

A Lightweight Quality-Aware and Explainable Deepfake Detection Framework for Video Verification

Prof. Anmol S. Budhewar¹, Ashish K. Shrivastva², Amruta S. Patil³, Kartik B. Kawale⁴,
Bhumika P. Jadhav⁵

¹Assistant Professor, Dept. of Computer Engineering, Sandip Institute of Technology and Research
Centre, Nashik

^{2,3,4,5}B. Tech, Dept. of Computer Engineering, Sandip Institute of Technology and Research Centre,
Nashik

Abstract—Deepfake videos threaten the reliability of visual evidence since modern face manipulation can co-exist with ordinary degradations such as blur, poor illumination, compression, occlusion, small facial regions and repeated frames. In this paper, we propose a framework-oriented design for lightweight and explainable video verification instead of reporting the unmeasured detection results. The framework integrates IFQA, adaptive key-frame selection, landmark-based face alignment, attention-enhanced EfficientNetV2-B0 classifier, quality-weighted video-level confidence aggregation, and Grad-CAM visualisation. IFQA explicitly scores candidate frames on sharpness, illumination, contrast, face visibility, face-area ratio, and compression quality, and discards low-quality and near-duplicate frames before expensive classification. The remaining facial crops are classified independently and fused with normalised quality weights, without the complexity of a recurrent temporal model. The paper formalises the working stages, equations, implementation workflow, leakage-resistant dataset protocol, planned ablation and robustness evaluation, computational analysis and web-based verification layer. The resulting design targets a clear trade-off between detection reliability, computational efficiency, interpretability and deployment simplicity.

Index Terms—deepfake detection, video verification, frame quality assessment, EfficientNetV2, explainable artificial intelligence

I. INTRODUCTION

Deepfakes are fake photos and videos of faces generated and manipulated using state-of-the-art deep learning. With the growing visual realism of online media and the loss of their evidentiary value, they also

pose dangers for misinformation, impersonation, fraud and digital forensics [1]– [3]. A practical detector should be able to cope with video capturing, transmission and recompression conditions, beside manipulation artefacts.

Video verification is different from still-image categorisation as a single clip can have hundreds of near-identical frames, momentary motion blur, low light, occlusion, position changes and small faces. A naive approach that classifies all sampled frames spends computation on redundant evidence and gives ambiguous frames as much weight as clear frames. Public benchmarks such as FaceForensics++ [4], DFDC [5], Celeb-DF [6], DeeperForensics-1.0 [7], and WildDeepfake [8] have demonstrated the large variation in manipulation types, compression levels, capture conditions, and domain shift between datasets. Robust spatial baselines [9]– [16] Compact CNNs [9]– [16] Temporal aggregation [9]– [16] Attention mechanisms [9]–[16] Artifact-based cues [9]–[16] Generalization-oriented approaches [9]–[16] However, the practical steps of evidence selection, face standardisation, lightweight classification, video-level fusion, visual explanation and deployment are mostly studied independently. In this research we fill the integration gap by developing a quality-aware framework, whose main contribution is the management and weighting of the video evidence, both before and after the classification.

The proposed contributions are: (1) a composite IFQA score ranking candidate facial frames based on measurable quality indicators; (2) an adaptive rejection of low-quality and near-duplicate frames while preserving temporal coverage; (3) an

EfficientNetV2-B0 spatial classifier [17] enhanced with compact channel attention; (4) quality-weighted aggregation of frame probabilities to a video-level decision; and (5) Grad-CAM-based [19] explanations and a reporting-oriented verification workflow. Provisional accuracy or runtime values are not part of the experimental results. The work is intentionally created as a methodology and validation design.

II. RELATED WORK AND RESEARCH GAP

Deepfake detection has moved from handcrafted forensic traces to learned spatial and temporal representations. MesoNet [9] was introduced for lightweight facial forgery detection and Xception [10] became a common convolutional baseline in manipulation benchmarks. Recurrent video methods aggregate per-frame spatial features over time [11], [12], but sequential modelling is subject to inference complexity and can be sensitive to the quality and ordering of frames. Attention-based manipulation detection [13] shows that focusing on informative facial regions can improve both discrimination and localization.

Generalization remains a central challenge. Face X-Ray [14] targets blending boundaries that are less specific to a single manipulation generator, and Wang et al. [15] studied features that transfer across unseen CNN generators.

