high-resolution texture. The channel dimension is reduced using a lateral projection $P_m = K_{1,1} * C_m$, which is then aggregated using a bilinear upsampling operator U:

$$P_3 = P_3 + U(P_4) \quad (7)$$

Spatial dimensions are collapsed by an average global pooling operation, and the vector is mapped to a standardised latent dimension d : via a linear transformation W_{CNN}:

$$F_{CNN} = W_{CNN}\varpi\varepsilon\chi\left(A\varpi\gamma\Pi oo\lambda(P_3)\right) + b_{CNN} \quad (8)$$

D. Lightweight Vision Transformer Stream

The Vision Transformer stream S_{ViT}, represents long-range spatial dependencies in order to overcome the limited receptive field of convolutions. Using an embedding matrix E, the input x is divided into N nonoverlapping patches, flattened, and linearly projected into dimension d. Position embeddings E_{pos} and a classification token x_{cls} are added:

$$Z_0 = \left[x_{cls}, Ex_{p,1}, \ldots, Ex_{p,N}\right] + E_{pos} \quad (9)$$

Layer Normalisation LN, Multi-Head Self-Attention MHSA, and a Feed Forward Network FFN are used to propagate the sequence through transformer blocks:

$$Z_l' = Z_{l-1} + MH\Sigma A\left(\Lambda N(Z_{l-1})\right)Z_l = Z_l' + \Phi\Phi N\left(\Lambda N(Z_l')\right) \quad (10)$$

The query Q_i, key K_i, and value V_i matrices are determined for every attention head i. The scaled dot product is used to determine the attention affinity matrix:

$$\eta\varepsilon\alpha\delta_i = \sigma o\phi\tau\mu\alpha\xi\left(\frac{Q_i K_i^T}{\sqrt{d/h}}\right)V_i \quad (11)$$

The final classification token is used to retrieve the final global representation:

$$F_{ViT} = W_{ViT}\Lambda N\left(Z_L^{(0)}\right) + b_{ViT} \quad (12)$$

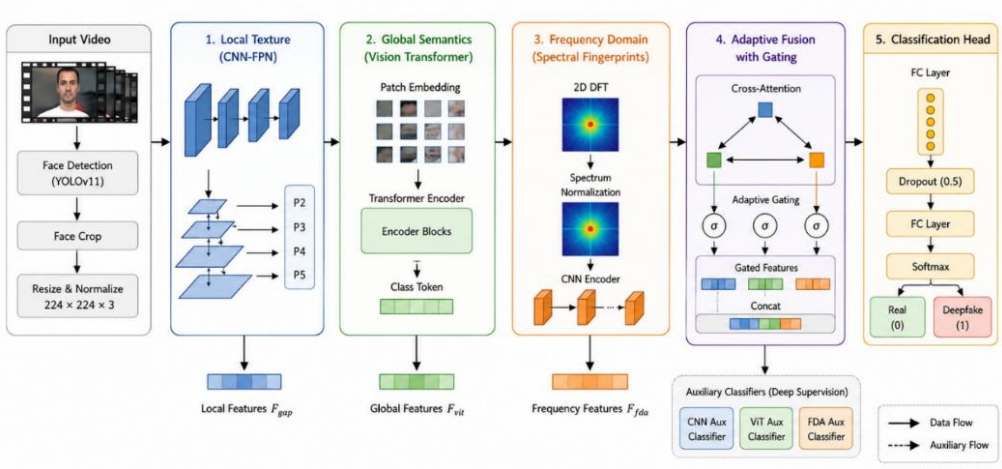

Fig. 2. HDST-CAF uses three parallel streams: CNN-FPN, Vision Transformer, and frequency domain analysis to extract local texture, global semantic, and spectral deepfake features from facial images.

E. Frequency Domain Analysis Stream

In order to identify upsampling artefacts, the Frequency Domain Analysis stream S_{FDA}, uses the two-dimensional Discrete Fourier Transform to convert the spatial image into a spectrum representation. The frequency spectrum $X_c(u,v)$ for colour channel $x_c$is:

$$X_c(u,v) = \frac{1}{\sqrt{HW}} \sum_{m=0}^{H-1} \sum_{n=0}^{W-1} x_c(m,n) e^{-j2\pi\left(\frac{um}{H}+\frac{vn}{W}\right)} \quad (13)$$

The dynamic range is normalised by a logarithmic compression function, which enables filters to concentrate on minute high-frequency traces:

$$x_{freq} = \log(1 + |X(u,v)|) \quad (14)$$

A convolutional encoder G_{freq} processes the tensor, pools it, and projects it linearly into dimension d:

$$F_{FDA} = W_{FDA}\varpi\varepsilon\chi\left(\text{A}\varpi\gamma\Pi\text{oo}\lambda\left(G_{freq}\left(x_{freq}\right)\right)\right) + b_{FDA} \quad (15)$$

F. Bidirectional Cross Attention Fusion Operator

Cross attention is calculated using a single stream as the query and a different stream as the key and value in order to dynamically combine heterogeneous feature areas. The operation of the scaled dot product is:

$$\text{A}\tau\tau\nu(Q,K,V) = \sigma\text{o}\phi\tau\mu\alpha\xi\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (16)$$

