

Advanced AI-Based Identification of Deepfake Content

¹Divyanshu Goyal, ² Mohnish Sachdeva

¹Student, Department of Artificial Intelligence and Data Science, Poornima Institute of Engineering and Technology, Jaipur, India

²Assistant Professor, Department of Artificial Intelligence and Data Science, Poornima Institute of Engineering and Technology, Jaipur, India

Abstract - Rapid progress in generative models, including generative adversarial networks (GANs) and diffusion models, has made it easy to synthesise face images and videos that people can no longer reliably tell apart from authentic recordings. This creates risks for misinformation, identity fraud, and the integrity of digital evidence. Detectors trained on a single dataset or generator often degrade on unseen manipulations and on compressed media, and many give no human-readable justification for their decisions. This paper proposes a unified, explainable framework for identifying deepfake images and videos. The framework combines (i) a transfer-learned Xception backbone for spatial artifact extraction from aligned face crops, (ii) a bidirectional LSTM with temporal attention for inter-frame inconsistencies in video, and (iii) a frequency-domain and statistical branch that captures spectral traces of generative up-sampling. A learned gated fusion module merges the three streams for real/fake classification, and Grad-CAM, temporal attention weights, and SHAP attributions provide complementary visual and feature-level evidence. We also specify a cross-dataset and robustness evaluation protocol covering compression, resizing, noise, unseen generators, and class imbalance. This is a design and methodology study: no new experiments were conducted, and every performance figure quoted is attributed to the study that reported it. The work consolidates recent literature into a coherent and reproducible blueprint for transparent deepfake forensics.

Keywords - Deepfake detection, deep learning, Xception, spatiotemporal analysis, frequency-domain analysis, explainable AI, Grad-CAM, image and video forensics.

I. INTRODUCTION

Generative models have progressed from producing blurry samples to synthesising photorealistic faces, speech, and video. Generative adversarial networks (GANs) [1] pit a generator against a discriminator, variational autoencoders (VAEs) [2] learn latent face representations that can be recombined, and denoising diffusion models [3] synthesise images by iteratively removing noise. These

families underpin face swapping, expression reenactment, and fully synthetic imagery, collectively known as deepfakes [4], [5]. In this paper a deepfake denotes image or video content in which a person's appearance, expression, or identity has been altered or synthesised by a learned model so that it appears authentic.

The same capabilities that benefit film production and accessibility also enable misuse: political misinformation, non-consensual imagery, impersonation and financial fraud, and the erosion of trust in photographs and video as evidence [4], [6], [7]. Freely available face-swapping applications have lowered the skill needed to produce convincing forgeries [7], so journalists, courts, and platform moderators now need dependable automated detection.

Early forensic techniques relied on hand-crafted cues such as blink rate, head-pose inconsistency, colour-filter-array traces, or compression analysis [5], [8]. These cues are brittle: generators quickly learn to suppress a known artifact, and hand-crafted features seldom survive the re-compression applied by social platforms. Deep learning replaces manual feature design with representations learned from data [5], and CNN-based detectors have reported strong results on benchmark datasets [7], [9], [10].

Three challenges nevertheless remain. The first is generalisation: detectors tend to overfit the generator and dataset used for training and degrade on unseen manipulations, a limitation repeatedly noted in recent studies [5], [6], [11]. The second is robustness: compression, resizing, and noise suppress the subtle high-frequency traces on which many detectors depend [6], [10]. The third is transparency: deep detectors behave as black boxes, which limits their acceptability in forensic and legal settings where a decision must be justified [6], [12].

Motivation and objectives. Spatial, temporal, and spectral evidence are complementary, yet most published detectors exploit only one or two of them (Section II). Our objectives are: (1) to design a single framework that fuses spatial, temporal, and frequency-domain evidence; (2) to

make the fusion adaptive, so that the model can down-weight a cue that is unreliable for a given input, for example spectral traces after heavy compression; (3) to attach explanations that are computed from, and consistent with, the model's own decision path; and (4) to define an evaluation protocol that measures generalisation and robustness rather than in-dataset accuracy alone.

Contributions. The main contributions of this paper are:

- 1) A critical synthesis of five recent studies [5]–[7], [11], [13] together with foundational work, leading to an explicit list of research gaps (Sections II and III).
- 2) A multi-stream detection architecture that couples an Xception spatial stream, a Bi-LSTM temporal stream with attention, and a frequency/statistical stream through a modality-aware gated fusion module that handles both images and videos.
- 3) An explainability design that combines Grad-CAM, temporal attention weights, and SHAP attributions over low-dimensional feature groups, together with quantitative checks of explanation quality.
- 4) A reproducible experimental protocol covering cross-dataset, cross-generator, compression, resizing, noise, and class-imbalance analysis, with explicit separation of literature-reported results from expected outcomes.

We stress that no experimental results are claimed for the proposed framework. The remainder of the paper is organised as follows. Section II reviews related work, Section III identifies research gaps, Sections IV and V present the methodology and system architecture, Sections VI and VII describe the experimental design and metrics, Section VIII discusses literature-reported and expected outcomes, and Sections IX to XI cover limitations, future work, and conclusions.

II. RELATED WORK

A. CNN-Based and Xception-Based Detection

CNNs are the dominant backbone for deepfake detection because they capture local texture and blending irregularities in face regions. MesoNet [9] showed that a compact network operating at a mesoscopic level can detect face tampering at low computational cost. Xception [14] replaces standard convolutions with depthwise separable convolutions and became a reference backbone in the FaceForensics++ benchmark [10]. Khatri and Gupta [7] fine-tune an ImageNet-pretrained Xception with an added dense head for AI-generated face images and report a validation accuracy of 93.37% on the Deepfake_Faces

dataset, crediting transfer learning and fine-tuning for the gain. The strengths of this line of work are parameter efficiency and effective transfer learning. Its limitations are image-level decisions without temporal modelling, evaluation on few datasets, and little explanation of individual decisions. Xception therefore serves as the spatial backbone of our framework.