Lip-based temporal cues [16] provide another generalization direction, but such cues may be weak or unavailable in short, silent, or profile-view videos. The proposed framework therefore keeps the classifier spatial and shifts the temporal decision problem toward quality-controlled evidence selection and weighted fusion.

Complementary video-forensics cues include optical-flow inconsistencies [20], abnormal eye-blinking patterns [21], capsule-based representations [22], and localized visual artifacts around manipulated facial regions [23].

These directions demonstrate that deepfake evidence can be encoded in motion, physiology, learned part relationships, and appearance artifacts. They also motivate a framework in which the input evidence is explicitly filtered for visibility and quality before any cue is trusted.

TABLE I REPRESENTATIVE DEEPPFAKE DETECTION DIRECTIONS

Direction	Representative basis	Typical limitation
Compact CNN	MesoNet [9]	Limited explicit frame-quality and video-level fusion.
Spatial CNN	Xception [10]	Can learn dataset- or compression-specific traces.
Temporal model	CNN-RNN [11], [12]	Higher sequential complexity; sensitive to weak frames.
Attention	Attention-based detection [13]	Attention alone does not control redundant video evidence.
Generalization	Face X-Ray [14]; CNN cues [15]	Transfer can still degrade under domain shift.
Proposed	IFQA + attention + weighted fusion	Requires validation-tuned quality weights and thresholds.

The research gap is therefore not the absence of another backbone network. It is the lack of a compact end-to-end verification workflow that explicitly measures frame utility, rejects redundancy, standardizes faces, learns manipulation-sensitive spatial features, weights evidence according to its quality, and exposes the basis of the final decision. The framework in Sections III–VI is organized around this gap.

III. PROPOSED FRAMEWORK ARCHITECTURE

Fig. 1 presents the complete architecture. Image input can be forwarded directly to face preprocessing, whereas video input is sampled into candidate frames with associated metadata. The IFQA module assigns a normalized utility score to each valid facial frame. A redundancy filter then retains temporally distributed key frames.

RetinaFace [18] is used for face localization and landmark-based alignment, after which the crop is classified by an attention-enhanced EfficientNetV2-B0. Frame probabilities are fused by quality weights, and Grad-CAM maps are generated for representative frames.

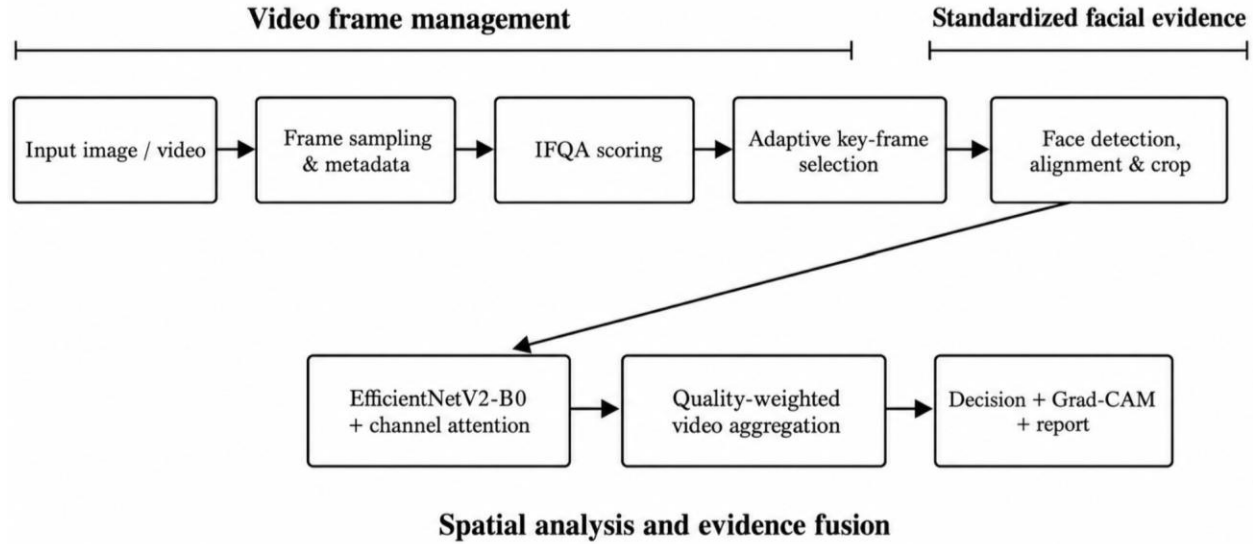


Fig. 1. End-to-end quality-aware and explainable deepfake video verification architecture.