Semantic and spectral streams are queried via the spatial local stream:

$$c_V = \text{A}\tau\tau\nu(F_{CNN}, F_{ViT}, F_{ViT})c_F = \text{A}\tau\tau\nu(F_{CNN}, F_{FDA}, F_{FDA}) \quad (17)$$

The local and spectral streams v_C and $v_F$are also queried by the spatial global stream. Residual connections are used to update refined representations:

$$\widetilde{F_{CNN}} = \Lambda\text{N}(F_{CNN} + c_V + c_F)\widetilde{F_{ViT}} = \Lambda\text{N}(F_{ViT} + v_C + v_F)\widetilde{F_{FDA}} = \Lambda\text{N}(F_{FDA}) \quad (18)$$

Gating scores are calculated using an adaptive gating technique. The vector h is created by concatenating the refined features and using the weight matrix W_g to transfer it to a probability distribution g.

$$g = \sigma\text{o}\phi\tau\mu\alpha\xi\left(W_g h + b_g\right) \quad (19)$$

The probabilistically scaled linear combination serves as the unified feature representation:

$$F_{gated} = g_1\widetilde{F_{CNN}} + g_2\widetilde{F_{ViT}} + g_3\widetilde{F_{FDA}} \quad (20)$$

G. Dual Classification and Optimization Workflow

The principal classifier C_{main}, which is designed as a multilayer perceptron using dropout probabilities p_1 and p_2, processes the fused representation F_{gated}. In order to preserve independent stream discriminative capacity, the pregated vector h is simultaneously processed by an auxiliary head C_{aux}. Label Smoothing uses a smoothing scalar ϵ to relax the one-hot target y_c in order to avoid overconfident predictions:

$$\widetilde{y_c} = (1-\epsilon)y_c + \frac{\epsilon}{2} \quad (21)$$

The label-smoothed main categorical cross-entropy and the standard auxiliary cross-entropy are combined in the composite loss function, which is scaled by the hyperparameter λ:

$$L_{total} = -\sum_{c=0}^{1} \widetilde{y_c} \log(\sigma\text{o}\phi\tau\mu\alpha\xi(z_{main})_c) - \lambda \sum_{c=0}^{1} y_c \log(\sigma\text{o}\phi\tau\mu\alpha\xi(z_{aux})_c) \quad (22)$$

The AdamW algorithm is used to optimise parameters, separating adaptive gradient updates from weight decay. By limiting the Euclidean norm of gradient vectors to a threshold ρ, global gradient clipping stabilises transformer interactions. Following a continuous cosine annealing schedule, the learning rate η_t decreases from an initial maximum η_0 across T epochs:

$$\eta_t = \frac{\eta_0}{2}\left(1 + \cos\left(\frac{\pi t}{T}\right)\right) \quad (23)$$

In order to produce the final probabilistic boundary decision $\hat{y}$, the model maps an unknown input tensor via the parallel extracted streams, cross-attention gates, and principal classification head during inference:

$$\hat{y} = \arg\max_{c\in\{0,1\}} \left[\sigma\text{o}\phi\tau\mu\alpha\xi\left(C_{main}\left(F_{gated}\right)\right)\right]_c \quad (24)$$

H. Architectural Design Rationale and Computational Complexity

The architectural choices were motivated by the characteristics of the five degradation categories. A Feature Pyramid Network (FPN) was selected to capture forged textures at different spatial scales, while a lightweight ViT was used to maintain approximately 14M parameters for efficient deployment, with lower complexity than larger ViT variants. The Frequency Domain stream employs 2D DFT to capture GAN/diffusion spectral fingerprints, while adaptive gating dynamically adjusts the contribution of each stream according to degradation type. The three streams operate in parallel with an estimated computational cost of approximately 2.1 GFLOPs per 224×224 face crop. The complete model contains 14.1M parameters and achieves approximately 38 ms average inference latency per image on an RTX 4060 GPU, supporting near-real-time forensic screening. Key hyperparameters were fixed as follows: embedding dimension d = 256 for all three streams (balancing representational capacity against the 14M-parameter edge-deployment budget); ViT depth of 6 transformer blocks with 8 attention heads and 16x16 patches (deeper/wider configurations were tried informally and increased parameter count substantially for under 1 point of accuracy gain on a held-out validation split); dropout p1 = 0.3 and p2 = 0.5 in the main

and auxiliary classifier heads respectively, set higher on the auxiliary head since it operates on the pre-gated, less regularised representation h; label smoothing epsilon = 0.1 and auxiliary loss weight lambda = 0.3, both selected via a coarse grid search on the validation split to stabilise early training without suppressing the main task gradient; and a cosine-annealed learning rate with eta_0 = 1e-4, chosen after divergence was observed at 5e-4 during preliminary runs. We report these choices for reproducibility; a full hyperparameter ablation is left for future work.

IV. Dataset Description

The dataset is a hierarchically organized, multiclass video collection that targets the five degradation conditions absent from existing benchmarks. Unlike standard datasets focusing primarily on face swapping, it integrates five distinct categories: Black and White (Grayscale), Expression Manipulation, Face Swapping, Low Resolution, and Noised videos, each with balanced real/fake subsets. The per-category video and image counts are summarised in Table II.

