Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An Integrated Spatio-Temporal Architecture for Deepfake Video Classification: An Evaluation on the FFIW-10K Multi-Person Benchmark

Sameena Javaid

sameenajavaid@mmu.edu.my

Multimedia University

Fazeel Zafar

Bahria University

Shahood Sajid

Bahria University

Safdar Rizvi

Bahria University

Muhammad Mansoor Alam

Multimedia University

Mazliham Mohd Su'ud

Multimedia University

Article

Keywords: deepfake video classification, FFIW-10K, EfficientNet-B0, bidirectional LSTM, temporal attention, Transformer encoder, multi-person face forgery

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11508304/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

An Integrated Spatio-Temporal Architecture for Deepfake Video Classification: An Evaluation on the FFIW-10K Multi-Person Benchmark

Sameena Javaid [1,*], Fazeel Zafar [2], Shahood Sajid [2], Safdar Rizvi [2,3], Muhammad Mansoor Alam [4], Mazliham Mohd Su'ud [1,*]

[1] Research Management Center, Multimedia University, 63100 Cyberjaya, Selangor, Malaysia.

[2] Department of Computer Sciences, School of Engineering and Applied Sciences, Bahria University, Karachi Campus, Karachi, Pakistan.

[3] Malaysian Institute of Information Technology (MIIT), University Kuala Lumpur, Kuala Lumpur, Malysia.

[4] Faculty of Computing, Riphah International University, Rawalpindi 46000, Islamabad.

* Corresponding Author: Sameena Javaid (sameenajavaid@mmu.edu.my)

and Mazliham Mohd Su'ud (mazliham@mmu.edu.my)

Abstract

Modern deepfake methods can create fake faces that are very difficult to detect just by watching a video. Therefore, practical deepfake-detection systems must be accurate while using limited computing resources. CNN-based methods analyse individual frames and can identify visual signs of manipulation, but they may miss unnatural changes between consecutive frames. Video Transformers can analyse these changes over time, but they usually require substantial computational power. A more practical approach is to combine CNNs for analysing facial details with recurrent and attention-based models for studying changes across frames. However, only a limited number of studies have tested such combined methods on realistic videos containing multiple people, where some faces may be manipulated while others remain genuine.

This paper describes and evaluates an integrated architecture for video-level deepfake classification. Spatial features are extracted by an EfficientNet-B0 backbone augmented with a multi-scale, attention-refined feature path; the resulting 1,280-dimensional frame embeddings are processed by a two-layer bidirectional long short-term memory (BiLSTM) network with an additive temporal attention mechanism that rescales each frame representation by a learned weight; and the weighted sequence is refined by a single-layer Transformer encoder with four attention heads before temporal averaging and binary classification. Training employs gradient accumulation, mixed-precision arithmetic, chunked frame processing, exponential moving averaging of parameters and differential learning rates, enabling 45-frame sequences to be trained on a single 12 GB consumer graphics processing unit.

The architecture was evaluated on a subset of the FFIW-10K benchmark using a held-out evaluation set of 1,987 instances, comprising 998 real (50.23 %) and 989 fake (49.77 %). It attained an accuracy

[OFFICIAL]

of 96.18 % (95 % Wilson score interval 95.24–96.93 %), precision, recall and F1-score of 0.96 for both classes, specificity of 96.19 %, a Matthews correlation coefficient of 0.9235, and an area under the receiver operating characteristic curve of 0.991. Errors were numerically balanced: 38 false positives (false-positive rate 3.81 %) and 38 false negatives (false-negative rate 3.84 %), giving an overall error rate of 3.82 %. The spatial backbone requires approximately 0.45 GFLOPs per frame, giving an estimated 20.25 GFLOPs per video for the spatial path at 45 frames.

The contribution is an integrated architecture together with a transparent single-configuration evaluation of it. The study reports one training configuration on one dataset, processes complete frames without face localisation, and does not include component-wise ablation, repeated-seed trials, matched baseline reproduction, cross-dataset validation or measured inference latency. The independent contribution of each architectural module was therefore not quantified, and the findings apply to the reported evaluation setting only.

Keywords: deepfake video classification; FFIW-10K; EfficientNet-B0; bidirectional LSTM; temporal attention; Transformer encoder; multi-person face forgery

1. Introduction

Facial manipulation methods based on generative modelling have reached a level of visual fidelity at which synthetic video is not reliably distinguishable by human observers [1], [2]. Such content circulates through social platforms, messaging systems and broadcast media, where it supports defamation, fraud, non-consensual imagery and disinformation. Automated detection has therefore become a component of media-authentication infrastructure rather than a purely academic problem, and its practical value depends jointly on discriminative capability and on operating within the computational budget of the systems that would host it.

Manipulation leaves traces at several semantic levels simultaneously. Generative up-sampling introduces characteristic high-frequency texture and noise-residual patterns; compositing a synthesised face into a source frame produces localised blending discontinuities at the mask boundary together with inconsistencies in skin tone and illumination; and identity substitution alters facial geometry and its consistency with head pose. A detector that reads only the terminal convolutional stage of a backbone consequently discards the representations in which the first two artefact classes are most legible. Because manipulation is applied frame by frame, temporal coherence is imposed weakly or not at all, and the observable consequences span two ranges. Locally, adjacent frames exhibit flicker in texture, illumination and mask boundary. Globally, identity and geometry drift across the sequence, and artefacts appear in bursts separated by many clean frames.

These characteristics map onto complementary weaknesses in the two dominant families of detector. Frame-level convolutional classifiers [3] are computationally economical and transfer readily from large-scale pretraining, but judge a video one frame at a time, so temporal incoherence is invisible to

[OFFICIAL]

them. Placing a recurrent network above a convolutional backbone [4] recovers local temporal information, and this pattern now dominates the deployability-oriented literature [5]–[8]; but the influence of a distant frame decays through the recurrence, and every sampled frame enters the sequence with equal weight, so frames carrying no manipulation evidence contribute as strongly as those that do. Self-attention removes the distance penalty by relating every position to every other directly [9], [10], and graph formulations achieve comparable capability at reduced parameter cost [11]; but full spatio-temporal attention scales quadratically in token count, and applying it to raw frame features expends capacity relating positions that carry no evidence. Frequency-domain approaches [12] exploit spectral signatures of generative up-sampling, but degrade under recompression, which redistributes spectral energy.

A further limitation cuts across all of these families: most published detectors are evaluated on single-subject benchmarks. The FFIW-10K benchmark [13] is constructed to reflect a harder and more realistic condition, containing 3.15 faces per frame on average ranging from one to fifteen with only a small proportion of the faces present manipulated in many videos. A detector operating on such footage must locate evidence occupying a small fraction of the frame, while the real faces in the same frame act as distractors. Progress in this setting is also impeded by a measurement problem. As Table 1 shows, six of the ten studies reviewed here report no computational-complexity figure at all, and those that do use incompatible accounting conventions, so accuracy–efficiency trade-offs across the literature cannot be compared directly.

The gap addressed in this paper is therefore the absence of a transparently reported evaluation, on a multi-person in-the-wild benchmark, of an architecture that combines multi-scale attention-refined spatial encoding with both local and global temporal modelling at a per-frame cost comparable to a lightweight convolutional backbone. The architecture evaluated here combines three mechanisms whose individual motivations are established in prior work: multi-scale feature aggregation, which preserves representations from several network depths [14]; attention-based feature refinement, which reweights representations according to learned relevance [15], [16]; and hierarchical temporal modelling, in which a recurrent stage supplies locally contextualised representations to a global self-attention stage [17], [10], [4]. The design intent is that these mechanisms be complementary. That intent was not tested experimentally in this study: no component-wise ablation was performed, and no claim is made about the independent contribution of any module. The architecture is presented as an integrated design whose components are theoretically motivated, and it is evaluated as a whole.