The architecture separates three concerns. First, video frame management reduces low-value evidence before classification. Second, spatial analysis produces manipulation probabilities from standardized facial crops. Third, decision and explanation combine the accepted evidence into a video score and expose representative regions that influenced prediction. This separation allows each component to be validated independently through ablation while preserving a deployable end-to-end flow.

IV. DETAILED METHODOLOGY AND MATHEMATICAL FORMULATION

A. Stage 1: Dataset Preparation and Leakage Control

Public datasets are organized at the source-video and identity levels before any frame extraction. Training, validation, and test partitions are created so that all frames derived from the same source video remain in a single partition; identity-disjoint partitioning is preferred when identity metadata permits. Labels and media readability should be verified before preprocessing. This design prevents nearly identical frames from leaking across splits and inflating evaluation estimates.

B. Stage 2: Video Decoding, Face Detection, and Candidate Sampling

For a given input video V , the decoder logs duration, frame rate, resolution and total number of frames. Instead of processing all frames, candidate

frames are sampled at a configurable rate. RetinaFace [18] provides a face bounding box and five-point landmarks for each candidate. Frames without a valid target face are excluded. The remaining face box is used to compute the face-area ratio and an initial detector confidence.

C. Stage 3: Intelligent Frame Quality Assessment

For candidate frame F_i , IFQA combines six normalized indicators: sharpness s_i , illumination l_i , contrast c_i , facial visibility f_i , face-area ratio a_i , and compression-quality proxy r_i . Each indicator is scaled to $[0,1]$ so that higher values denote more usable facial evidence.

The composite score is

$$q_i = w_s \hat{s}_i + w_l \hat{l}_i + w_c \hat{c}_i + w_f \hat{f}_i + w_a \hat{a}_i + w_r \hat{r}_i \quad (1)$$

where the non-negative weights are fixed using validation data and satisfy

$$0 \leq w_k \leq 1, \quad \sum_k w_k = 1. \quad (2)$$

Sharpness can be estimated from the variance of the Laplacian over the face crop; illumination from the distance of mean intensity from an acceptable range; contrast from intensity standard deviation; visibility from detector confidence and landmark validity; face size from the facial area divided by frame area; and compression from a no-reference quality or blockiness proxy. A practical no-reference alternative is BRISQUE, which models spatial natural-scene

statistics for distortion-generic quality assessment [24]. A frame is quality-eligible when $q_i \geq \tau_q$, where τ_q is selected on validation data.

TABLE II IFQA COMPONENTS AND OPERATIONAL MEASUREMENTS

Component	Operational measurement	Purpose
Sharpness	Variance of Laplacian	Reject defocus/motion blur.
Illumination	Mean-intensity validity	Reject dark/overexposed faces.
Contrast	Face-crop intensity deviation	Estimate usable facial detail.
Visibility	Detector confidence + landmarks	Reject partial or missing faces.
Face size	Face area / frame area	Reject very small facial evidence.
Compression	Blockiness/no-reference proxy	Penalize severe coding artifacts.

D. Stage 4: Adaptive Key-Frame Selection

Quality-eligible frames are compared with the most recently accepted key frame. Let $\phi(F_i)$ be a perceptual embedding of the aligned face. Redundancy is measured by cosine distance

$$d_{ij} = 1 - [\phi(F_i)^T \phi(F_j)] / [||\phi(F_i)||_2 ||\phi(F_j)||_2]. \quad (3)$$

A frame is retained when $d_{ij} \geq \tau_d$, subject to a maximum temporal gap that prevents long segments from being represented by a single frame. The output is a small set of high-quality, non-redundant, temporally distributed key frames. Fig. 2 shows the quality-first selection sequence.

E. Stage 5: Facial Standardization and Training Augmentation

Accepted faces are aligned using facial landmarks, cropped with a small context margin, resized to the network input, converted to RGB, and normalized using the statistics required by the pretrained backbone.

Training-only augmentation may include horizontal flipping, mild rotation, brightness/contrast perturbation, Gaussian blur, JPEG recompression, and limited random erasing.