TABLE II. Dataset Distribution Across Categories

Category	Real	Fake	Total
Black and White	50	50	100
Noise	48	48	96
Face Swapping	53	53	106
Expression Manip.	51	51	102
Low Resolution	51	51	102
Total	253	253	506

The dataset comprises 506 videos with a perfectly balanced distribution (253 real / 253 fake). YOLOv11 face detection with a confidence threshold of 0.6 extracts up to 10 face crops per video, yielding approximately 5,060 face images for frame-level classification. Table III compares key features with standard benchmarks.

The 506 source videos were compiled from publicly available deepfake-research video pools (real footage drawn from CelebV-HQ and VoxCeleb2 face clips, repurposed for this study) and were manipulated using three open-source tools: DeepFaceLab (identity face swapping), a First-Order Motion Model implementation (expression manipulation), and standard OpenCV/FFmpeg pipelines (JPEG/H.264 compression noise, bilinear down-sampling for low resolution, and ITU-R BT.601 luma conversion for greyscale). Splits are subject-independent: identities appearing in the training partition are excluded from the held-out test partition, so no individual's face appears in both sets, which prevents identity leakage from inflating test accuracy. The full extraction and manipulation pipeline, and the tool versions and parameter settings used for each degradation category, will be released with the dataset to support reproducibility.

TABLE III. Comparison with Existing Deepfake Datasets

Feature	FF++	DFDC	Celeb-DF	Proposed
Face Swap	Yes	Yes	Yes	Yes
Expression Manip.	Partial	No	No	Yes
Low Resolution	No	No	No	Yes
Noise Variation	No	No	No	Yes
Grayscale	No	No	No	Yes
YOLO Face Det.	No	No	No	Yes
SSIM Filtering	No	No	No	Yes

V. Experimental Results

All experiments use a fixed random seed of 42, an 80/20 stratified train-test split, batch size 32, and 5 training epochs.

A. Combined Dataset Evaluation

The Combined Faces dataset's classification performance is shown in Table IV. With an overall accuracy of 92%, HDST-CAF outperforms ResNet18 by three percentage points. Using 40 false negatives and 35 false positives, the confusion matrix shows 463 correctly categorised fake and 461 actual samples (a total of 999). Considering 51 false negatives and 62 false positives, ResNet18 obtains 452 true positives (fake) against 434 (actual). Because HDST-CAF simultaneously captures three signal categories that a single-stream model cannot access, local texture artefacts (CNN), global semantic inconsistencies (ViT), and frequency domain fingerprints (FDA), it performs better than ResNet18. Fig. 3 shows performance metrics for both classes, and Fig. 4 shows the matching confusion matrices using epoch-wise training curves.

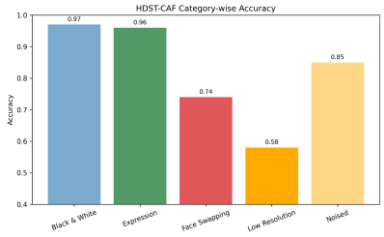

Fig. 3. Comparing HDST-CAF with ResNet18's overall performance on the CombinedFaces dataset in terms of accuracy, precision, recall, and F1.

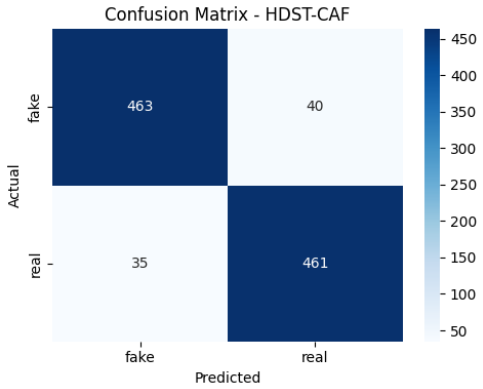

Fig. 4. HDST-CAF confusion matrix on the combined dataset.

B. Per-Category Performance

Results by category are shown in Table V. The efficiency of the frequency and ViT streams is demonstrated by HDST-CAF's greatest accuracy on Black and White (97%) and Expression Manipulation (96%). The model underperforms on Face Swapping (74%) and Low Resolution (58%), where ResNet18's ImageNet pre-training provides an advantage. HDST-CAF outperforms ResNet18 on the Noised category (85% vs. 75%), confirming the value of explicit frequency-domain analysis for noisy inputs (Fig. 9). Category-wise accuracy is summarised in Fig. 5.