B. Spatial–Temporal Detection

Frame-by-frame manipulation leaves temporal traces such as flicker, unnatural blinking, and lip-motion inconsistency. Güera and Delp [15] paired a CNN frame encoder with an LSTM [16] to model such inconsistencies, a pattern followed by many later studies [5]. Lee and Kim [17] analyse the rate of change between frames, and Sagar and Arukonda [18] propose a CNN-LSTM variant; the latter is described in [6] as computationally heavy, dependent on large labelled video sets, and sensitive to video quality. Nandhini et al. [6] combine a CNN-LSTM with Grad-CAM and SHAP on material drawn from FaceForensics++, DFDC, and Celeb-DF and deploy it through a web interface. Their evaluation, however, uses a 100-video test set and the reported accuracy values are not consistent across the text and tables (0.75 in the metric tables, 95% in the narrative), so the quantitative evidence is hard to interpret. The study is nonetheless a useful template for coupling temporal modelling with explanations.

C. Frequency-Domain and Statistical Approaches

Up-sampling layers in generators leave periodic spectral artifacts. Durall et al. [19] showed that GAN-generated images systematically fail to reproduce the spectral distribution of real images, and Frank et al. [20] demonstrated that DCT-domain representations support reliable detection. F3-Net [21] mines frequency-aware clues and is particularly helpful on compressed media, while Giudice et al. [22] identify GAN-specific DCT anomalies and Sadhya and Qi [23] integrate wavelet transforms into CNNs. Zaidi et al. [11] propose a hybrid pipeline that merges deep features with statistical analysis (wavelet, frequency-domain, and anomaly-detection measures). Frequency cues are cheap and relatively interpretable and complement semantic cues, but spectral signatures are generator-specific and fragile under resizing and compression, and [11] gives limited detail on how the statistical features are fused with the learned ones.

D. Transformer-Based Detection

Vision Transformers (ViTs) [24] split an image into patches and model long-range dependencies by self-attention, which may capture global inconsistencies such as lighting or face-boundary blending that local convolutions miss. M2TR [25] applies multi-scale transformers with frequency modelling to forgery detection. Zaidi et al. [11] use a ViT with 16×16 patches and report ablations over patch size, depth, number of attention heads, dropout, and augmentation; a hybrid ViT-plus-SVM image-forgery framework is summarised in [5]. Transformers are data-hungry and costly, and the ablations in [11] rely on a single image dataset. We therefore retain transformers as an alternative temporal encoder and future backbone rather than a core component.

E. GAN- and Diffusion-Generated Image Detection

Wang et al. [26] showed that a classifier trained on one GAN with careful augmentation generalises surprisingly well to other CNN-based generators. Diffusion models change the picture: Corvi et al. [27] report that detectors trained on GAN images transfer poorly to diffusion-generated images, and Ojha et al. [28] propose classifying with frozen features of a large pretrained vision model to generalise across generators. Dedicated datasets now exist, including CIFAKE [29] and a diffusion-based face set summarised in [5] and [30]. The review in [5] also notes that human observers misclassify a substantial share of AI-generated images, underlining the difficulty of the task.

F. Explainable AI for Deepfake Detection

Grad-CAM [31] localises the image regions that drive a class score, and SHAP [32] attributes a prediction to input features through Shapley values. Studies on explainable deepfake detection include [33] and the multimodal survey of Momin et al. [12], which, as summarised in [6], stresses the difficulty of producing consistent explanations and of balancing explainability against accuracy. CIFAKE [29] also pairs detection with explanation. Saliency maps can be unfaithful to the model, however [34], so explanations need quantitative validation, which most deepfake studies omit.

G. Hybrid Frameworks and Image Forgery Detection

Abdelmaksoud et al. [5] review manipulated-image forgeries (splicing, copy-move) and AI-generated images (VAE, GAN, diffusion) and conclude that most detectors

target a specific forgery type and generalise poorly to unseen generators. Chen et al. [35] use a multi-graph attention network and report 97.89% accuracy on CIFAKE, as summarised in [11]. Ravi Kumar et al. [13] embed a deepfake checker in a full-stack image platform that relies on external AI services and report 95.2% accuracy on FaceForensics++ and Celeb-DF; the detector architecture and protocol are not specified, which limits reproducibility, but the work illustrates the deployment context for such detectors. Table I summarises the approaches discussed.

TABLE I COMPARISON OF REPRESENTATIVE DEEPPFAKE DETECTION APPROACHES

Ref.	Approach	Strengths	Limitations (as reported or observed)	Relevance here
[7]	Fine-tuned Xception, image-level	Transfer learning; low parameter count	No temporal cues; few datasets; little explanation	Spatial backbone
[6]	CNN-LSTM with Grad-CAM and SHAP (video)	Temporal modelling; explanations; web interface	Small test set; inconsistent reported metrics; extra compute	Temporal and XAI template
[11]	ViT with statistical analysis (hybrid)	Global attention; spectral/statistical cues; ablations	Single image dataset; fusion details limited; weak on high-quality fakes	Hybrid-fusion motivation
[13]	Platform with externally served detector	Deployment and scalability focus	Detector architecture and protocol unspecified	Deployment context
[5]	Review: CNN, transfer learning, ViT+SVM, ensembles	Broad taxonomy and dataset survey	Methods target specific forgery types; weak generalisation	Gap evidence; datasets
[21], [19]	Frequency-aware and spectral analysis	Robust on compressed media; detects up-sampling traces	Generator-specific; fragile under resizing	Frequency stream
[26], [28]	Cross-generator generalisation of image detectors	Good transfer across GANs (augmentation); frozen pretrained features	Diffusion and compression remain difficult	Generalisation protocol

III. RESEARCH GAP

The literature reviewed above, and the five uploaded studies in particular, point to the following gaps, which motivate the proposed framework.

G1. Generalisation to unseen generators and datasets. Detectors trained on one source degrade on others [5], [6], [11]; GAN-trained detectors transfer poorly to diffusion images [27].

G2. High-quality and diffusion-generated content. Photorealistic outputs contain fewer artifacts, and [11] itself lists highly sophisticated GAN output as a limitation; diffusion-based face sets are comparatively new [5], [30].