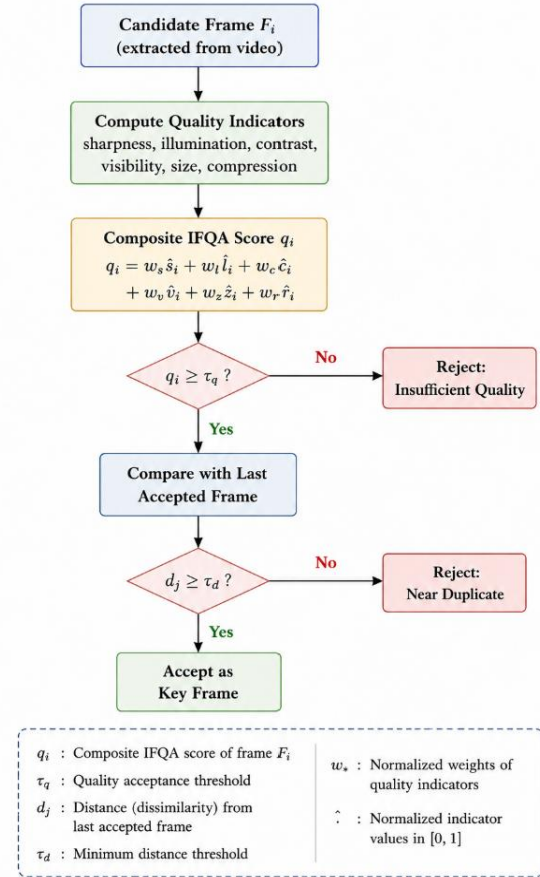


Fig. 2. IFQA and adaptive key-frame selection procedure.

The purpose is to reduce sensitivity to benign capture and transmission variations without creating label-changing artifacts.

F. Stage 6: Attention-Enhanced EfficientNetV2-B0

EfficientNetV2-B0 is selected as the lightweight spatial backbone because EfficientNetV2 was designed for improved training speed and parameter efficiency [17].

Let the final convolutional feature tensor be $X \in \mathbb{R}^{H \times W \times C}$. A compact channel-attention block first performs global average pooling

$$z_c = (1/HW) \sum_u \sum_v X_{\{u,v,c\}}, \quad (4)$$

and produces channel coefficients

$$a = \sigma(W_2 \delta(W_1 z)). \quad (5)$$

Here $\delta(\cdot)$ is ReLU and $\sigma(\cdot)$ is sigmoid. The attended tensor is $\tilde{X}_{\{u,v,c\}} = a_c X_{\{u,v,c\}}$. This squeeze-and-recalibration design is consistent with lightweight channel/spatial attention mechanisms such as

Squeeze-and-Excitation and CBAM [25], [26]. A binary classification head outputs frame-level deepfake probability $p_i \in [0,1]$. The attention block is deliberately small so that the framework's efficiency is not dominated by an auxiliary module.

G. Stage 7: Quality-Weighted Video Decision and Explainability

For K selected frames, the video-level deepfake confidence is the normalized quality-weighted mean $P_V = [\sum_{i=1}^K q_i p_i] / [\sum_{i=1}^K q_i + \epsilon]$,

(6)

where ϵ prevents division by zero. The final binary decision is

$$\hat{y} = I(P_V \geq \tau_c). \quad (7)$$

The threshold τ_c is calibrated on validation data. For image input, the single-frame probability is used directly. Grad-CAM [19] is generated from the final convolutional stage for representative accepted frames to indicate facial regions associated with the class score. These heatmaps are treated as model explanations, not as causal or legally conclusive forensic evidence.

H. Stage 8: Edge-Case Handling and Final Output

The workflow is deterministic for weak inputs. Instead of a forced real/fake label, the output is “insufficient facial evidence” if no valid target face is found. If there is only one eligible frame, the video score is the frame probability.

Multi-face videos need a target-face rule defined or face-track separation before being aggregated. The last entry logs the decision, confidence, selected frame identifiers, quality scores, processing statistics, representative Grad-CAM maps, model version and threshold configuration.

I. Stage 9: Complete Video-Inference Algorithm

The system can be algorithmically implemented as a deterministic evidence pipeline rather than a monolithic black box. The main ordering constraint is that quality assessment must precede expensive categorization and that video aggregation must only use accepted frames. This maintains the interpretation of q_i as evidence quality weight, not a learned class score.