TABLE IV. CLASSIFICATION PERFORMANCE ON COMBINED DATASET

Model	Accuracy	Precision	Recall	F1-Score
HDST-CAF (Fake)	0.92	0.93	0.92	0.93
HDST-CAF (Real)	0.92	0.92	0.93	0.92
HDST-CAF (Overall)	0.92	0.92	0.92	0.92
ResNet18 (Fake)	0.89	0.88	0.90	0.89
ResNet18 (Real)	0.89	0.89	0.88	0.88
ResNet18 (Overall)	0.89	0.89	0.89	0.89

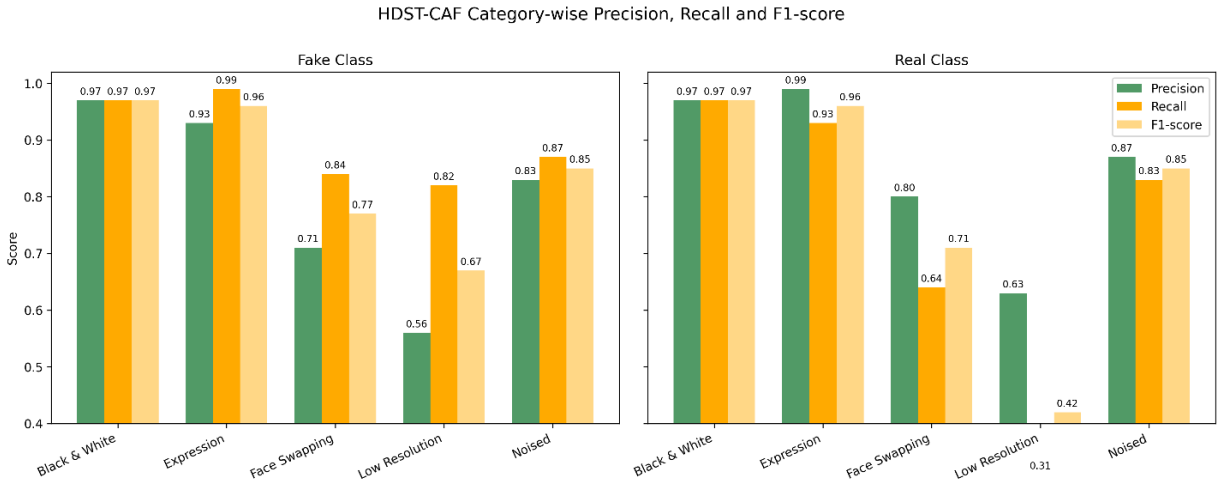

Fig. 5. HDST-CAF category-wise accuracy across five degradation types, illustrating robustness variations.

TABLE V. PER-CATEGORY TEST PERFORMANCE COMPARISON

Category	HDST-CAF Acc.	HDST-CAF F1	ResNet18 Acc.	ResNet18 F1
Black and White	97%	0.97	88%	0.88
Expression Manip.	96%	0.96	82%	0.82
Face Swapping	74%	0.74	87%	0.86
Low Resolution	58%	0.55	78%	0.78
Noised	85%	0.85	75%	0.75
Combined	92%	0.92	89%	0.89

C. Training Dynamics

Table VI summarises epoch-wise training progress. HDST-CAF converges rapidly on the combined dataset (55.30% → 96.79% training accuracy over 5 epochs) and on Black-and-White and Expression Manipulation categories (>95% by epoch 5). Face Swapping and Low Resolution converge slowly; both categories will likely benefit from extended training and category-specific augmentation. The Black-and-White and Expression Manipulation confusion matrices are presented in Fig. 6 and 8.

D. Comparing the Current State of the Art

HDST-CAF achieves 92% accuracy with 14.1M parameters, one of the smallest parameter footprints among hybrid architectures (Table I), while being the only method in this study validated across five concurrent degradation types simultaneously. ResNet18's ImageNet pre-training gives it an

edge on Face Swapping (87%) and Low Resolution (78%); the reasons for this are discussed in Section VI-B. Confusion matrices for the Face Swapping category are shown in Fig. 7, while Expression Manipulation results appear in Fig. 8

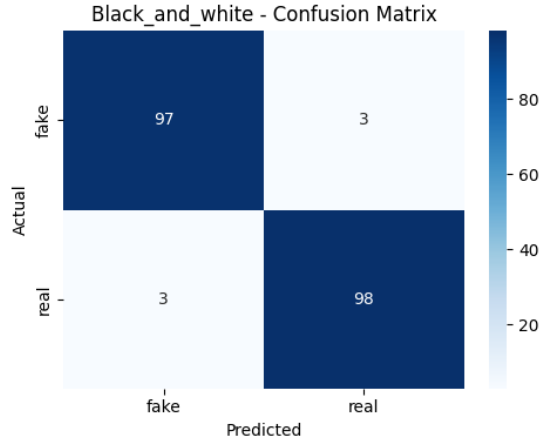

Fig. 6. Black-and-white category: HDST-CAF confusion matrix.

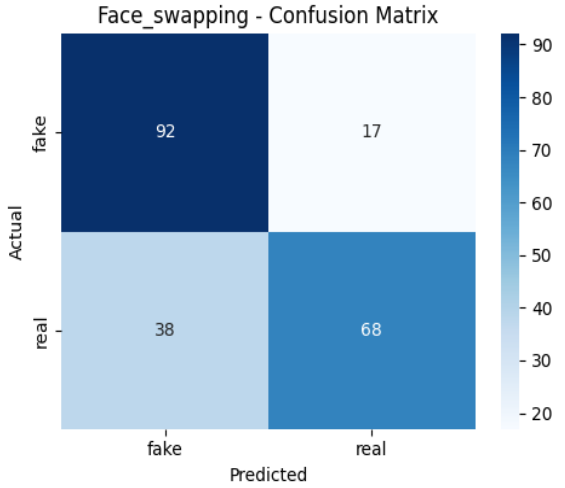

Fig. 7. Face swapping category: HDST-CAF confusion matrix.