G3. Robustness to compression and quality changes. Video-quality sensitivity is a recurring weakness [6],

[18], and compression suppresses the traces many detectors use [10], [21].

G4. Fragmented use of cue types. Image-level detectors such as [7] and [11] ignore temporal evidence, while video-level detectors such as [6] rarely exploit spectral cues; [11] observes that robust fusion of learned and statistical features is still limited.

G5. Explainability without validation. Several detectors give no justification for decisions [11], [36], and where explanations exist they are seldom evaluated quantitatively [12], [34].

G6. Dataset imbalance, bias, and evaluation reliability. Benchmarks are skewed toward fakes (for example Celeb-DF has 590 real and 5,639 fake videos [37]), demographic fairness varies across detectors [38], and some reported evaluations use very small test sets or contain internally inconsistent figures [6], [7].

G7. Computational cost and deployment. Complex detectors incur high compute [6], [18], and lightweight, real-time models are still needed [5].

The framework proposed in the next section targets gaps G1 to G5 directly in its design and gaps G6 and G7 through its evaluation protocol and implementation configuration.

IV. PROPOSED METHODOLOGY

A. System Overview

The framework accepts an image or a video clip and returns a real/fake decision, a calibrated confidence, and an evidence report. After preprocessing and face extraction, three parallel streams produce feature vectors of equal dimension: a spatial stream (z_s) that detects local artifacts, a temporal stream (z_t) that models inter-frame inconsistency, and a frequency/statistical stream (z_f) that captures spectral and noise irregularities. A gated fusion module weights and merges the streams, an MLP classifier produces the fake probability, and an explainability module attributes the decision to image regions, frames, and spectral features. For still images the temporal stream is masked. The end-to-end architecture is shown in Fig. 1 (Section V).

B. Data Collection and Dataset Preparation

The framework is designed for publicly available benchmarks with complementary properties: FaceForensics++ [10] for multi-method training and compression levels, Celeb-DF [37] and DFDC [39] for cross-dataset testing, and diffusion-oriented image sets such as DFF and CIFAKE [29], [30] for generator-shift testing (Section VI, Table IV). Splits are made at the identity/video level so that no subject appears in more than one partition, preventing identity leakage. Frame-level labels are inherited from the parent video. These datasets are proposed for future experiments; we make no claim that any experiment has been run on them.

C. Data Preprocessing

Frame extraction. For a video, $T = 16$ frames are sampled uniformly over the clip. **Face detection and alignment.** Faces are located with MTCNN [40], cropped with a $1.3\times$ margin so that the blending boundary stays visible, and aligned using the detected landmarks. **Resizing and normalisation.** Crops are resized to 299×299 pixels, the native Xception input size [14], and pixel values are scaled to $[-1, 1]$. **Quality normalisation.** Colour space is standardised to sRGB and gross illumination differences are reduced by gain normalisation. We deliberately avoid denoising and

smoothing filters, which some pipelines apply to improve visual quality [6], because they suppress the high-frequency traces exploited by the frequency stream. **Augmentation.** Training uses horizontal flips, random crop and scale, colour jitter, Gaussian blur and noise, and simulated JPEG and H.264 compression. Compression and resizing augmentation targets robustness, in line with the practice of [26].

D. Spatial Feature Extraction

The spatial stream uses an Xception network [14] initialised from ImageNet weights [41]. Xception stacks depthwise separable convolutions in entry, middle, and exit flows with residual connections, and has roughly 20.9 million parameters without its classification head [7]. Global average pooling yields a 2048-dimensional vector $\mathbf{f}_t$ per frame. The mean over frames is projected by a fully connected layer to $\mathbf{z}_s \in \mathbb{R}^{512}$. Training is two-stage: the backbone is frozen while the new layers converge, then the last blocks are unfrozen with a lower learning rate, a fine-tuning practice also adopted in [7].

E. Temporal Feature Analysis

For video, the per-frame vectors $\mathbf{f}_1, \dots, \mathbf{f}_T$ are passed to a bidirectional LSTM [16] with 256 hidden units per direction, producing $\mathbf{h}_t \in \mathbb{R}^{512}$. Because inconsistencies are often confined to a few frames, a soft attention layer pools the sequence:

$$\alpha_t = \frac{\exp(\mathbf{w}^T \tanh(\mathbf{W}\mathbf{h}_t))}{\sum_{t=1}^T \exp(\mathbf{w}^T \tanh(\mathbf{W}\mathbf{h}_t))} \quad (1)$$

and the temporal vector is $\mathbf{z}_t = \sum_t \alpha_t \mathbf{h}_t$, projected to $\mathbb{R}^{512}$. The weights α_t double as a frame-importance signal for explanation. A two-layer Transformer encoder with eight heads is a candidate replacement in the ablation plan. Frames without a detected face are masked out of the attention. For still images, $\mathbf{z}_t$ is set to zero and its gate is masked.

F. Frequency-Domain and Statistical Analysis

GAN up-sampling produces characteristic spectral artifacts [19], [20], whereas diffusion images show different traces [27]. The frequency stream therefore supplies complementary rather than sufficient evidence. For the grey-level crop $g(x,y)$ of size $H\times W$, the 2-D discrete Fourier transform is

$$F(u,v) = \sum_{x=0}^{H-1} \sum_{y=0}^{W-1} g(x,y) e^{-j2\pi(ux/H + vy/W)} \quad (2)$$

The shifted log-magnitude spectrum is reduced to a one-dimensional radial profile by azimuthal averaging over rings R_r of equal radius:

$$P(r) = \frac{1}{|R_r|} \sum_{(u,v) \in R_r} \log(1 + |F(u,v)|) \quad (3)$$

with $r = 1, \dots, 64$ radial bins. Sixteen statistical descriptors are appended: high-frequency energy ratio, spectral flatness, and the mean, variance, skewness, and kurtosis of a high-pass residual, computed globally and over the eye, mouth, and boundary regions. A one-class anomaly score is added by modelling the descriptor vector $\mathbf{d}$ of authentic training faces with mean μ_R and covariance Σ_R , in the spirit of one-class detection discussed in [5]:

$$s_a = \sqrt{(\mathbf{d} - \mu_R)^T \Sigma_R^{-1} (\mathbf{d} - \mu_R)} \quad (4)$$

Profile, descriptors, and score are standardised with training-set statistics and mapped by a small MLP (128 units) and a linear layer to $z_f \in \mathbb{R}^{512}$. Spectral traces are fragile under compression, so the fusion stage is allowed to down-weight this stream.