TABLE V COMPACT QUALITY-AWARE VIDEO INFERENCE ALGORITHM

Step	Operation
1	Read video V ; validate format and collect FPS, duration, resolution, and frame count.
2	Sample candidate frames at the configured temporal interval.
3	Detect the target face and facial landmarks; reject frames without valid facial evidence.
4	Compute normalized IFQA indicators and composite quality score q_i .
5	Reject frames with $q_i < \tau_q$.
6	Compare each eligible frame with the most recently accepted key frame using d_{ij} .
7	Reject near-duplicates with $d_{ij} < \tau_d$ while enforcing temporal coverage.
8	Align/crop accepted faces and predict frame-level fake probability p_i .
9	Compute P_V from quality-weighted aggregation and assign label using τ_c .
10	Generate representative Grad-CAM maps and export verification metadata/report.

The same pipeline handles still images by bypassing sampling, redundancy filtering, and multi-frame aggregation. For videos, the maximum number of retained key frames can be bounded to protect latency and memory, but the bound should be declared and validated because overly aggressive truncation can reduce temporal coverage.

V. IMPLEMENTATION AND PLANNED EVALUATION PROTOCOL

A. Dataset Roles and Training Configuration

FaceForensics++ [4] is proposed as the main controlled training benchmark as it contains multiple face manipulation methods and compression levels. Celeb-DF [6] has a more challenging cross-dataset test; DFDC [5] has diverse identities and recording conditions; DeeperForensics-1.0 [7] enables perturbation-oriented robustness evaluation; WildDeepfake [8] spans unconstrained web content. Exact subset sizes are only reported after acquisition and pre-processing; the framework paper does not assume counts that are not verified.

TABLE III DATASET ROLES IN THE PLANNED EVALUATION

Dataset	Role	Reason
FaceForensics++ [4]	Primary training / compression	Multiple manipulation methods and compression settings.
Celeb-DF [6]	Cross-dataset test	Higher-quality, visually challenging deepfakes.
DFDC [5]	Diverse evaluation subset	Varied identities and acquisition conditions.
DeeperForensics-1.0 [7]	Perturbation robustness	Designed with real-world degradations.
WildDeepfake [8]	In-the-wild test	Unconstrained videos and domain shift.

A reproducible implementation record should disclose input resolution, pretrained weights, optimizer, learning rate and schedule, batch size, epoch budget, augmentation, loss, class-balancing policy, early stopping, random seed, framework version, hardware, IFQA weights, and thresholds. A reference configuration may use 224×224 aligned RGB crops, ImageNet-pretrained EfficientNetV2-B0, AdamW, class-balanced binary cross-entropy, cosine learning-rate scheduling, and validation-based early stopping. These are implementation settings, not performance claims.

B. Planned Evaluation, Ablation, and Robustness

Evaluation is planned at frame and video levels using accuracy, precision, recall, specificity, F1-score, and ROC-AUC. Efficiency measures include candidate-frame count, selected-frame count, frame-reduction ratio, model inference time, and peak memory. All threshold selection and IFQA weight tuning are restricted to validation data; the test set is reserved for final evaluation.

TABLE IV PLANNED ABLATION AND VALIDATION MATRIX

Study	Comparison	Question answered
Attention	Backbone vs. attention variant	Does attention add discriminative value?
IFQA	With vs. without quality rejection	Does low-quality rejection improve evidence reliability?
Redundancy	With vs. without similarity filter	How much computation is removed?

Aggregation	Mean vs. quality-weighted score	Does quality weighting improve video fusion?
Robustness	Original vs. degraded inputs	Stability under blur, light change, resize, recompression.
Cross-dataset	In-domain vs. unseen dataset	Generalization under domain shift.

Robustness testing should include recompression, resizing, Gaussian or motion blur, low illumination, occlusion, adversarial perturbation, and previously unseen manipulation conditions. Cross-dataset testing is especially important because a detector may otherwise exploit dataset-specific camera, synthesis, or compression traces [14], [15]. Adversarial perturbations have also been shown to degrade deepfake detectors [27], while explicit GAN-forensics studies reinforce the importance of unseen-generator generalization [28]. When multiple training runs are feasible, mean and standard deviation should be reported across random seeds; video-level bootstrap intervals are preferable to frame-level resampling because frames from the same clip are not independent.