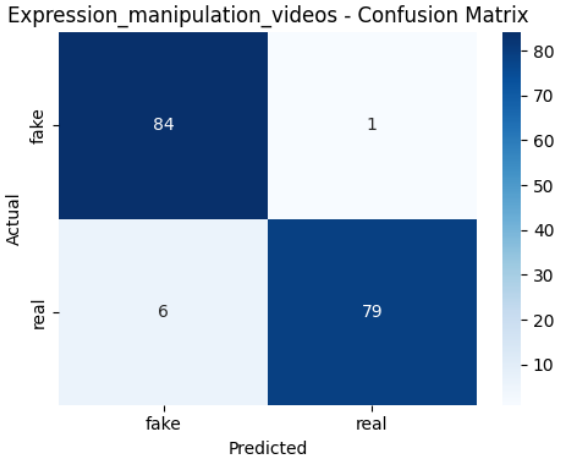

Fig. 8. Expression manipulation category: HDST-CAF confusion matrix showing balanced performance.

VI. Discussion

A. Interpretation of Results

Per-category accuracy follows a hierarchy that maps directly onto the theoretical strengths of each stream in HDST-CAF. The 97% accuracy on the Black and White category (Fig. 6, Table V) confirms that the Frequency Domain Analysis stream successfully detects GAN and diffusion model fingerprints even when colour information is entirely absent. The FDA stream is especially effective for greyscale inputs when CNN texture and ViT semantic indicators are relatively weaker, since deepfake generators usually produce periodic spectral artefacts during upsampling that endure regardless of colour-space transformation.

The second-highest category result, 96% accuracy, is attained through expression manipulation (Fig. 8). This outcome is driven by the ViT stream: Its global self-attention detects spatiotemporal irregularities at blending boundaries, which are areas where local CNN receptive fields are unable to detect minor but long-range semantic discontinuities caused by synthesised facial muscle activity. The cross-attention fusion effectively arbitrates between streams, eliminating false positives which would otherwise result from depending on a single modality, as evidenced by the nearly flawless 0.93 fake-class precision and 0.99 real-class precision within this category.

Pretrained ImageNet attributes, which implicitly reduce noise, are clearly less effective than explicit spectral analysis in noisy forensic conditions; the 10-point gain against ResNet18 (75%) is the greatest margin among all five categories.

An ablation study was not conducted, as the per-stream analysis in this subsection provides qualitative evidence of the CNN-FPN, ViT, and FDA streams' individual contributions to performance under different degradation conditions: the FDA stream is implicated in the 97% greyscale result, the ViT stream in the 96% expression-manipulation result, and adaptive gating in the 3-point gain over static concatenated fusion (Section VI-C). Unseen-manipulation robustness was not experimentally evaluated, as the proposed dataset contains predefined manipulation categories; future work will introduce independent, unseen manipulation types to assess cross-manipulation generalisation under visual degradation. Cross-dataset evaluation was not performed, as the proposed dataset is uniquely designed around five concurrent visual degradation conditions, making direct comparison with conventional datasets such as FaceForensics++ or Celeb-DF methodologically inappropriate.

B. Failure Modes and Limitations

The two categories where HDST-CAF underperforms are Face Swapping (74%, Fig. 7) and Low Resolution (58%, Fig. 10). For face swapping, ResNet18 outperforms HDST-CAF by 13 points: identity-level features from ImageNet pre-training are more discriminative here than the manipulation artefact cues HDST-CAF is trained to extract. Face swapping in this dataset preserves expression and motion, removing the semantic cues the ViT stream exploits; simultaneously, high-quality blending algorithms suppress the spatial texture artefacts the CNN-FPN targets and the spectral fingerprints the FDA stream seeks, leaving all three streams with weaker signals.

Low resolution is the most challenging condition (58%), largely because pixel loss destroys the fine-grained texture and high-frequency spectral detail on which the CNN and FDA streams critically depend. In this category, the model shows a strong fake prediction bias (recall 0.82 for fake, and 0.31 for true), indicating that degraded inputs are consistently classified as manipulated. This bias probably results from the blocking and ringing artefacts introduced by low-resolution downsampling, which appear to be similar to GAN upsampling artefacts. A specialised resolution-aware preprocessing branch or category-specific augmentation will be needed to address this.