This study follows the authors' earlier work [18], which evaluated an EfficientNet-B0 backbone combined with a feature pyramid and a temporal convolutional network on FFIW-10K at an approximate per-frame cost of 0.45 GFLOPs, reporting 92.45 % accuracy. The two studies overlap in dataset, backbone family, complexity budget and authorship. The present architecture differs in three respects: attention-based refinement is applied to each scale before fusion, rather than fusing pyramid

[OFFICIAL]

features directly; temporal modelling is recurrent and bidirectional with an explicit per-frame attention weighting, rather than convolutional; and a global self-attention stage is added above the recurrent stage. The accuracy reported in [18] is used here only as a literature-based reference point and was not reproduced under the present protocol; the two evaluation protocols cannot be shown to be equivalent, because the split construction used here is not fully documented (Section 3.1.2). No text, figure, table or equation from [18] is reused.

The contributions of this paper are as follows.

1. An integrated deepfake video-classification architecture combining EfficientNet-B0 spatial feature extraction with a multi-scale attention-refined feature path, bidirectional recurrent temporal modelling with additive frame-level attention, and Transformer-based sequence refinement, specified with complete tensor dimensions throughout (Section 3.3, Figures 1 and 3).
2. An evaluation of the architecture on the FFIW-10K multi-person benchmark under a single training configuration and a single held-out evaluation protocol (Section 4.1).
3. A quantitative analysis of classification behaviour derived from the reported confusion matrix, receiver operating characteristic curve and per-class metrics, including specificity, per-class error rates, the Matthews correlation coefficient and an interval estimate for the observed accuracy proportion (Sections 4.2–4.3).
4. A treatment of computational cost that separates per-frame backbone complexity from estimated per-video processing cost, correcting a conflation present in the prior literature on this architecture family (Section 3.6).
5. An explicit statement of the study's evidential limits, including the absence of ablation, repeated-seed evaluation, matched baseline reproduction, cross-dataset validation and latency measurement (Section 5.2).

The remainder of the paper is organised as follows. Section 2 reviews related work thematically. Section 3 describes the dataset, data preparation, architecture, training configuration, computational-cost estimation and evaluation metrics. Section 4 reports the results. Section 5 interprets them, states the limitations and threats to validity, and discusses practical implications and future work. Section 6 concludes.

2. Related Work

Frame-level convolutional detectors. The most widely deployed approach classifies individual frames with a convolutional network, treating manipulation detection as a texture and noise-residual problem. Sharma and Selwal [3] exemplify the modern form, showing that preprocessing contrast-limited adaptive histogram equalisation, denoising and sharpening recovers accuracy on degraded inputs, reporting 92.5 % on CelebA images paired with images generated by a deep convolutional generative adversarial network. Detectors of this family are economical and transfer readily from

[OFFICIAL]

ImageNet pretraining, but they cannot observe temporal incoherence, and they read only the terminal convolutional stage, discarding earlier representations.

Recurrent spatio-temporal detectors. Placing a recurrent network above a convolutional backbone allows frame embeddings to be consumed as a sequence, an approach introduced for this task by Güera and Delp [4]. Four of the studies in Table 1 follow this pattern with lightweight backbones chosen for deployability. AlMuhaideb et al. [5] combine MobileNet-v1 with a convolutional block attention module and a gated recurrent unit, using multi-task cascaded convolutional network (MTCNN) face cropping [19] and structural-similarity frame selection, reporting 98.2 % on Celeb-DF v2 [20] and 95.0 % on the DeepFake Detection Challenge (DFDC) dataset [21]. Maheswari et al. [7] and Anandhasivam et al. [8] pair MobileNet variants with long short-term memory (LSTM) networks, reporting 96.3 % and 91.8 % respectively. Ghosh et al. [6] add a micro-expression branch to a ResNet-50 and BiLSTM combination, reporting 95.6 % on low-quality and 97.0 % on high-quality FaceForensics++ [22] inputs. This family recovers local temporal information, but the influence of distant frames decays through the recurrence, and all sampled frames are weighted equally.

Transformer and graph-based detectors. Self-attention relates every position to every other directly. Bastos et al. [9] combine a video vision Transformer with handcrafted texture, illumination and spectral descriptors, reporting 96.71 % on FaceForensics++; their ablation found that adding raw red–green–blue input *degraded* performance, indicating that constraining what an attention stage observes can improve rather than harm results. Liu et al. [11] represent the video as a spatial–spectral–temporal graph and apply graph attention, reporting competitive accuracy with up to 42.4× fewer parameters than the models compared against. The limiting factor for this family is computational: full spatio-temporal attention scales quadratically in token count, and the graph formulation of [11] is an explicit response to that cost.

Frequency-domain and handcrafted-feature approaches. A parallel line exploits the fact that generative up-sampling leaves signatures more separable in the frequency domain than in pixel space. Amen and Ranam [12] combine a fast Fourier transform stage with MobileNet and an attention layer for mobile deployment, reporting 94.2 % accuracy and 93.8 % F1-score with a 27.5 % improvement in computational efficiency over conventional convolutional approaches. The handcrafted descriptors used by [9] belong to the same tradition. Frequency cues are effective against the generators they were tuned on but degrade under recompression.

Efficiency-oriented detectors and the measurement problem. Deployability is a shared concern across [5], [6], [7], [8], [11], [12], expressed as mobile-class backbones, parameter reduction and edge inference. The obstacle to cumulative progress here is measurement rather than method. Six of the ten studies in Table 1 report no complexity figure, and those that do use incompatible conventions: the 63.25 GFLOPs recorded for a MobileNet-v1-based detector [5] cannot be a per-frame figure for that

[OFFICIAL]

backbone and is presumably a per-clip aggregate, whereas the 0.45 GFLOPs of [18] is a per-frame figure. The present paper addresses this for its own architecture in Section 3.6 by reporting both bases explicitly.

Related work in image forensics. Aryan et al. [23] propose a patch-based self-attention network for image forgery detection, evaluated on CASIA v1, DEFACTO, Columbia and MICC-F2000. This addresses splicing and copy–move localisation in still images a different task with a different artefact model and is included for methodological context only; its reported accuracies are not comparable with video deepfake results.

Position of the present study. The architecture evaluated here combines multi-scale attention-refined spatial encoding, recurrent temporal modelling with learned frame weighting, and global self-attention over the weighted sequence. Its closest antecedent is the authors' earlier work [18], which shares the dataset, backbone and complexity budget but uses direct pyramid fusion and temporal convolution in place of attention-refined fusion and recurrent modelling with attention.

Table 1. Studies referenced in this review. Accuracy figures are those reported by the cited authors on the datasets named in the same row. They are not comparable across rows, because datasets, subsets, splits, evaluation units and preprocessing differ, and they are not comparable with the result reported in this paper. Task codes: VDD, video deepfake detection; GAN, image-level generative-adversarial-network content detection; IFL, image forgery localisation.

Ref.	Year	Task	Dataset(s)	Architecture	Face pipeline	Reported accuracy (%)	Complexity	Pretraining
[3]	2025	GAN	CelebA (real) + DCGAN (fake)	MesoNet + preprocessing optimisation	CLAHE, denoising, sharpening	92.5 [b]	Not reported	MesoNet
[5]	2025	VDD	DFDC, Celeb-DF v2	MobileNet-v1 + CBAM + GRU	MTCNN crop, 224 × 224, SSIM frame selection	98.2 (Celeb-DF v2); 95.0 (DFDC)	63.25 GFLOPs [a]	MTCNN; ImageNet
[6] [d]	2025	VDD	FaceForensics++	ResNet-50 + BiLSTM + micro-expression CNN	Face crop, 224 × 224 and 64 × 64 ROIs	95.6 (LQ); 97.0 (HQ) [b]	7.82 GFLOPs [a]	ResNet-50 (ImageNet)
[7]	2025	VDD	FF++, Celeb-DF v2, DFDC	MobileNet + LSTM	Frame extraction, resize, normalise	96.3 [b]	Not reported	MobileNet (ImageNet)
[8]	2024	VDD	DFDC, DFD	MobileNet-v3 + LSTM	Frame extraction, resize, normalise	91.8 [b]	Not reported	MobileNet-v3 (ImageNet)
[9]	2025	VDD	FaceForensics++, Replay-Attack	ViViT + handcrafted features	Canny-based frame sampling	96.71 (FF++) [b]	Not reported	None
[11]	2026	VDD	FF++, Celeb-DF, DFDC, Wild-DF	Spatial–spectral–temporal GNN + GAT	Frame patching; no explicit face crop	95.6 (FF++) [b]	Not reported; 42.4× fewer parameters than compared models	ResNet-50 (partial)
[12]	2025	VDD	FF++, Celeb-DF, DFDC [b]	MobileNet + attention + FFT	Grayscale, FFT, normalisation	94.2 [c]	Not reported	MobileNet (ImageNet)