C. Planned Metrics and Statistical Reporting

For the positive class 'deepfake', threshold-dependent measures are computed from true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). Precision, recall, F1-score, and specificity complement ROC-AUC because a single accuracy value can conceal class imbalance or a poor operating threshold.

$$\text{Precision} = TP / (TP + FP) \quad (8)$$

$$\text{Recall} = TP / (TP + FN), \quad (9)$$

$$F1 = 2(\text{Precision} \cdot \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (10)$$

Frame-selection efficiency is reported as the reduction from N candidate frames to K accepted key frames, $R_f = 100(1 - K/N) \%$. (11)

All metrics should be reported with the exact split and decision threshold. Where training can be repeated, the manuscript should report mean $\pm$ standard deviation and a video-level bootstrap confidence interval for paired video predictions. Failure cases should be categorised by blur, illumination, occlusion, pose, compression, face-detection failure, and domain shift so that model errors are not reduced to a single aggregate score.

D. Reproducibility Record

TABLE VI REPRODUCIBILITY CHECKLIST FOR THE PLANNED STUDY

Item	Required disclosure
Data split	Video/identity-disjoint rules, counts, class balance, and random seed.
Sampling	Frame rate/interval, maximum retained frames, temporal-gap rule.
Face processing	Detector version, landmarks, alignment, crop margin, resize, normalization.
IFQA	Indicator definitions, normalization, weights, τ_q , and failure policy.
Selection	Embedding/similarity representation, τ_d , and temporal coverage constraint.
Training	Initialization, optimizer, schedule, loss, batch size, epochs, augmentation.
Evaluation	Frame/video metrics, τ_c policy, robustness perturbations, confidence intervals.
Compute	Hardware, software versions, precision mode, and runtime measurement method.

VI. COMPUTATIONAL ANALYSIS, DEPLOYMENT AND DISCUSSION

A. Computational Model

Let N be the number of sampled candidate frames, K be the number of frames after IFQA and redundancy filtering, C_q be the per-frame cost of quality/similarity processing and C_m be the cost of one classifier inference.

The naive sampled-frame classifier requires $O(NC_m)$, while the proposed pipeline is about $O(NC_q + K C_m)$. The efficiency gain should be obtained when $K \ll N$ and $C_q \ll C_m$. So, the selection stage is not free of charge. Classification must be profiled separately from face identification, alignment, quality scoring and duplicate comparison.

A helpful implementation report should include the time for decoding, face detection, IFQA, redundancy filter, classifier, Grad-CAM and report generation. This per-component measure prevents us from attributing all the runtime savings to the backbone, and indicates whether face localisation becomes the main bottleneck on high-definition video.

B. Web Verification and Reporting Layer

The deployment layer accepts an image or video, validates format, records non-sensitive media metadata, displays candidate and selected frames, and presents the frame-level probabilities and video-level confidence. The report includes a timestamp, model version, IFQA and decision thresholds, number of sampled/retained frames, confidence score, representative Grad-CAM maps, and a clear limitation statement. Retention of uploaded facial media should be minimized and access controlled. Related applied work on SafeSpace AI demonstrates the practical value of combining AI-based content analysis with an explicit web moderation and review workflow [29].

C. Module-Level Data Contract and Failure Controls

For implementation, each stage should expose an explicit input, output, and failure state. This is important because a deployable verification system should not silently replace missing evidence with default values. A frame rejected because of missing landmarks is different from a frame rejected because of low sharpness, and both are different from a model prediction near the decision threshold. The platform therefore stores stage-level status codes alongside the final result.

TABLE VII MODULE-LEVEL INPUT, OUTPUT, AND FAILURE CONTROL

Module	Primary output	Failure / fallback
Media validation	Validated image/video + metadata	Reject unsupported/corrupt media.
Face localization	Target-face box + landmarks	No face $\rightarrow$ insufficient evidence.
IFQA	q_i and component scores	Below $\tau_q \rightarrow$ quality rejection.
Redundancy filter	Temporally distributed key frames	Near duplicate $\rightarrow$ omit frame.
Classifier	Frame probability p_i	Invalid crop $\rightarrow$ exclude with log.
Aggregation	Video confidence P_V	No valid frames $\rightarrow$ no binary label.
XAI/report	Heatmaps + audit metadata	XAI failure does not alter class score.