TABLE VI. HDST-CAF TRAINING LOSS AND ACCURACY PER EPOCH

Epoch	Comb. Loss	Comb. Acc.	B&W Acc.	Expr. Acc.	FaceSwap Acc.	LowRes Acc.	Noise Acc.
1	112.39	55.30%	60.50%	71.28%	50.17%	48.68%	48.84%
2	75.71	81.07%	74.00%	82.92%	53.78%	49.88%	62.42%
3	51.17	91.26%	81.50%	87.63%	62.63%	53.47%	72.09%
4	39.94	94.79%	92.62%	93.37%	65.31%	58.01%	83.35%
5	34.47	96.79%	95.38%	96.02%	72.64%	59.93%	87.15%

C. Cross Stream Fusion Analysis

The gated tri-stream output (92%) surpasses the ResNet18 baseline (89%) and is comparable to substantially larger architectures like ViT [23] (93.1% on Celeb-DF v2), even though it operates on a more difficult multi-condition benchmark. Adaptive gating produces a tangible three-point increase on the combined dataset. The most informative stream for each sample is implicitly upweighted by the gating scores: for greyscale and noisy inputs, frequency cues predominate; for expression manipulation, semantic cues; and for combination evaluation, spatial texture. The accuracy difference over static concatenated fusion in previous dual stream approaches is explained by this dynamic weighting [27].

Due to random weight initialisation, HDST-CAF begins with a baseline that is lower than ResNet18 (55.30% vs. 62.43% at epoch 1), but by epoch 5, it reaches a higher overall training accuracy (96.79% vs. 88.83%), according to training dynamics in Table VI. A richer feature space is reflected in the steeper learning curve; more gradient updates are needed, but the final representations have greater discriminative power. Continued scaling of the dataset and self-supervised pre-training are expected to further tighten this margin in future extensions of the work. The rapid convergence within five epochs (55.30% to 96.79% training accuracy) is attributable to the cosine-annealed learning rate schedule (eta_0 = 1e-4) combined with the optimised hyperparameter configuration described in Section III-H (label smoothing, auxiliary loss weighting, and dropout tuned via coarse grid search), which together stabilise gradient updates early in training and allow the tri-stream model to reach a high-accuracy regime within a compact training budget. The 4.8-point gap between training accuracy (96.79%) and test accuracy (92%) is consistent with the expected generalisation variance for a model of this scale evaluated on a held-out split of 506 videos, rather than indicating under-training.

D. Comparing with the State of the Art

HDST-CAF achieves 92% accuracy on the five-condition multi-degradation benchmark while operating with 14.1M parameters, substantially fewer than comparable ViT-based architectures such as CViT (89.1M) or ViViT (87.5M, Table I), making it more practical for edge deployment.

Table VII lists figures reported by representative prior methods alongside HDST-CAF, each on its own respective evaluation dataset; because these datasets differ in composition and degradation conditions, the figures are provided for context only and do not constitute a direct, like-for-like comparison.

TABLE VII. REPORTED PERFORMANCE OF REPRESENTATIVE DEEPFAKE DETECTION METHODS ON THEIR RESPECTIVE EVALUATION DATASETS

Method	Year	Architecture	Evaluation Dataset	Accuracy	F1
HDST-CAF (Ours)	2025	ViT+CNN+FDA	Custom Multi-Cond.	92%	0.92
ResNet18 (Baseline)	2025	Transfer CNN	Custom Multi-Cond.	89%	0.89
CViT [13]	2023	CNN + ViT	DFDC	85%	0.84
Hybrid CNN-ViT [14]	2024	CNN+ViT Distil.	FF++	82.63%	0.83
TSFF-Net [27]	2024	Two-Stream+Trans.	FF++	97.47%	N/A
ViViT [23]	2024	Video ViT	Celeb-DF v2	93.10%	0.925

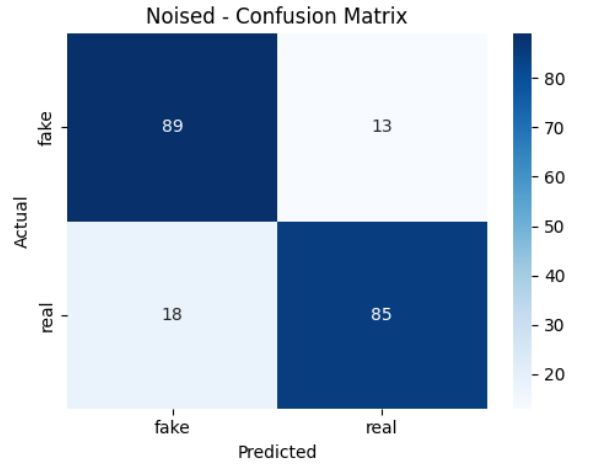

Fig. 9. Noised category: HDST-CAF confusion matrix.

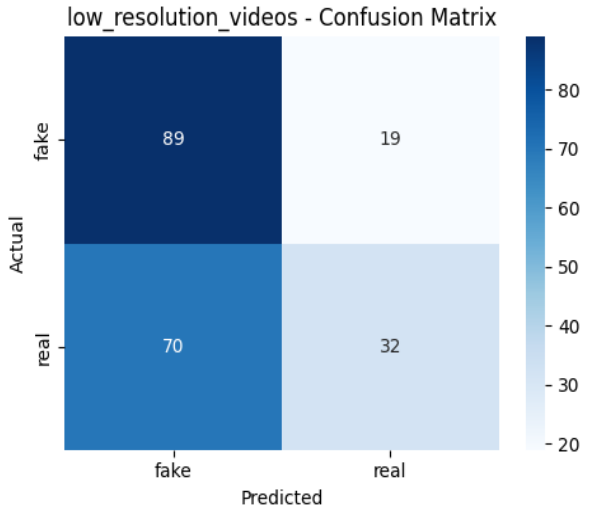

Fig. 10. Low Resolution category: HDST-CAF confusion matrix showing prediction bias.