[OFFICIAL]

[18]	2025	VDD	FFIW-10K	EfficientNet-B0 + FPN + TempCNN	MTCNN detect and align	92.45 [c]	0.45 GFLOPs per frame	EfficientNet-B0 (ImageNet)
[23]	2025	IFL	CASIA v1, DEFACTO, Columbia, MICC-F2000	ViT (4 Transformer blocks)	Resize 224 × 224, 16 × 16 patches	93.98 / 99.14 / 82.65 / 96.06 [b]	Not reported	None
This study	—	VDD	FFIW-10K (subset)	EfficientNet-B0 + multi-scale attention path + BiLSTM with temporal attention + Transformer encoder	**Whole frames; no face detection (Section 3.2.1)**	**96.18**	0.45 per frame; ≈ 20.25 per video (spatial path, T = 45)	EfficientNet-B0 (ImageNet)

[a] Accounting basis not stated in the source. The 63.25 GFLOPs figure is inconsistent with a per-frame MobileNet-v1 cost and is presumed to be a per-clip aggregate.

[b] Not independently verified during preparation of this manuscript.

[c] Corrected during preparation of this manuscript; the source reports 94.2 % accuracy and 93.8 % F1-score.

[d] Unrefereed preprint.

[e] Reported by the cited authors on FFIW-10K. Used here as a literature-based reference point only; not reproduced under the present protocol (Sections 1 and 4.5).

Abbreviations used in Table 1: CBAM, convolutional block attention module; CLAHE, contrast-limited adaptive histogram equalisation; CNN, convolutional neural network; DCGAN, deep convolutional generative adversarial network; FF++, FaceForensics++; FFT, fast Fourier transform; FPN, feature pyramid network; GAT, graph attention network; GNN, graph neural network; GRU, gated recurrent unit; HQ, high quality; LQ, low quality; ROI, region of interest; SSIM, structural similarity index measure; TempCNN, temporal convolutional neural network; ViT, vision Transformer; ViViT, video vision Transformer.

3. Materials and Methods

3.1 Dataset

3.1.1 FFIW-10K

The experiments were conducted using the Face Forensics in the Wild 10K (FFIW-10K) dataset introduced by Zhou et al. [13], which contains 10,000 high-quality manipulated videos involving complex multi-person scenes. It was the sole dataset used in this study. The published benchmark comprises 20,000 videos in total the 10,000 manipulated videos paired with 10,000 real source videos with an official partition of 16,000 training, 500 validation and 3,500 test videos. Frames contain 3.15

[OFFICIAL]

faces on average, ranging from one to fifteen, and in many videos only a small proportion of the faces present are manipulated. The benchmark was selected because these multi-person, in-the-wild conditions are more representative of deployment than single-subject benchmarks, and because it is the benchmark used in the authors' earlier work [18].

A subset of approximately 3,000 videos was used rather than the full benchmark, for reasons of computational budget: the available hardware was a single 12 GB graphics processing unit (Table 2). This choice reduces comparability with results reported on the full benchmark and is recorded as a limitation in Section 5.2.

3.1.2 Partitioning and the held-out evaluation set

The training procedure used a 70 % / 15 % / 15 % training, validation and evaluation partition of the subset, with the stated intention of maintaining subject-level separation between partitions.

All reported metrics were computed on a held-out evaluation set of **1,987 instances**, comprising 998 real (50.23 %) and 989 fake (49.77 %) a class ratio of 1.009 : 1 and an imbalance of nine instances, or 0.45 percentage points. The set is therefore close to balanced, which is the basis on which accuracy is treated as an informative primary metric in Section 3.7.

Two properties of this evaluation set could not be established from available material and are disclosed here rather than resolved by assumption. First, its **unit** is not established: a 15 % partition of approximately 3,000 videos would contain approximately 450 videos, and 1,987 corresponds to no partition of a 3,000-video pool under the stated proportions, nor to either official FFIW-10K partition of 500 or 3,500 videos. The 1,987 instances may therefore be videos, temporal sequences, clips or frames, and they are referred to throughout as evaluation instances. Second, the partition from which they were drawn validation or test is not established, and the set is accordingly described as a held-out evaluation set rather than a test set. Both properties bear directly on the interpretation of the reported figures and are revisited in Section 5.2.

Because the partitioning key whether splitting was performed on source identity, source video or extracted sequence is not documented and disjointness was not verified, identity-level and sequence-level separation between partitions cannot be confirmed in this study. This is treated as a threat to internal validity in Section 5.2 and is not claimed as an achieved property.

3.2 Data preparation

3.2.1 Frame handling: whole frames without face localisation

The pipeline processes complete video frames. Each sampled frame is resized directly to 224 × 224 pixels, as indicated in Figure 1. No face detection, face cropping, face tracking, landmark alignment, bounding-box extraction or person-level selection is applied at any stage.

[OFFICIAL]

This is stated explicitly because it is consequential. As noted in Section 3.1.1, FFIW-10K frames contain 3.15 faces on average, and only a subset of those faces is manipulated in many videos. Under whole-frame processing at 224 × 224, each face occupies a small fraction of the input, and the model produces a single video-level decision without localising which face or which region carries the evidence. The architecture therefore performs video-level classification, not person-level forgery localisation, and no localisation capability is claimed. The associated risk that the classifier responds to scene-level rather than face-level cues is analysed in Section 5.2.

3.2.2 Frame sampling and sequence formation

Each video is represented as an ordered sequence of frames:

$$\mathcal{V} = \{S_1, S_2, \ldots, S_N\} \tag{1}$$

where N is the number of frames in the source video and S_i is a red–green–blue frame. A fixed number T of frames is sampled uniformly from each video, so that videos of differing duration produce tensors of uniform shape. The reported configuration uses $T = 45$ throughout, as reflected in the tensor dimensions of Figure 1. Small random offsets are applied to the sampled frame indices during training, so that the temporal modules are exposed to variation in inter-frame spacing.

3.2.3 Frame preprocessing

Each sampled frame is transformed as

$$S_i' = \text{Augment}\left(\text{Resize}\left(\text{Normalize}(S_i)\right)\right) \tag{2}$$

where Normalize(·) applies per-channel standardisation, Resize(·) maps to 224 × 224 pixels, and Augment(·) denotes the stochastic augmentation pipeline of Section 3.2.4, applied during training only. The resulting input tensor has shape (B, 45, 3, 224, 224), where B is the batch size (Figure 1).

3.2.4 Data augmentation

Three pixel-space augmentation families and one temporal augmentation were applied during training, all disabled at evaluation.

MixUp [24] constructs synthetic examples by convex combination of pairs of sequences and their labels:

$$\tilde{x} = \lambda x_i + (1\text{-}\lambda)x_j, \quad \tilde{y} = \lambda y_i + (1\text{-}\lambda)y_j \tag{3}$$

where x_i and x_j are two training sequences with labels y_i and y_j, and $\lambda \sim \text{Beta}(\alpha,\alpha)$ with α in the range 0.1–0.2, chosen so that mixing is conservative and temporal coherence within each sequence is preserved.

[OFFICIAL]

Random erasing [25] removes a rectangular region with probability $p = 0.5$, with area fraction sampled from 0.02–0.2 and aspect ratio from 0.5–2.0, encouraging the model to distribute evidence across the frame rather than relying on a single region.