A structured data contract also helps with reproducibility. For each selected frame, the stored record may contain the source timestamp, bounding box, detector confidence, IFQA components, composite q_i , similarity distance to the previous accepted frame, class probability p_i , and whether a Grad-CAM map was generated.

For the video: N , K , R_f , P_V , τ_q , τ_d , τ_c , model identifier and software version. These values are sufficient to reconstruct the reason of a frame acceptance and its contribution to the final score without the need of keeping the whole uploaded video forever.

D. Threshold Calibration and Evidence Policy

The three thresholds serve different purposes and should not be tuned jointly on the test set. The quality threshold τ_q controls minimum facial usability; the redundancy threshold τ_d controls diversity among accepted frames; and the classification threshold τ_c maps the aggregated confidence to an operational label. A validation-only search can first select IFQA weights and τ_q using a frame-utility objective, then select τ_d using a performance-versus-frame-count trade-off, and finally select τ_c from the validation ROC or precision-recall curve according to the intended use. This staged calibration prevents the danger of a threshold compensating for a badly tuned earlier step.

Also, the interface should make a policy about evidence on the binary decision. A practical policy can distinguish between high confidence forecasts with many good quality key frames and predictions with only one marginal frame. The framework therefore proposes to report the aggregated class confidence along with the key-frame count, mean/median q_i , temporal coverage and an insufficient-evidence indication. This does not add a new classifier, it just stops showing a single probability value with no context.

Security & privacy are part of the deployment design. Validate media file type and size before decoding. . Use random names for temporary files. Logs should not contain unnecessary personally identifying information. Generate reports with model version and limitations. Where media is retained for research purposes, the terms of consent, access control and retention should be specified. These constraints do not improve the classification accuracy, but are necessary

for a responsible verification approach when using human facial data.

E. Discussion, Limitations, and Responsible Use

The core hypothesis of the framework is that frame quality can be treated as evidence quality to benefit video verification instead of treating all frames uniformly. The IFQA removes obvious weak observations, the redundancy filter discards repetitive evidence, the attention focuses on the manipulation-sensitive spatial channels and the quality-weighted aggregation propagates the ranking of evidence to the final video decision.

This organization also allows the framework to be testable as each component can be removed without affecting the remaining training/evaluation split. The broader work of Budhewar et al. highlights robustness, interpretability and computational-complexity trade-offs in multi-feature deep learning [30], while neighbour-aware decentralised learning underscores adaptive selection under resource constraints [31]. These considerations further justify a modular evidence-selection design.

There remain important limitations. A purely spatial classifier can miss temporal inconsistencies exploited by recurrent or multimodal systems [11], [12], [16]. IFQA thresholds may need to be adjusted if camera conditions or datasets change. Identity-aware tracking is necessary for aggregation of multi-face videos. Grad-CAM can indicate where a network is sensitive, but does not confer causal authenticity. Finally, high model confidence does not equate to provenance verification, metadata authentication, or professional forensic certification. Therefore, the platform should be used as a research and pre-verification tool.

F. Threats to Validity

Internal validity is compromised when the same source video or identity appears in training and testing, when thresholds are estimated on test data, or when there are different pre-processing and augmentation budgets than baselines.

Testing a detector on a single generator family or compression regime limits its external validity. The quality score relates to the labels in the dataset, not to generic visual usability, thus endangering construct validity. These issues result in identity/video-disjoint splits, validation-only calibration, matched baseline

budgets, cross-dataset testing, and explicit reporting of discarded frames.

The system should not operationally undermine low-evidence States. In cases of high compression or when only a small portion of the face is visible, forcing a binary label when no face is detected would not be a true representation of the evidence available. The interface should reflect a lack of evidence and preserve the difference between model confidence and media authenticity. The difference is especially significant in journalistic, editorial and investigative contexts where there may be a tendency to rely too much on the automated output.

VII. CONCLUSION AND FUTURE WORK

In this paper, we present a seven-stage research methodology for lightweight and explainable deepfake video verification. The proposed design consists of quality-aware frame management, adaptive key-frame selection, standardised face pre-processing, Attention-enhanced EfficientNetV2-B0 classifier, quality-weighted video aggregate, Grad-CAM explanation and reporting-oriented deployment layer. It describes the technique, equations, leakage-resistant data protocol, ablation plan, robustness tests, and computational model, but unlike a results article, it does not report preliminary engineering targets as measurable results.