G. Feature Fusion

Each stream is projected to a common 512-dimensional space, giving $\hat{z}_s$, $\hat{z}_v$, and $\hat{z}_f$ (the projected vectors in Fig. 2). A gating network computes per-sample modality weights:

$$g = \text{softmax}(W_g [\hat{z}_s; \hat{z}_v; \hat{z}_f] + b_g) \quad (5)$$

and the fused representation is $z = \sum_m g_m \hat{z}_m$ over $m \in \{s, t, f\}$. For images the temporal logit is masked before the softmax. During training, whole streams are randomly dropped (modality dropout, probability 0.15) so that the model learns to operate with missing or unreliable inputs. Compared with plain concatenation, gating exposes which cue drove each decision, which also feeds the explanation report.

H. Classification

The fused vector passes through a 256-unit fully connected layer with ReLU and dropout (0.5), then a sigmoid output giving $p_{\text{fake}} = P(\text{fake} | \text{input})$. Training minimises a class-weighted binary cross-entropy with optional auxiliary heads on each stream so that no stream is ignored:

$$L = -\frac{1}{N} \sum_{i=1}^N [w_1 y_i \log p_i + w_0 (1 - y_i) \log (1 - p_i)] + \lambda \sum_m L_m^{\text{aux}} \quad (6)$$

where w_1 and w_0 are inverse-frequency class weights that counter imbalance. Video-level scores average the calibrated clip scores. A decision threshold τ is selected on the validation set for a target operating point (for example a fixed false positive rate), and temperature scaling is used for calibration. An abstention band around τ labels uncertain cases as inconclusive and routes them to human review.

I. Explainable AI

Three complementary explanations are produced (Fig. 3). First, Grad-CAM [31] computes, for the fake logit y^c and the feature maps A^k of the last Xception block, channel weights and a localisation map:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (7)$$

$$L^c = \text{ReLU}(\sum_k \alpha_k^c A^k) \quad (8)$$

Frame maps are aggregated using the temporal attention weights α_t , which also rank key frames. Second, SHAP [32] assigns each feature j of a feature set F a Shapley value:

$$\varphi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f(S \cup \{j\}) - f(S)] \quad (9)$$

Because pixel-level SHAP is expensive, it is applied to low-dimensional feature groups: the radial-profile bins, the statistical descriptors, and the three fused streams. This indicates which spectral bands or which stream pushed the decision. Third, the gate weights g_m report the modality contributions. Explanation quality is assessed with deletion and insertion curves, comparison with ground-truth manipulation masks where available, and parameter-randomisation sanity checks [34]. Explanations show what drives the model, not proof of manipulation, and are reported as such.

J. Robustness and Generalisation

Generalisation is addressed by (i) pretrained initialisation and multi-method training, (ii) leave-one-manipulation-out validation so that model selection rewards transfer, (iii) compression, resizing, and noise augmentation, (iv) modality dropout and gated fusion, and (v) calibration with threshold selection under domain shift. Evaluation uses the perturbations in Table II, applied at test time to every compared model. Adversarial robustness is probed with white-box gradient attacks [42]; adversarial training is left to future work.

TABLE II PLANNED TEST-TIME PERTURBATIONS (DESIGN VALUES)

Perturbation	Planned levels	Purpose
JPEG compression	Quality factor 90, 70, 50, 30	Social-media re-compression
Resize (down then up)	Scale 0.75, 0.5, 0.25	Loss of high-frequency detail
Gaussian noise / blur	Increasing standard deviation (three levels each)	Sensor noise and defocus
H.264 video compression	CRF 23, 28, 33, 40	Video-codec degradation
Colour and brightness shift	Mild and strong settings	Lighting and camera variation
Gradient-based attack	Small-budget white-box perturbation	Adversarial vulnerability probe

V. SYSTEM ARCHITECTURE

A. Overall Pipeline

Fig. 1 shows the end-to-end flow: input → preprocessing → face/frame extraction → spatial feature extraction → temporal/frequency analysis → feature fusion → deepfake classification → explainability → final decision. The stages are:

- 1) the input image or video is received and decoded;
- 2) preprocessing standardises size, range, and colour space;
- 3) frames are sampled and faces detected and aligned;

Image mode: temporal stream bypassed (gate masked)

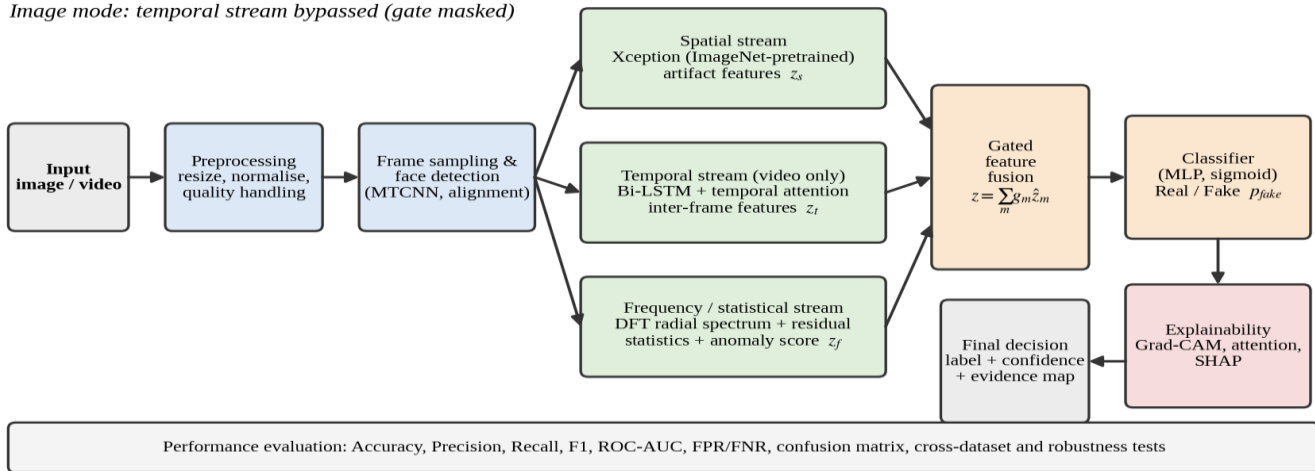


Fig. 1. Proposed framework for advanced AI-based identification of deepfake content, from input to explainable decision. In image mode the temporal stream is bypassed.