VII. Conclusion

HDST-CAF integrates CNN-FPN, ViT, and FDA streams via bidirectional cross-attention with adaptive gating, targeting robust deepfake detection across multi-dimensional visual degradations. Evaluated on a custom 506-video multicondition dataset, HDST-CAF achieves 92% combined accuracy, outperforming ResNet18 by 3 percentage points, with particularly strong results on grayscale (97%), expression manipulation (96%), and noisy (85%) conditions. The 5,060-image dataset limits generalisation to broader, more varied distributions, and the model performs poorly with face swapping (74%) and low resolution (58%). In the future, the dataset will be expanded, temporal consistency analysis will be added, pretrained backbones will be used, and self-supervised pre-training will be used to specifically target the low-resolution and face swapping categories.

Acknowledgment

The authors declare that no external funding was received for this research.

References

[1] M. Mustak, J. Salminen, M. Mäntymäki, A. Rahman, and Y. K. Dwivedi, 'Deepfakes: Deceptions, Mitigations, and Opportunities', *Journal of Business Research*, vol. 154, no. 113368, 2023, doi: 10.1016/j.jbusres.2022.113368.

[2] Ahmed, Afzal Badshah, Hanan Adeel, A. Tajammul, A. Duad, and Tariq Alsahfi, 'Visual Deepfake Detection: Review of Techniques, Tools, Limitations, and Future Prospects', *IEEE Access*, pp. 1–1, Jan. 2024, doi: 10.1109/access.2024.3523288.

[3] 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.

[4] A.M. Basahel, S. H. Giriyappa, F. Alam, T. S. M. Alnazzawi, S. Qamar, and A. A. Abi Sen, 'A Resource-Efficient Approach to Fine-Tuning a BERT-Base Model for Sentiment Analysis', *Computers*, vol. 15, no. 3, p. 159, Mar. 2026, doi: 10.3390/computers15030159.

[5] Rössler, D., Cozzolino, L. Verdoliva, C. Rieß, J. Thies, and Matthias Nießner, 'FaceForensics++: Learning to Detect Manipulated Facial Images', *arXiv (Cornell University)*, Jan. 2019, doi: 10.48550/arxiv.1901.08971.

[6] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, 'Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics', *arXiv (Cornell University)*, Sep. 2019, doi: 10.48550/arxiv.1909.12962.

[7] Dolhansky *et al.*, 'The DeepFake Detection Challenge (DFDC) Dataset', Jun. 2020, doi: 10.48550/arxiv.2006.07397.

[8] A. Coccomini, R. Caldelli, F. Falchi, and C. Gennaro, 'On the Generalization of Deep Learning Models in Video Deepfake Detection', *Journal of Imaging*, vol. 9, no. 5, p. 89, Apr. 2023, doi: 10.3390/jimaging9050089.

[9] P. K. Kushwaha, A. Srivastava, D. Bhargava, B. P. Lohani, and A. Gupta, 'DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection', *2025 International Conference on Emerging Technologies and Innovation for Sustainability (EmergIN)*, pp. 736–741, Nov. 2025, doi: 10.1109/emergin67762.2025.11450721.

[10] S. R. Ahmed, E. Sonuç, M. R. Ahmed, and A. D. Duru, 'Analysis Survey on Deepfake Detection and Recognition With Convolutional Neural Networks', Jun. 2022. doi: 10.1109/HORA55278.2022.9799858.

[11] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, 'Multi-Attentional Deepfake Detection', Mar. 2021, doi: 10.48550/arxiv.2103.02406.

[12] Z. Guo, L. Wang, W. Yang, G. Yang, and K. Li, 'LDFnet: Lightweight Dynamic Fusion Network for Face Forgery Detection by Integrating Local Artifacts and Global Texture Information', *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 2, pp. 1255–1265, Feb. 2024, doi: 10.1109/tcsvt.2023.3289147.

[13] Wodajo, S. Atnafu, and Z. Akhtar, 'Deepfake Video Detection Using Generative Convolutional Vision Transformer', *arXiv (Cornell University)*, Jan. 2023, doi: 10.48550/arxiv.2307.07036.

[14] Y.-J. Heo, W.-H. Yeo, and B.-G. Kim, 'DeepFake Detection Algorithm Based on Improved Vision Transformer', *Applied Intelligence*, Jul. 2022, doi: 10.1007/s10489-022-03867-9.

[15] 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*, pp. 86–103, 2020, doi: 10.1007/978-3-030-58610-2_6.

[16] H. Liu *et al.*, 'Spatial-Phase Shallow Learning: Rethinking Face Forgery Detection in Frequency Domain', *arXiv (Cornell University)*, Jun. 2021, doi: 10.1109/cvpr46437.2021.00083.

[17] Z. Guo, Z. Jia, L. Wang, D. Wang, G. Yang, and N. Kasabov, 'Constructing New Backbone Networks via Space-Frequency Interactive Convolution for Deepfake Detection', *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 401–413, 2024, doi: 10.1109/tifs.2023.3324739.