Colour-space augmentation applies random adjustment of brightness, contrast, saturation and hue, simulating variation in capture conditions and sensor response.

Temporal jitter applies the frame-index offsets described in Section 3.2.2.

Figure 2 shows the effect of colour-space augmentation and random erasing on a representative frame, alongside the unaugmented input. MixUp and temporal jitter are not shown, for the reasons given in the caption.

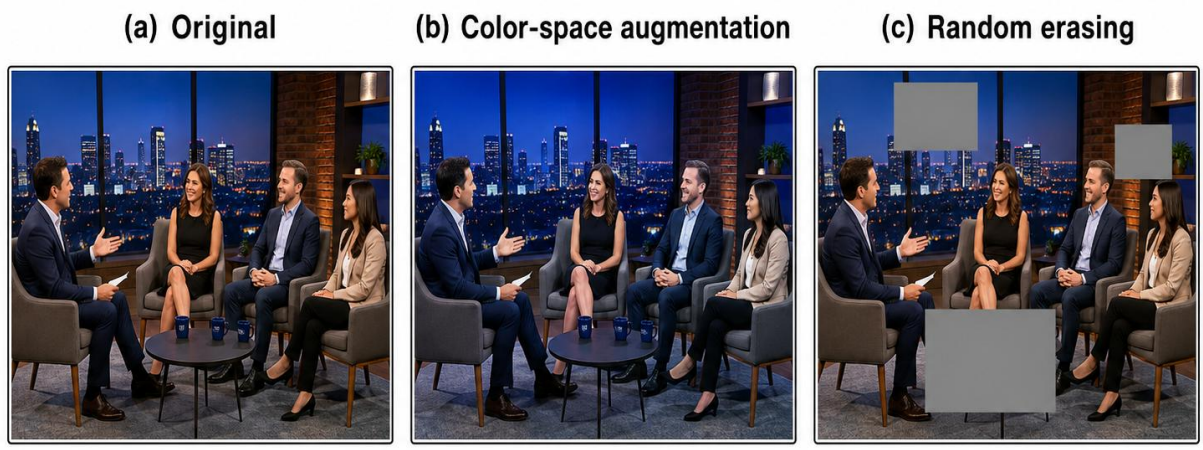

Figure 2. Examples of training-time data augmentation. Effect of the pixel-space augmentations applied during training, shown on a representative multi-person frame: (a) the unaugmented frame; (b) colour-space augmentation, applying random adjustment of brightness, contrast, saturation and hue; (c) random erasing, in which rectangular regions are removed with probability $p = 0.5$ and area fraction 0.02–0.2. All augmentations are disabled at evaluation. Frames are shown at full scene extent, reflecting the whole-frame processing described in Section 3.2.1: the scene contains several faces, of which a manipulation may affect only a subset. MixUp (Equation 3) and temporal jitter (Section 3.2.2) are not illustrated, because neither produces a visually informative effect on a single still frame MixUp because the mixing coefficients used are deliberately conservative, and temporal jitter because it alters frame selection rather than frame content.

3.3 Proposed architecture

The architecture comprises a spatial feature extractor, a recurrent temporal stage with attention, a Transformer encoder and a classification head. Figure 1 shows the complete data flow with tensor dimensions at every stage.

[OFFICIAL]

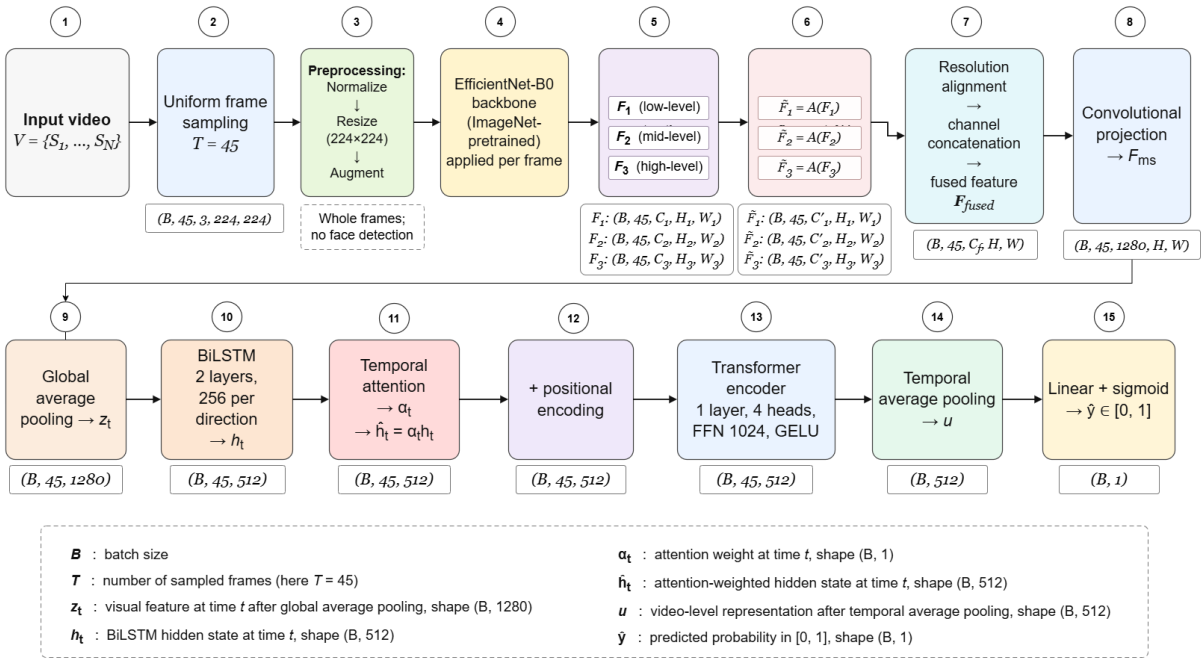

Figure 1. Overall architecture of the proposed framework. Overall architecture of the proposed framework. Each of the $T = 45$ uniformly sampled frames is preprocessed and passed through an ImageNet-pretrained EfficientNet-B0 backbone, from which three feature maps F_1, F_2 and F_3 are extracted at low, intermediate and high semantic levels (block 5). Each is refined by the attention operation $\mathcal{A}(\cdot)$ of Equation (5) (block 6), aligned to a common resolution and concatenated (Equation 6, block 7), projected by a convolution (Equation 7, block 8) and reduced by global average pooling to the frame embedding z_t (Equation 8, block 9). The embedding sequence is processed by a two-layer bidirectional long short-term memory network (Equations 10–11, block 10) and rescaled by the temporal attention weights α_t (Equations 12–14, block 11), then combined with a positional encoding (Equation 15, block 12) and refined by a single-layer Transformer encoder with four attention heads (Equations 16–20, block 13). Temporal average pooling (Equation 21, block 14) and a linear layer with sigmoid activation (Equation 22, block 15) produce the manipulation probability $\hat{y}$. Tensor shapes are given beneath each stage, with B the batch size. Complete frames are processed; no face detection or cropping is applied (Section 3.2.1). BiLSTM, bidirectional long short-term memory; FFN, feed-forward network; GELU, Gaussian error linear unit.

3.3.1 Spatial feature extraction

Spatial features are extracted by an ImageNet-pretrained EfficientNet-B0 backbone [26] applied independently to each of the $T = 45$ frames, augmented with a multi-scale, attention-refined feature path. Three intermediate feature maps are taken from different depths of the backbone,

$$F_1 \in \mathbb{R}^{H_1 \times W_1 \times C_1}, \quad F_2 \in \mathbb{R}^{H_2 \times W_2 \times C_2}, \quad F_3 \in \mathbb{R}^{H_3 \times W_3 \times C_3} \tag{4}$$

where H_i, W_i and C_i denote the height, width and channel count of the i-th map. The three depths correspond to low-level detail (texture, noise residual, edges), mid-level structure (localised artefact patterns, region boundaries) and high-level semantics (facial geometry, global coherence), as indicated in Figure 1. The candidate tap points are fixed by the backbone geometry: for a 224 × 224 input, the successive EfficientNet-B0 stages emit feature maps of 112 × 112 × 16, 56 × 56 × 24, 28 × 28 × 40, 14 × 14 × 80, 14 × 14 × 112, 7 × 7 × 192 and 7 × 7 × 320 [26], so the low-, mid- and high-level taps are drawn from the early, intermediate and terminal groups of this progression.