B. Feature Extraction and Fusion Architecture

Fig. 2 details the layer dimensions. Xception features feed both the mean-pooled spatial path and the temporal path, so

- 4) the Xception stream extracts spatial features per frame;
- 5) the Bi-LSTM stream models temporal dynamics (video only);
- 6) the frequency/statistical stream computes spectral and noise features;
- 7) the gated module fuses the three feature vectors;
- 8) the classifier outputs p_{fake} and the real/fake label;
- 9) the explainability module produces heatmaps, key frames, and feature attributions;
- 10) the decision, confidence, and evidence are reported, and performance is evaluated offline with the metrics of Section VII.

the backbone is shared and computed once per frame. The frequency path starts from the grey-level crop and is cheap to compute. Gated fusion yields a 512-dimensional vector for

the classifier. The dimensions in Fig. 2 and Table III are initial design choices, not tuned values.

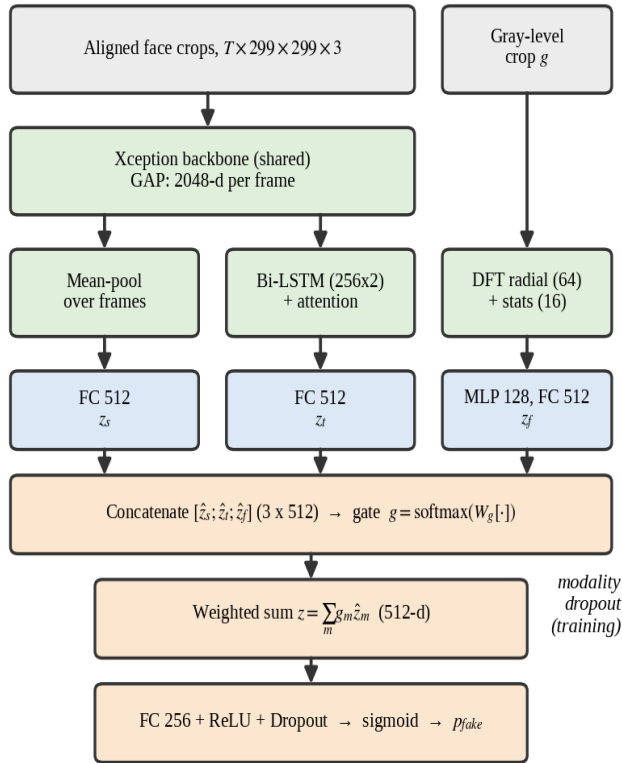


Fig. 2. Feature extraction and fusion architecture with layer dimensions.

C. Explainability Workflow

The explainability workflow (Fig. 3) runs after classification. It reuses the detector's activations and gradients (Grad-CAM), its temporal attention weights, and its gate weights, and adds SHAP on low-dimensional features. The evidence composer overlays heatmaps on the face, ranks key frames, and lists the spectral bands and streams that contributed most. Reports are generated for human review and flagged inconclusive when the score lies in the abstention band.

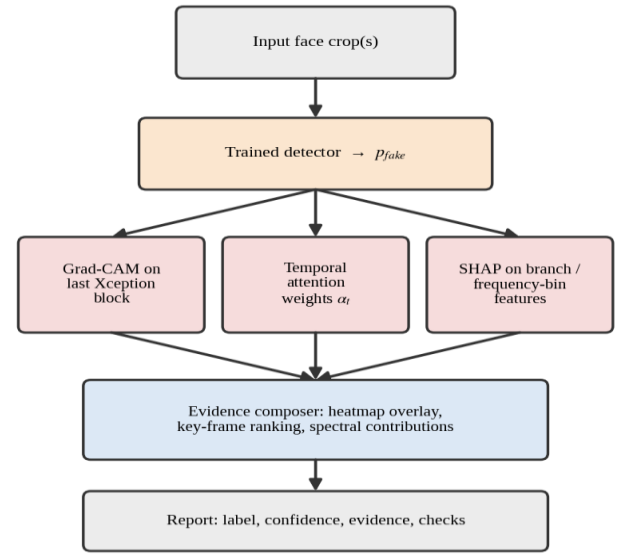


Fig. 3. Explainable AI workflow producing visual and feature-level evidence.

D. Implementation Configuration

Table III lists the proposed starting configuration. Because no experiments were run, these values are design defaults to be tuned on a validation split; they are partly informed by common practice and by settings reported in [6], [7], [11].

TABLE III PROPOSED INITIAL CONFIGURATION (NOT TUNED)

Component	Initial setting
Input	16 frames per clip, 299×299 face crops
Spatial backbone	Xception, ImageNet initialisation, two-stage fine-tuning
Temporal encoder	Bi-LSTM, 256 units per direction, attention pooling
Frequency stream	64 radial bins + 16 statistics + anomaly score
Fusion	Gated, 512-d, modality dropout 0.15
Classifier	FC 256, ReLU, dropout 0.5, sigmoid
Optimiser	Adam-type, lower rate for backbone than heads
Loss	Class-weighted BCE + auxiliary stream losses
Early stopping	On leave-one-method-out validation AUC

The model can be wrapped in a lightweight web service, as is done in [6] and [13], to accept uploads and return decisions and evidence.

VI. DATASETS AND EXPERIMENTAL DESIGN

This section describes a proposed experimental design. No experiments have been performed by the authors, and Table IV lists dataset properties as reported in the cited sources. Details should be re-verified against the official releases before use.