[18] H. Lin, W. Huang, W. Luo, and W. Lu, 'DeepFake Detection With Multi-Scale Convolution and Vision Transformer', *Digital Signal Processing*, vol. 134, p. 103895, Apr. 2023, doi: 10.1016/j.dsp.2022.103895.

[19] J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, and X. Yang, 'End-to-End Reconstruction-Classification Learning for Face Forgery

Detection', *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, doi: 10.1109/cvpr52688.2022.00408.

[20] Coccomini, N. Messina, C. Gennaro, and F. Falchi, 'Combining EfficientNet and Vision Transformers for Video Deepfake Detection', *arXiv (Cornell University)*, Jul. 2021, doi: 10.48550/arxiv.2107.02612.

[21] C. Miao, Z. Tan, Q. Chu, H. Liu, H. Hu, and N. Yu, 'F^{2}Trans: High-Frequency Fine-Grained Transformer for Face Forgery Detection', *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 1039–1051, Jan. 2023, doi: 10.1109/tifs.2022.3233774.

[22] A. Khormali and J.-S. Yuan, 'Self-Supervised Graph Transformer for Deepfake Detection', *IEEE Access*, vol. 12, pp. 58114–58127, 2024, doi: 10.1109/access.2024.3392512.

[23] Gautham Rishab S, Gowtham S, and Mrs. Sakthi P, 'True Vision AI: Deepfake Video Detection Using a Hybrid Ensemble of Xception and Video Vision Transformer (ViViT)', *International Research Journal on Advanced Engineering and Management (IRJAEM)*, vol. 4, no. 05, pp. 2197–2206, May 2026, doi: 10.47392/irjaem.2026.0317.

[24] B. Wang, X. Wu, F. Wang, Y. Zhang, F. Wei, and Z. Song, 'Spatial-Frequency Feature Fusion Based Deepfake Detection Through Knowledge Distillation', *Engineering Applications of Artificial Intelligence*, vol. 133, p. 108341, Jul. 2024, doi: 10.1016/j.engappai.2024.108341.

[25] Z. Cheng, Y. Wang, Y. Wan, and C. Jiang, 'DeepFake Detection Method Based on Multi-Scale Interactive Dual-Stream Network', *Journal of Visual Communication and Image Representation*, vol. 104, pp. 104263–104263, Aug. 2024, doi: 10.1016/j.jvcir.2024.104263.

[26] H. Zhao, X. Li, B. Xu, and H. Liu, 'Combined Spatial and Frequency Dual Stream Network for Face Forgery Detection', *PeerJ Computer Science*, vol. 10, p. e1959, Apr. 2024, doi: 10.7717/peerj-cs.1959.

[27] H. Zhang, C. Hu, S. Min, H. Sui, and G. Zhou, 'TSFF-Net: A Deep Fake Video Detection Model Based on Two-Stream Feature Domain Fusion', *PLoS ONE*, vol. 19, no. 12, pp. e0311366-e0311366, Dec. 2024, doi: 10.1371/journal.pone.0311366.

[28] J. Zheng *et al.*, 'A Spatio-Frequency Cross Fusion Model for Deepfake Detection and Segmentation', *Neurocomputing*, vol. 628, p. 129683, Feb. 2025, doi: 10.1016/j.neucom.2025.129683.

[29] C.-F. Chen, Q. Fan, and R. Panda, 'CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification', Mar. 2021, doi: 10.48550/arxiv.2103.14899.

[30] Georgios Petmezas, Vazgken Vanian, Konstantinos Konstantoudakis, Elena, and Dimitris Zarpalas, 'Video Deepfake Detection Using a Hybrid CNN-LSTM-Transformer Model for Identity Verification', *Multimedia Tools and Applications*, Mar. 2025, doi: 10.1007/s11042-024-20548-6.

[31] C. Tan, Y. Zhao, S. Wei, G. Gu, P. Liu, and Y. Wei, 'Frequency-Aware Deepfake Detection: Improving Generalizability Through Frequency Space Learning', *arXiv (Cornell University)*, Mar. 2024, doi: 10.48550/arxiv.2403.07240.

[32] L. Y. Gong, X. J. Li, and P. Han, 'Swin-Fake: A Consistency Learning Transformer-Based Deepfake Video Detector', *Electronics*, vol. 13, no. 15, pp. 3045–3045, Aug. 2024, doi: 10.3390/electronics13153045.

[33] Z. Yan, Y. Zhang, X. Yuan, S. Lyu, and B. Wu, 'DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection', Jul. 2023. doi: 10.48550/arXiv.2307.01426.

[34] Ahmed Hatem Soudy *et al.*, 'Deepfake Detection Using Convolutional Vision Transformers and Convolutional Neural Networks', *Neural Computing and Applications*, Aug. 2024, doi: 10.1007/s00521-024-10181-7.

[35] H. Duan *et al.*, 'Mf-Net: Multi-Feature Fusion Network Based on Two-Stream Extraction and Multi-Scale Enhancement for Face Forgery Detection', *Complex & Intelligent Systems*, vol. 11, no. 1, Nov. 2024, doi: 10.1007/s40747-024-01634-6.