Each map is refined by an attention operation:

$$\widetilde{F}_i = \mathcal{A}(F_i), \quad i \in \{1, 2, 3\} \tag{5}$$

[OFFICIAL]

where $\mathcal{A}(\cdot)$ is a learnable attention function that increases the relative weight of manipulation-relevant activations. The operation preserves spatial resolution and maps C_i channels to C_i'. Attention-based refinement of convolutional features in this style follows an established line of work [15], [16].

The refined maps are aligned to a common spatial resolution and concatenated along the channel axis:

$$F_{\text{fused}} = \text{Concat}(\widetilde{F_1}, \widetilde{F_2}, \widetilde{F_3}) \in \mathbb{R}^{H \times W \times C_f} \tag{6}$$

where C_f is the summed channel count of the three aligned maps, then projected by a convolution to a fixed width:

$$F_{\text{ms}} = \text{Conv}(F_{\text{fused}}) \in \mathbb{R}^{H \times W \times 1280} \tag{7}$$

Multi-scale aggregation of this kind follows the feature-pyramid literature [14]. Global average pooling over the spatial dimensions produces the frame-level embedding:

$$z_t = \text{GAP}(F_{\text{ms}}) \in \mathbb{R}^{1280} \tag{8}$$

where z_t is the embedding of the t-th sampled frame. The embedding sequence therefore has shape (B, 45, 1280), as shown in Figure 1.

3.3.2 Recurrent temporal modelling with attention

The frame embeddings form the sequence

$$Z = \{z_1, z_2, \ldots, z_T\} \tag{9}$$

of shape (B, 45, 1280), which is processed by a two-layer bidirectional LSTM [17], [27] with 256 units per direction, reading the sequence in both directions:

$$\overrightarrow{h_t} = \text{LSTM}_f\left(z_t, \overrightarrow{h_{t-1}}\right), \quad \overleftarrow{h_t} = \text{LSTM}_b\left(z_t, \overleftarrow{h_{t+1}}\right) \tag{10}$$

$$h_t = \text{Concat}\left(\overrightarrow{h_t}, \overleftarrow{h_t}\right) \in \mathbb{R}^{512} \tag{11}$$

giving a hidden-state sequence of shape (B, 45, 512). Bidirectional processing is used because manipulation evidence is often clearest at a transition the onset or offset of an artefact burst which a unidirectional model observes from one side only.

An additive attention mechanism [28] assigns each position a scalar score,

$$e_t = w^{\top} \tanh(W_a h_t + b_a) \tag{12}$$

normalised across the sequence,

$$\alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^{T} \exp(e_k)}, \quad \sum_{t=1}^{T} \alpha_t = 1 \tag{13}$$

where $W_a \in \mathbb{R}^{da \times 512}$, $b_a \in \mathbb{R}^{da}$ and $w \in \mathbb{R}^{da}$ are learnable and d_a is the attention projection dimension. Each hidden state is then rescaled by its normalised weight,

$$\hat{h}_t = \alpha_t h_t \in \mathbb{R}^{512}, \quad t = 1, \ldots, T \tag{14}$$

producing the weighted sequence $\hat{H} = \{\hat{h}_1, \ldots, \hat{h}_T\}$ of shape (B, 45, 512), which is passed to the Transformer encoder. The mechanism is element-wise rescaling rather than pooling: the sequence

[OFFICIAL]

length is preserved, and positions carrying weak evidence are attenuated before global modelling rather than being summed away. Figure 3 shows the module in detail.

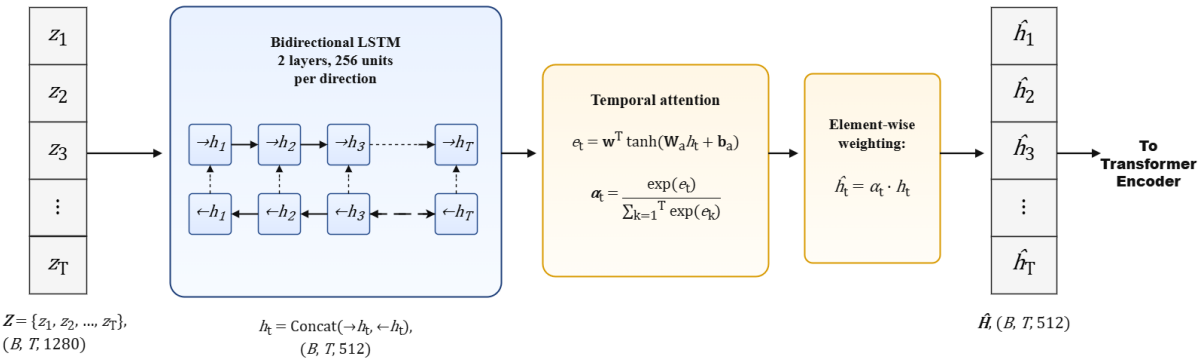

Figure 3. Temporal attention mechanism. Temporal attention mechanism. The frame-embedding sequence $Z = \{z_1,\ldots,z_T\}$ of shape (B, T, 1280) is processed by a two-layer bidirectional long short-term memory network with 256 units per direction; the forward and backward hidden states are concatenated to give h_t of shape (B, T, 512) (Equation 11). An additive attention function assigns each position a score e_t, normalised across the sequence to α_t (Equations 12–13). Each hidden state is then rescaled element-wise by its weight, $\hat{h}_t = \alpha_t \cdot h_t$ (Equation 14), producing the weighted sequence $\hat{H}$ of shape (B, T, 512) that is passed to the Transformer encoder. The operation preserves sequence length: positions carrying weak evidence are attenuated rather than summed away. B, batch size; T, sequence length ($T = 45$ in the reported configuration).

3.3.3 Transformer encoder

Positional information is added to the weighted sequence, since self-attention is permutation-invariant:

$$H^{(0)} = \hat{H} + P \in \mathbb{R}^{T \times d_{\text{model}}}, \quad d_{\text{model}} = 512 \tag{15}$$

where $P \in \mathbb{R}^{45 \times 512}$ is the positional-encoding matrix. Scaled dot-product attention [10] is computed as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_h}}\right)V \tag{16}$$

where Q, K, $V \in \mathbb{R}^{T \times dh}$ are the query, key and value matrices and d_h is the per-head dimension; division by $\sqrt{(d_h)}$ keeps the dot products in a range where the softmax gradient does not vanish. Attention is computed in $n_h = 4$ parallel heads,

$$\text{head}_j = \text{Attention}\left(H^{(0)}W_j^Q, H^{(0)}W_j^K, H^{(0)}W_j^V\right) \tag{17}$$

$$\text{MHA}\left(H^{(0)}\right) = \text{Concat}\left(\text{head}_1, ..., \text{head}_{n_h}\right)W^O \tag{18}$$

with W_j^Q, W_j^K, $W_j^V \in \mathbb{R}^{512 \times dh}$ and $W^O \in \mathbb{R}^{nh\,dh \times 512}$, where $d_h = d_{\text{model}}/n_h = 512/4 = 128$. A single encoder layer applies multi-head attention and a position-wise feed-forward network, each within a residual connection followed by layer normalisation:

$$U = \text{LayerNorm}\left(H^{(0)} + \text{MHA}\left(H^{(0)}\right)\right) \tag{19}$$

$$H^{(1)} = \text{LayerNorm}\left(U + \text{FFN}(U)\right), \quad \text{FFN}(u) = W_2\gamma(W_1u + b_1) + b_2 \tag{20}$$

where γ is the Gaussian error linear unit activation and the feed-forward width is 1,024, a twofold expansion of d_{model}. Residual connections preserve gradient flow and layer normalisation stabilises the activation scale. The encoder output retains the shape (B, 45, 512).