A. Candidate Datasets

FaceForensics++ [10] contains 1,000 original videos manipulated with four methods (DeepFakes, Face2Face,

FaceSwap, NeuralTextures) at raw, high-quality (c23), and low-quality (c40) compression, which makes it suitable for training and for compression studies. Celeb-DF [37] contains 590 real and 5,639 fake videos with improved visual quality. DFDC [39] is a large set of videos from paid actors, whose preview release has 1,131 real and 4,113 fake videos [5]. Diffusion-generated faces are covered by the DFF set (30,000 real and 90,000 fake images, as summarised in [5]), and general synthetic images by CIFAKE [29] (60,000 real and 60,000 synthetic images). Cross-GAN image generalisation can be studied with the protocol of [26].

TABLE IV CANDIDATE DATASETS AND THEIR PLANNED ROLES

Dataset	Content	Real / fake (as reported)	Generation methods	Planned role
FaceForensics++ [10]	Face video	1,000 / 4,000 (1,000 per method)	DeepFakes, Face2Face, FaceSwap, NeuralTextures	Training; in-domain test; compression study
Celeb-DF [37]	Face video	590 / 5,639	Improved face-swap synthesis	Zero-shot cross-dataset test
DFDC [39]	Face video	Preview: 1,131 / 4,113 [5]	Several synthesis methods	Large-scale cross-dataset test
DFF [30]	Face image	30,000 / 90,000 [5]	Diffusion-based generation	Diffusion generalisation test
CIFAKE [29]	General image	60,000 / 60,000	Latent-diffusion synthesis	Image-mode generalisation test
GAN image set [26]	General image	Per protocol	ProGAN-trained; GANs at test	Cross-GAN generalisation

B. Training, Validation, and Testing Strategy

Training uses FaceForensics++ videos split at identity level into training, validation, and test partitions (approximately 70/15/15). Validation data is used for early stopping, threshold selection, and calibration. Four protocols are planned (Fig. 4): P1, in-domain testing; P2, zero-shot transfer to Celeb-DF and DFDC; P3, leave-one-manipulation-out on FaceForensics++ and testing on diffusion image sets in image mode; and P4, robustness testing under the perturbations of Table II. Image-mode models are trained on image datasets separately.

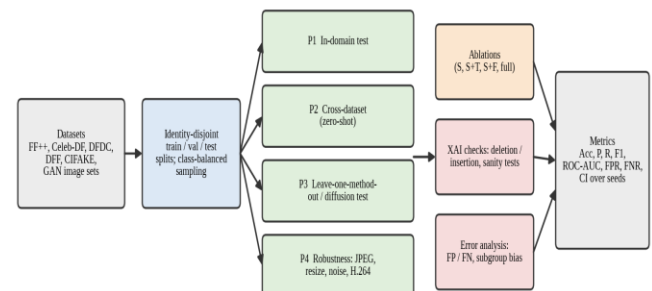


Fig. 4. Proposed experimental evaluation framework: datasets and splits, evaluation protocols P1 to P4, ablation, XAI, and error analysis, feeding the reported metrics.

C. Imbalance, Preprocessing, and Augmentation

Class imbalance (for example in Celeb-DF) is handled through class-weighted loss and balanced mini-batch sampling. Because accuracy can be misleading under imbalance, precision, recall, F1, balanced accuracy, and ROC-AUC are reported together. Preprocessing and augmentation follow Section IV-C.

D. Ablation and Evaluation Protocol

The ablation variants of Table V isolate the contribution of each stream and of the fusion mechanism. Each configuration is trained with at least three random seeds and reported as mean $\pm$ standard deviation with bootstrap confidence intervals, at both frame and video level. All baselines (Xception only, CNN-LSTM, frequency only) are re-trained and tested under identical splits and perturbations so that comparisons are fair.

TABLE V PLANNED ABLATION VARIANTS

ID	Configuration	Question addressed
V1	Spatial stream only	Baseline Xception performance
V2	Spatial + temporal	Value of temporal evidence
V3	Spatial + frequency	Value of spectral/statistical cues
V4	All streams, concatenation	Fusion without gating
V5	All streams, gated fusion	Full proposed model
V6	V5 without compression augmentation	Effect of robustness training
V7	V5 with Transformer temporal encoder	Bi-LSTM versus attention encoder

VII. PERFORMANCE EVALUATION

Fake is the positive class. Let TP, TN, FP, and FN denote true positives, true negatives, false positives (authentic

content flagged as fake), and false negatives (fakes accepted as real). The confusion matrix over these four counts underlies all metrics.

Accuracy is the share of correct decisions:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (10)$$

Precision measures how many fake flags are correct, and recall (true positive rate) how many fakes are found:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (11)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (12)$$

The F1-score is their harmonic mean:

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (13)$$

The false positive rate and false negative rate expose the two error types separately:

$$\text{FPR} = \frac{FP}{FP+TN} \quad (14)$$

$$\text{FNR} = \frac{FN}{FN+TP} = 1 - \text{Recall} \quad (15)$$

ROC-AUC summarises ranking quality across all thresholds:

$$\text{AUC} = \int_0^1 \text{TPR} d(\text{FPR}) \quad (16)$$

A false positive wrongly accuses authentic content, while a false negative lets a forgery pass; their costs differ by application, so the operating threshold is reported together with TPR at fixed low FPR values and the equal error rate. Calibration is assessed with expected calibration error. Under imbalance, balanced accuracy, $(\text{TPR} + \text{TNR})/2$, and precision-recall curves complement accuracy. No numerical values are presented here, because none were measured.

VIII. RESULTS AND DISCUSSION

A. Results Reported in the Literature

Table VI lists results reported by the cited studies. They were produced by those authors with different datasets, splits, and protocols and are not results of this work. They are not directly comparable and are shown only to illustrate the state of the field.