Future work should validate the framework on identity-disjoint and cross-dataset splits, calibrate IFQA weights on varied capture situations and compare the proposed spatial fusion to lightweight temporal alternatives. Extensions may include face tracking for many persons in films, frequency domain or physiological cues, uncertainty calibration, and provenance-aware metadata analysis, while at the same time respecting the idea of explicit evidence quality in the framework.

REFERENCES

[1] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," *Information Fusion*, vol. 64, pp. 131–148, 2020.

[2] L. Verdoliva, "Media forensics and DeepFakes: An overview," *IEEE Journal of Selected Topics*

in Signal Processing, vol. 14, no. 5, pp. 910–932, 2020.

[3] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," *ACM Computing Surveys*, vol. 54, no. 1, Art. no. 7, 2021.

[4] 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. Computer Vision (ICCV)*, 2019, pp. 1–11.

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

[6] 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. Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3207–3216.

[7] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, "DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2889–2898.

[8] B. Zi, M. Chang, J. Chen, X. Ma, and Y.-G. Jiang, "WildDeepfake: A challenging real-world dataset for deepfake detection," in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 2382–2390.

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

[10] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1251–1258.

[11] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *Proc. IEEE Int. Conf. Advanced Video and Signal Based Surveillance (AVSS)*, 2018, pp. 1–6.

[12] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops*, 2019, pp. 80–87.

- [13] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain, "On the detection of digital face manipulation," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5781–5790.
- [14] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face X-Ray for more general face forgery detection," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5001–5010.
- [15] 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. Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8695–8704.
- [16] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5039–5049.
- [17] M. Tan and Q. V. Le, "EfficientNetV2: Smaller models and faster training," in Proc. 38th Int. Conf. Machine Learning (ICML), Proc. Machine Learning Research, vol. 139, 2021, pp. 10096–10106.
- [18] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-shot multi-level face localisation in the wild," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5203–5212.
- [19] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2017, pp. 618–626.
- [20] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, "DeepFake video detection through optical flow-based CNN," in Proc. IEEE/CVF Int. Conf. Computer Vision Workshops (ICCVW), 2019, pp. 1205–1207.
- [21] Y. Li, M.-C. Chang, and S. Lyu, "In Ictu Oculi: Exposing AI created fake videos by detecting eye blinking," in Proc. IEEE Int. Workshop Information Forensics and Security (WIFS), 2018, pp. 1–7.
- [22] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-Forensics: Using capsule networks to detect forged images and videos," in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 2307–2311.
- [23] F. Matern, C. Riess, and M. Stamminger, "Exploiting visual artifacts to expose deepfakes and face manipulations," in Proc. IEEE Winter Applications of Computer Vision Workshops (WACVW), 2019, pp. 83–92.
- [24] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," IEEE Transactions on Image Processing, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.
- [25] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7132–7141.
- [26] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in Proc. European Conf. Computer Vision (ECCV), 2018, pp. 3–19.
- [27] A. Gandhi and S. Jain, "Adversarial perturbations fool deepfake detectors," in Proc. Int. Joint Conf. Neural Networks (IJCNN), 2020, pp. 1–8.
- [28] X. Xuan, B. Peng, W. Wang, and J. Dong, "On the generalization of GAN image forensics," in Biometric Recognition (CCBR 2019), Lecture Notes in Computer Science, vol. 11818, 2019, pp. 134–141.
- [29] A. Budhewar, T. M. Agrawal, R. P. Shinde, R. R. Yadav, and R. K. Deore, "SafeSpace AI: Intelligent Content Moderation Platform," International Journal of Advanced Computer Engineering and Communication Technology, vol. 14, no. 1, pp. 658–662, 2025.
- [30] A. S. Budhewar, P. G. Patil, R. Bodkhe, S. Ghuge, D. Bhavar, and G. Shirsath, "A comprehensive survey on ensemble multi featured deep learning models: Applications, challenges, and future directions," Open Access International Journal of Science and Engineering, vol. 7, no. 7, pp. 23–30, 2024.
- [31] A. S. Budhewar, P. G. Patil, and S. M. Kale, "Neighbour-Aware Cooperation for Semi-Supervised Decentralized Machine Learning," Educational Administration: Theory and Practice, vol. 30, no. 5, pp. 2039–2047, 2024, doi: 10.53555/kuey.v30i5.3271.