[OFFICIAL]

3.3.4 Classification head

The encoder output is averaged over time and classified:

$$u = \frac{1}{T}\sum_{t=1}^{T} h_t^{(1)} \in \mathbb{R}^{512} \tag{21}$$

$$\hat{y} = \sigma(w_c^{\top} u + b_c) \in [0, 1] \tag{22}$$

where $h_t^{(1)}$ is the t-th row of $H^{(1)}$, $\sigma(\cdot)$ is the logistic sigmoid, and $\hat{y}$ is the predicted probability that the input is fake. The video-level representation u has shape (B, 512) and the output (B, 1).

3.4 Training objective and optimisation

The model is trained with binary cross-entropy:

$$\mathcal{L}_{\text{BCE}} = -[y\log\hat{y} + (1-y)\log(1-\hat{y})] \tag{23}$$

where $y \in \{0,1\}$ is the ground-truth label, with 1 denoting fake, and $\hat{y}$ the predicted probability of Equation (22).

Three measures address the memory cost of training on video sequences. Gradients are accumulated over K micro-batches before each parameter update, giving an effective batch size $K \times B$ without a proportional increase in activation memory:

$$g = \frac{1}{K}\sum_{k=1}^{K} \nabla_\theta \mathcal{L}^{(k)} \tag{24}$$

where $\mathcal{L}^{(k)}$ is the loss on micro-batch k and θ the model parameters; with $B = 6$ and $K = 3$ the effective batch size is 18. An exponential moving average of the parameters [29] is maintained,

$$\theta_{\text{ema}} \leftarrow \beta\theta_{\text{ema}} + (1-\beta)\theta \tag{25}$$

with smoothing factor $\beta = 0.999$, which suppresses the parameter oscillation induced by a small micro-batch size. Mixed-precision arithmetic [30] and chunked frame processing further reduce activation memory, the latter by applying the backbone to sub-sequences of frames rather than to all 45 simultaneously. A differential learning rate assigns the pretrained backbone a lower rate than the newly initialised modules, so that ImageNet-derived low-level filters are perturbed only slightly while randomly initialised modules converge at a normal rate.

The effect of these measures on peak memory was not measured, and no quantitative memory-reduction claim is made.

3.5 Training configuration

Table 2 gives the training configuration. Values are reported only where documented; conventional values have not been substituted for undocumented ones.

Table 2. Training and implementation configuration. Dimensions marked † are determined by the architecture and are confirmed by the tensor annotations in Figures 1 and 3.

Category	Parameter	Value
Input	Sequence length T	45
	Frame resolution	224 × 224

[OFFICIAL]

	Input tensor	(B, 45, 3, 224, 224) †
Backbone	Architecture	EfficientNet-B0, ImageNet-pretrained, with multi-scale attention-refined path
	Frame embedding width	1,280 †
Temporal	BiLSTM layers	2
	Hidden units per direction	256
	Hidden-state width	512 †
	Dropout	0.2
Global	Transformer layers	1
	Model dimension dmodel	512 †
	Attention heads nh	4
	Per-head dimension dh	128 †
	Feed-forward width	1,024
	Activation	Gaussian error linear unit
Optimisation	Optimiser	AdamW [31]
	Weight decay	5×10^{-5}
	Learning rate, backbone	2.5×10^{-6}
	Learning rate, other modules	5×10^{-5}
	Schedule	Cosine annealing with warm-up [32]
	Epochs	30
	Micro-batch size B	6
	Accumulation steps K	3
	Effective batch size	18
	Gradient clipping	1.0 (global norm)
Regularisation	Dropout	0.2
	Exponential moving average decay β	0.999
Augmentation	MixUp α	0.1–0.2
	Random erasing	p = 0.5; area 0.02–0.2; aspect ratio 0.5–2.0
	Colour jitter	Brightness, contrast, saturation, hue
	Temporal jitter	Random frame-index offset
Memory	Mixed precision	Enabled
	Chunked frame processing	Enabled
Environment	Graphics processing unit	NVIDIA RTX 3060, 12 GB
	Central processing unit / memory	AMD Ryzen 5 5600 / 32 GB
	Framework	PyTorch [33] with CUDA acceleration

3.6 Computational-cost estimation

The EfficientNet-B0 backbone requires approximately 0.45 GFLOPs for a single 224 × 224 frame. Because the backbone is applied independently to each of the T sampled frames (Figure 1), the spatial path's cost per video scales linearly:

$$C_{\text{spatial,video}} \approx T \times C_{\text{frame}} = 45 \times 0.45 = 20.25 \text{ GFLOPs} \quad (26)$$

This figure is an analytical estimate of the spatial path only. It excludes the bidirectional LSTM, the temporal attention, the Transformer encoder, the classification head and all preprocessing. No end-to-end profiling was performed, and no inference latency, throughput or peak memory was measured. The value should therefore be read as a lower bound on per-video arithmetic cost, not as a measured system cost. Reporting both bases matters because the per-frame figure alone understates the cost of a

[OFFICIAL]

video-level system by more than an order of magnitude, a conflation present in earlier reporting of this architecture family (Section 2).

3.7 Evaluation metrics

Let TP, TN, FP and FN denote true positives, true negatives, false positives and false negatives, with **fake as the positive class**, matching the axis labels of Figure 4. Predictions are thresholded at 0.5. The reported metrics are:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{27}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{28}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{29}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{30}$$

$$F_1 = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{31}$$

Accuracy is treated as an informative primary metric because the evaluation set is close to balanced (Section 3.1.2). The Matthews correlation coefficient, which summarises a binary confusion matrix in a single value insensitive to the designation of the positive class [34], is

$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \tag{32}$$

The area under the receiver operating characteristic curve [35] is computed from the continuous predicted probabilities across all decision thresholds.

An interval estimate for the observed accuracy proportion $\hat{p}$ on n evaluation instances is obtained by the Wilson score method [36]:

$$\text{CI}_{95\%} = \frac{\hat{p} + \frac{z^2}{2n} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}, \quad z = 1.96 \tag{33}$$

Scope of Equation (33). This interval describes the sampling variability of a single observed proportion on one evaluation set of fixed size. It does not describe variability across random initialisations, data orderings or repeated training runs, none of which were assessed in this study, and it is not used to support any comparative claim, since no matched predictions, repeated runs or paired statistical test are available.

4. Results

This section reports a single training configuration evaluated once on a single held-out evaluation set. No repeated runs, component ablations, matched baseline reproductions or statistical hypothesis tests were performed, and none is reported.

[OFFICIAL]

4.1 Primary evaluation results

Table 3 gives the per-class classification report and Table 4 the summary metrics. On 1,987 evaluation instances the integrated framework attained an accuracy of 96.18 %.

Table 3. Per-class classification report on the held-out evaluation set. Fake is the positive class. Class names match the axis labels of Figure 4. Values are given to four decimal places to expose the small inter-class differences that two-decimal rounding conceals.

Class	Precision	Recall	F1-score	Support
Real	0.9619	0.9619	0.9619	998
Fake	0.9616	0.9616	0.9616	989
Accuracy			**0.9618**	**1,987**
Macro average	0.9618	0.9618	0.9618	1,987
Weighted average	0.9618	0.9618	0.9618	1,987

Precision, recall and F1-score coincide within each class, and the macro and weighted averages coincide with the overall accuracy. This is a structural consequence of the confusion matrix rather than a coincidence: because the false-positive and false-negative counts are equal at 38 each, each class's row total equals its column total, so precision and recall share a denominator within each class. The property is noted because it means the three per-class metrics carry no independent information in this evaluation, and they are not treated as though they did.

Table 4. Summary performance on the held-out evaluation set. All values are computed from the confusion matrix of Figure 4, except the area under the curve, which is taken from the receiver operating characteristic output of Figure 5.