TABLE VI RESULTS REPORTED BY CITED STUDIES (NOT RESULTS OF THIS WORK)

Source	Method	Dataset	Reported result	Remark
[7]	Improved Xception	Deepfake_Faces (images)	93.37% validation accuracy	Illustrative confusion matrix has only ten samples
[6]	CNN-LSTM + Grad-CAM/SHAP	FF++/DFDC/Celeb-DF material; 100 test videos	Accuracy 0.75 in tables; 95% in text	Internally inconsistent; small test set
[11]	ViT + statistical (HDLST)	30,000-image set	Precision 0.98, recall 0.97, accuracy 0.96	ROC value (0.96) below VGG baseline (0.98) despite stated improvement
[13]	Detector module (unspecified)	FaceForensics++, Celeb-DF	Accuracy 95.2%; FPR 2.1%; FNR 2.7%	Architecture not described
[5]	One-class VAE (OC-FakeDect)	FF++ NeuralTextures subset	97.5% accuracy	Trained on real faces only
[5]	CNN on CIFAKE	CIFAKE	92.98% accuracy	After hyperparameter tuning
[5]	NA-VGG	Celeb-DF	AUC 85.7%	Reported as AUC, not accuracy
[5]	Human vs. model perception	Fake2M benchmark	Human error 38.7%; best model 13%	Shows difficulty for humans

Three observations follow. First, in-dataset accuracies are generally high (above 90% in several studies), whereas results on harder or cross-source data are lower (for example 85.7% AUC on Celeb-DF for NA-VGG, as summarised in [5]), consistent with the generalisation gap in Section III. Second, human observers perform poorly on modern synthetic images [5], which supports automated detection. Third, several reports have very small test sets or internally inconsistent numbers [6], [7], [11], so we treat them as indicative only and do not build performance claims on them. This also motivates the stricter reporting protocol of Section VI.

B. Expected Outcomes of the Proposed Framework

The following are hypotheses to be tested, not findings. Each could fail, and negative results (for example a frequency stream that does not help on diffusion images) would be informative.

- H1. Adding the frequency stream and gated fusion improves cross-dataset AUC over the spatial-only baseline (V5 versus V1), especially on GAN-based images.
- H2. The temporal stream helps most on video manipulations with frame-to-frame instability, and adds little on single images.

- H3. Compression augmentation reduces the performance drop under JPEG and H.264 (V5 versus V6), and the gate assigns lower weight to the frequency stream as compression increases.

- H4. Grad-CAM maps concentrate on blending boundaries, eyes, and mouth for face-swap fakes, and pass deletion/insertion tests better than random maps.

C. Expected Limitations and Generalisation

The framework is expected to remain vulnerable to generators that leave minimal artifacts and to adversarial perturbations [11], [42]. Spectral cues may not transfer from GANs to diffusion models [27]. Face-centric processing limits applicability to non-face or full-scene forgeries covered in [5]. Performance on demographic subgroups must be examined explicitly [38].

D. Computational Requirements

Cost is dominated by the Xception backbone (about 20.9 million parameters [7]) applied to T frames per clip. The frequency stream adds one FFT and a small MLP per frame, and the Bi-LSTM and gate add a small number of further parameters. Cost can be reduced by sampling fewer frames, batching, half-precision inference, or substituting a lighter

backbone, but this has not been measured. Training is expected to need a single modern GPU, and inference on short clips to be feasible near real time on a GPU; both statements require empirical confirmation.

E. Explainability

The multi-view explanation design offers regional (Grad-CAM), temporal (attention), and feature-level (SHAP, gate) evidence. Their value will be judged by the quantitative checks in Section IV-I rather than by visual plausibility alone.

IX. CHALLENGES AND LIMITATIONS

Evolving generators. New generators erase known artifacts, so detectors need continual re-evaluation [5], [27]. **Adversarial attacks.** Detectors can be evaded by small crafted perturbations [42]. **Dataset bias and domain shift.** Benchmarks differ in subjects, scenes, and compression, and models can learn dataset shortcuts; fairness across demographic groups is uneven [38]. **Computational cost and real-time use.** Multi-stream models are costlier than single-frame CNNs [5], [6]. **High-quality deepfakes.** Near-artifact-free outputs reduce the signal available [11]. **Video compression.** Codec artifacts mask or mimic forgery traces [10]. **Privacy.** Analysing faces involves biometric data, so storage, consent, and access control must be addressed in deployment. **Explainability limits.** Saliency maps may not faithfully reflect model reasoning [34], and an explanation is not proof that content is fake. Finally, this paper reports no experiments, so every performance-related statement about the framework remains a hypothesis.

X. FUTURE WORK

Planned extensions are: implementing and benchmarking the framework under the protocol of Section VI; multimodal detection that adds audio-video synchronisation analysis; Vision Transformer and multimodal Transformer backbones [24], [25]; detectors targeting diffusion-generated content using pretrained features [28]; real-time and edge deployment through distillation, pruning, and quantisation; continual learning to absorb new generators without forgetting; federated learning to train across institutions without sharing sensitive media; adversarial training for robustness [42]; and improved, quantitatively validated explanations [34].

XI. CONCLUSION

Deepfakes produced by GANs, VAEs, and diffusion models are increasingly hard to distinguish from authentic media, and current detectors are limited by poor generalisation,

sensitivity to compression, and opaque decisions. Drawing on five recent studies and foundational work, this paper identified these gaps and proposed an explainable multi-stream framework that integrates a transfer-learned Xception spatial stream, a Bi-LSTM temporal stream with attention, and a frequency/statistical stream through gated fusion, with Grad-CAM, attention, and SHAP evidence. A cross-dataset, cross-generator, and robustness-oriented protocol was specified to test it fairly. The contribution is a design and an evaluation blueprint; no accuracy is claimed because no experiments were run. The next step is implementation and empirical validation, after which the framework may support more transparent and trustworthy media forensics.

REFERENCES

- 1) I. J. Goodfellow et al., "Generative adversarial nets," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 27, 2014.
- 2) D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *arXiv:1312.6114*, 2013.
- 3) J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 6840–6851.
- 4) R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "DeepFakes and beyond: A survey of face manipulation and fake detection," *Inf. Fusion*, vol. 64, pp. 131–148, 2020.
- 5) M. Abdelmaksoud, B. Youssef, K. Wassif, and R. A. El-Khoribi, "A review on image forgery detection using deep learning techniques," in *Proc. 12th Int. Conf. Intell. Comput. Inf. Syst. (ICICIS)*, 2025, pp. 178–185, doi: 10.1109/ICICIS66182.2025.11313120.
- 6) S. Nandhini, R. Yasir Abdullah, R. Rathika, P. Dinakaran, T. Dharani Sree, and K. Magesh, "Explainable AI based framework for accurate deepfake video detection system," in *Proc. 4th Int. Conf. Artif. Intell. Mach. Learn. Appl. (AIMLA)*, 2026, doi: 10.1109/AIMLA67915.2026.11522405.
- 7) A. Khatri and N. Gupta, "Detection of AI-generated fake images using Xception model," in *Proc. 7th Int. Conf. Inf. Syst. Comput. Netw. (ISCON)*, Mathura, India, Sep. 2025, pp. 1–5, doi: 10.1109/ISCON65210.2025.11341496.
- 8) L. Verdoliva, "Media forensics and DeepFakes: An overview," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 5, pp. 910–932, Aug. 2020.

- 9) D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS), 2018, pp. 1–7.
- 10) A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2019, pp. 1–11.
- 11) U. F. Zaidi, R. Kumar, and M. I. Khan, "AI model for deepfake detection – A hybrid deep learning and statistical approach for GAN-generated visuals," in Proc. Int. Conf. Eng. Innov. Technol. (ICoEIT), 2025, pp. 131–137, doi: 10.1109/ICoEIT63558.2025.11211537.
- 12) M. S. Momin, A. H. Sufian, D. Barman, M. Leo, C. Distant, and N. Damer, "Explainable deepfake detection: Methods and challenges across multiple modalities," *Image Vis. Comput.*, vol. 163, p. 105738, 2025.
- 13) J. M. S. V. Ravi Kumar, D. V. Babu, B. Dheeraj, B. A. Kiran, G. Gowtham, and M. Srikanth, "Image search engine with recognition," in Proc. 4th Int. Conf. Sentiment Anal. Deep Learn. (ICSADL), 2025, pp. 1217–1223, doi: 10.1109/ICSADL65848.2025.10933116.
- 14) F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 1251–1258.
- 15) D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in Proc. IEEE Int. Conf. Adv. Video Signal-Based Surveill. (AVSS), 2018, pp. 1–6.
- 16) S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
- 17) G. Lee and M. Kim, "Deepfake detection using the rate of change between frames based on computer vision," *Sensors*, vol. 21, no. 21, p. 7367, 2021.
- 18) N. K. Sagar and S. Arukonda, "A new CNN-LSTM technique for effective deepfake detection," *Procedia Comput. Sci.*, vol. 258, pp. 1844–1855, 2025.
- 19) R. Durall, M. Keuper, and J. Keuper, "Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 7890–7899.
- 20) J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging frequency analysis for deep fake image recognition," in Proc. Int. Conf. Mach. Learn. (ICML), 2020, pp. 3247–3258.
- 21) Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2020, pp. 86–103.
- 22) O. Giudice, L. Guarnera, and S. Battiato, "Fighting deepfakes by detecting GAN DCT anomalies," *J. Imaging*, vol. 7, no. 8, p. 128, 2021, doi: 10.3390/jimaging7080128.
- 23) S. Sadhya and X. Qi, "Deepfake detection with wavelet-integrated convolutional networks," in Proc. IEEE Int. Conf. Big Data, Dec. 2024, pp. 4540–4547, doi: 10.1109/bigdata62323.2024.10825484.
- 24) A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learn. Represent. (ICLR), 2021.
- 25) J. Wang, Z. Wu, W. Ouyang, X. Han, J. Chen, Y.-G. Jiang, and S.-N. Li, "M2TR: Multi-modal multi-scale transformers for deepfake detection," in Proc. ACM Int. Conf. Multimedia Retrieval (ICMR), 2022, pp. 615–623.
- 26) S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot... for now," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 8695–8704.
- 27) R. Corvi et al., "On the detection of synthetic images generated by diffusion models," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), 2023, pp. 1–5.
- 28) U. Ojha, Y. Li, and Y. J. Lee, "Towards universal fake image detectors that generalize across generative models," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023, pp. 24480–24489.
- 29) J. J. Bird and A. Lotfi, "CIFAKE: Image classification and explainable identification of AI-generated synthetic images," *IEEE Access*, vol. 12, pp. 15642–15650, 2024.
- 30) H. Song, S. Huang, Y. Dong, and W.-W. Tu, "Robustness and generalizability of deepfake detection: A study with diffusion models," *arXiv preprint*, 2023.
- 31) R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based

- localization," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 618–626.
- 32) S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017.
- 33) N. Mansoor and A. I. Iliev, "Explainable AI for DeepFake detection," Appl. Sci., vol. 15, no. 2, p. 725, Jan. 2025, doi: 10.3390/app15020725.
- 34) J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 31, 2018.
- 35) G. Chen, C. Du, Y. Yuan, H. Hu, H. Duan, and H. Zhu, "A deepfake image detection method based on a multi-graph attention network," Electronics, vol. 14, no. 3, p. 482, Jan. 2025, doi: 10.3390/electronics14030482.
- 36) H. S. Alhaji, Y. Celik, and S. Goel, "ACO-PSO feature based deepfake video detection via deep learning," Electronics, vol. 13, no. 12, p. 2398, 2024.
- 37) Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for DeepFake forensics," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 3207–3216.
- 38) L. Trinh and Y. Liu, "An examination of fairness of AI models for deepfake detection," in Proc. 30th Int. Joint Conf. Artif. Intell. (IJCAI), 2021, pp. 567–574.
- 39) B. Dolhansky et al., "The DeepFake Detection Challenge (DFDC) dataset," arXiv:2006.07397, 2020.
- 40) K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," IEEE Signal Process. Lett., vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
- 41) J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2009, pp. 248–255.
- 42) N. Carlini and H. Farid, "Evading deepfake-image detectors with white- and black-box attacks," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 2020, pp. 658–659.