Metric	Value	Source
Accuracy	96.18 %	Equation (27), Figure 4
95 % Wilson score interval for accuracy	[95.24 %, 96.93 %]	Equation (33); see scope statement, Section 3.7
Macro precision / recall / F1-score	0.96	Equations (28), (29), (31)
Specificity	96.19 %	Equation (30)
Sensitivity (recall, fake)	96.16 %	Equation (29)
False-positive rate	3.81 %	Figure 4
False-negative rate	3.84 %	Figure 4
Overall error rate	3.82 %	Figure 4
Balanced accuracy	96.18 %	Equations (29), (30)
Matthews correlation coefficient	0.9235	Equation (32)
Area under the ROC curve	0.991	Figure 5

4.2 Confusion-matrix analysis

Figure 4 presents the confusion matrix. Of 998 real instances, 960 were classified correctly and 38 were classified as fake. Of 989 fake instances, 951 were classified correctly and 38 were classified as real. The total error is 76 instances of 1,987, an error rate of 3.82 %.

[OFFICIAL]

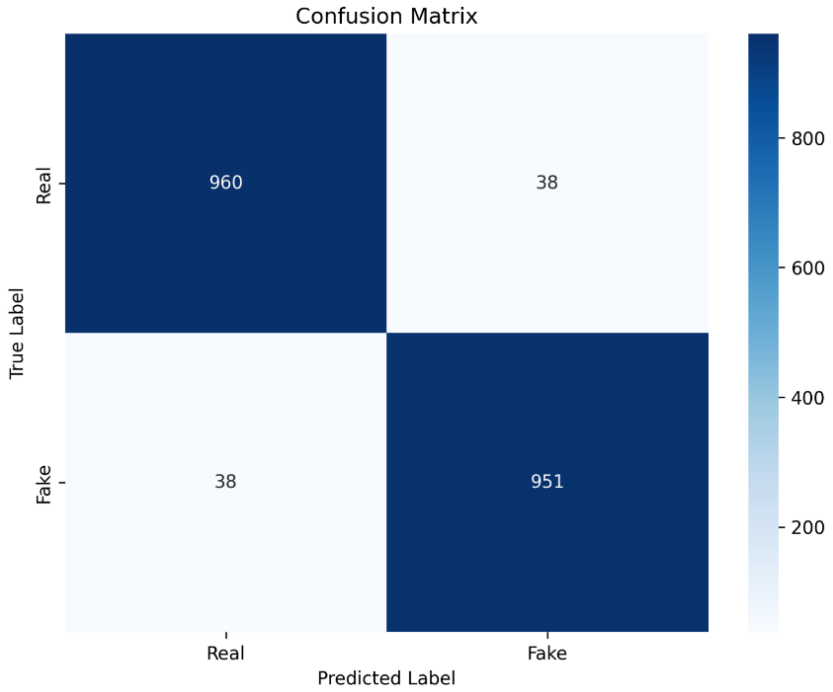

Figure 4. Confusion matrix on the held-out evaluation set. Confusion matrix on the held-out evaluation set (n = 1,987) at a decision threshold of 0.5, with fake as the positive class. Rows denote true labels, columns predicted labels, and cell values are instance counts. Of 998 real instances, 960 were classified correctly and 38 were classified as fake; of 989 fake instances, 951 were classified correctly and 38 were classified as real. These counts yield an accuracy of 96.18 %, a false-positive rate of 3.81 % and a false-negative rate of 3.84 % (Table 4).

The two error types are numerically equal, giving a false-positive rate of 3.81 % and a false-negative rate of 3.84 % a difference of 0.03 percentage points. At the 0.5 decision threshold the classifier therefore shows no measurable directional bias on this evaluation set. This balance follows from the near-equal class frequencies of the evaluation set (50.23 % against 49.77 %) together with the operating threshold; the training objective was unweighted binary cross-entropy (Equation 23), so no loss-weighting mechanism contributed to it.

The Matthews correlation coefficient of 0.9235 provides a threshold-fixed summary that accounts for all four matrix cells and is invariant to the choice of positive class. Its value is consistent with the balanced per-class rates.

4.3 Receiver operating characteristic analysis

Figure 5 presents the receiver operating characteristic curve, which describes the trade-off between the true-positive and false-positive rates across decision thresholds. The area under the curve is 0.991.

The curve rises steeply in the low-false-positive region, remaining above a true-positive rate of approximately 0.9 at false-positive rates below approximately 0.02. This region is the one relevant to a screening application, in which a low false-alarm rate is required.

The area under the curve exceeds the fixed-threshold accuracy of 96.18 %, indicating that the ranking induced by the model's continuous scores separates the classes more completely than the 0.5 threshold

[OFFICIAL]

decision extracts. The area under the curve is reported as read from the stored evaluation output and is not derived from the confusion matrix; a confusion matrix describes a single operating point and does not determine the area under the full curve.

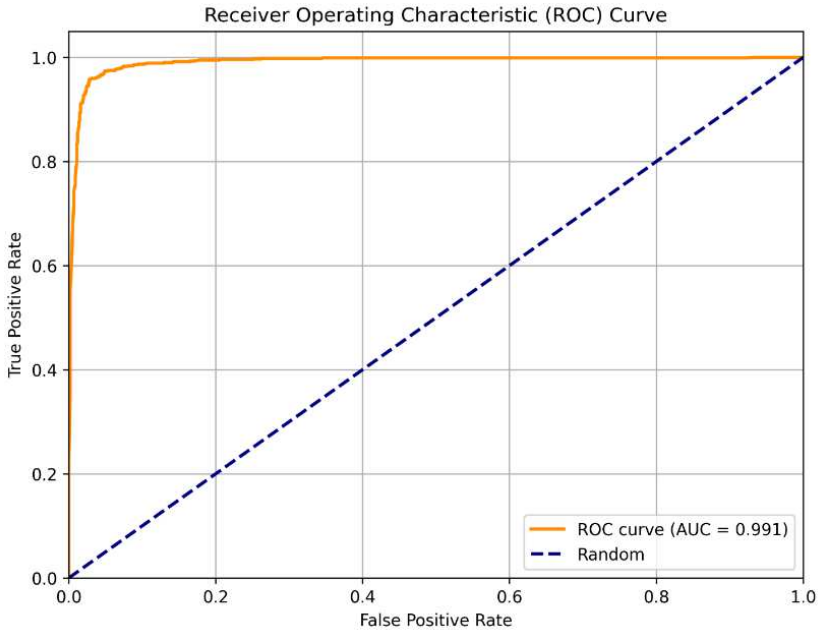

Figure 5. Receiver operating characteristic curve. Receiver operating characteristic curve on the held-out evaluation set, showing the trade-off between the true-positive and false-positive rates across decision thresholds. The dashed diagonal denotes chance performance. The area under the curve is 0.991, computed from the model's continuous output probabilities and not derived from the confusion matrix of Figure 4.

4.4 Training behaviour

Figure 6 shows the loss and accuracy curves across the 30 training epochs. Both losses decline without oscillation or divergence, and validation accuracy rises from approximately 78 % at initialisation to approximately 96 % in the final epochs, where it plateaus. Validation loss does not rise at any point during training, indicating that the model did not enter an overfitting regime within the 30-epoch budget.

Validation loss remains below training loss throughout, and validation accuracy exceeds training accuracy for the majority of the run. This ordering is consistent with the training configuration: MixUp, random erasing, colour jitter and dropout are active when the training metrics are computed and inactive when the validation metrics are computed, so the training metrics are evaluated on a systematically harder distribution.

[OFFICIAL]

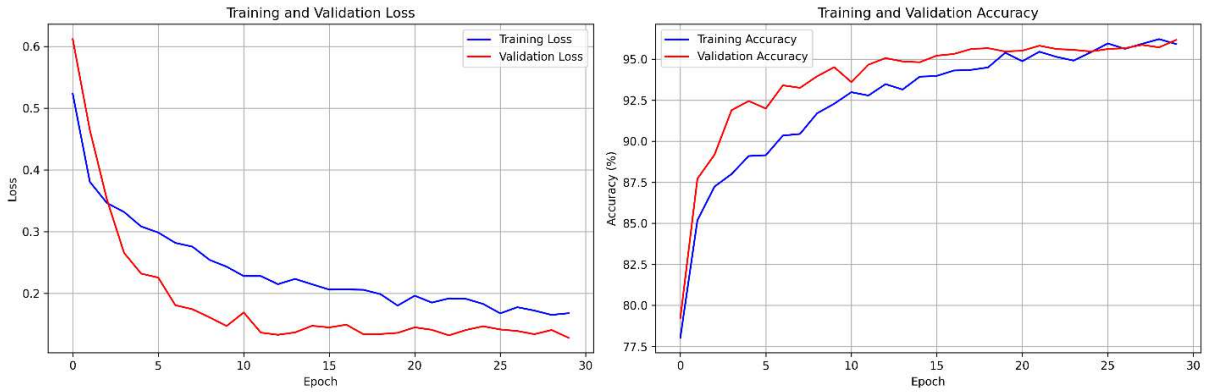

Figure 6. Training and validation performance curves. Training and validation performance across 30 epochs: (left) training and validation loss; (right) training and validation accuracy. Validation accuracy rises from approximately 78 % at initialisation to approximately 96 % in the final epochs, where it plateaus, and validation loss does not rise at any point. Training metrics are computed with MixUp, random erasing, colour jitter and dropout active, whereas validation metrics are computed with all augmentation and dropout disabled; the training metrics are therefore evaluated on a systematically harder distribution, which accounts for the validation curves lying favourably relative to the training curves throughout (Section 4.4).

4.5 Comparison with previous studies

Table 1 records the results reported by the studies reviewed in Section 2. Those figures are not comparable with the result reported here. They were obtained on different datasets, on different subsets and splits, at different evaluation units, and under different preprocessing pipelines most of them using face-cropped input, whereas the present pipeline processes whole frames (Section 3.2.1). No baseline architecture was retrained under the present protocol in this study, and no matched comparison is offered.

One study warrants specific treatment. The authors' earlier work [18] evaluated an EfficientNet-B0 backbone with a feature pyramid and a temporal convolutional network on FFIW-10K, reporting 92.45 % accuracy at approximately 0.45 GFLOPs per frame. That result was taken from the earlier publication and is included as a literature-based reference point rather than as a newly reproduced baseline. It was not rerun under the present protocol. Furthermore, the two studies' evaluation protocols cannot be shown to be equivalent, because the subset construction, partitioning key and evaluation unit used here are not fully documented (Section 3.1.2). The 3.7-percentage-point difference between 92.45 % and 96.18 % cannot therefore be attributed to the architectural differences described in Section 1. It is reported as a difference between two separately conducted studies on the same benchmark, and no causal interpretation is offered.

5. Discussion

5.1 Interpretation and comparison with previous work

The integrated framework separated the two classes on the held-out evaluation set with an accuracy of 96.18 %, a 95 % Wilson score interval of 95.24–96.93 %, and an area under the receiver operating characteristic curve of 0.991. Errors were numerically balanced at 38 in each direction, giving a false-

[OFFICIAL]

positive rate of 3.81 %, a false-negative rate of 3.84 % and an overall error rate of 3.82 %, with a Matthews correlation coefficient of 0.9235. Validation loss did not rise across 30 epochs. These are the study's direct findings, and each follows from the confusion matrix of Figure 4, the receiver operating characteristic output of Figure 5 or the training curves of Figure 6.

Two further observations follow directly from these numbers. First, the area under the curve exceeds the accuracy attainable at the 0.5 threshold, so the score ranking is stronger than the fixed-threshold decision realises; threshold calibration on a held-out partition would be expected to improve the operating point, although no such calibration was performed here. Second, the interval on the accuracy proportion spans 1.69 percentage points. This is a useful scale reference: differences between systems smaller than approximately this width cannot be distinguished on an evaluation set of this size even under repeated measurement of the same quantity, quite apart from run-to-run training variance, which was not assessed.

Beyond these direct findings, interpretation must be separated from evidence. The architecture combines multi-scale attention-refined spatial encoding with two temporal mechanisms whose ranges are complementary in principle: recurrence supplies locally contextualised representations, and self-attention relates positions without a distance penalty. It is plausible that this combination contributes to the observed separation, and plausible that rescaling frame representations before global attention (Equation 14) concentrates capacity on positions carrying evidence. Neither interpretation was evaluated through ablation. The observed result is a property of the integrated system and cannot be decomposed; the findings suggest that the integrated architecture is capable of separating the classes in this setting, but the independent contribution of each module was not quantified in the current study. Similarly, the equality of the false-positive and false-negative counts is consistent with a well-placed threshold on a near-balanced evaluation set, and should not be read as evidence of a robustness property, since class balance is a characteristic of the evaluation data rather than of the model.

Direct numerical comparison with the studies in Table 1 is not valid, for the reasons given in Section 4.5. Three structural observations can nonetheless be made. The accuracy reported here falls within the band reported by the recurrent spatio-temporal detectors of Section 2 (91.8–98.2 %); however, those results were obtained on FaceForensics++, Celeb-DF v2 and DFDC, which are predominantly single-subject benchmarks, whereas FFIW-10K averages 3.15 faces per frame [13]. The two settings pose different problems, and the present result should be read neither as competitive with nor as inferior to those figures. Next, the finding of Bastos et al. [9] that adding raw red–green–blue input to a video vision Transformer with handcrafted descriptors degraded performance is consistent in direction with the design principle motivating Equation (14), namely that constraining what a global attention stage observes may help rather than hinder; this is a consistency of motivation, not of evidence, since no experiment here tests it. Finally, almost all comparable detectors in Table 1 apply face cropping, while the present pipeline does not. This is the most consequential methodological difference between this

[OFFICIAL]

study and the literature it sits alongside, and it bears directly on how the result should be interpreted, as the following section sets out.

Three questions remain open. Whether the Transformer stage contributes beyond the recurrent representation is undetermined. Whether the temporal attention learns a peaked weight distribution as its design intends or a near-uniform one is also undetermined, since the learned weights were not examined. And whether the multi-scale spatial path contributes beyond a single-scale backbone is untested. Each requires an experiment not conducted here.

5.2 Limitations and threats to validity

The limitations below constrain the interpretation of the result; they do not affect the reported measurements, all of which were recomputed from the confusion matrix and verified during preparation of this manuscript.

Whole-frame processing and the risk of scene shortcuts. This is the study's principal threat to construct validity. The pipeline resizes complete frames to 224 × 224 without face detection, cropping or tracking (Section 3.2.1). On FFIW-10K, where frames contain 3.15 faces on average and only a subset is manipulated in many videos [13], each face occupies a small fraction of a whole-frame input. Under these conditions the classifier may be responding to cues other than facial manipulation evidence: scene composition, background content, compression and re-encoding signatures, editing artefacts, or programme- and identity-specific characteristics correlated with the label. No control experiment was conducted to exclude this, no face-masked or background-only condition was evaluated, and no saliency or attention visualisation was produced. The reported accuracy therefore establishes that the model separates the two classes on this evaluation set; it does not establish that the separation is driven by manipulation artefacts. This distinction is material for any inference about deployment.

Unresolved evaluation unit and partition. The 1,987 evaluation instances cannot be reconciled with the stated 70/15/15 partition of an approximately 3,000-video subset, and their unit is not established (Section 3.1.2). The partition from which they were drawn validation or test is likewise not established. If the reported figures are best-epoch validation results on a partition that also guided early stopping and hyperparameter selection, they carry a selection-induced optimistic bias, since the maximum over epochs of a fluctuating quantity is biased upward. A per-sequence or per-frame accuracy would also not be directly comparable with the per-video accuracies reported elsewhere in the literature.

Unverified partition disjointness. The partitioning key and the mechanism by which subject-level separation was maintained are not documented, and disjointness was not verified (Section 3.1.2). If partitioning was performed at a level below the source identity or the source video for example at the level of extracted sequences then material derived from the same source could appear in more than one partition, which would inflate the reported figures by an unknown margin. This possibility cannot be

[OFFICIAL]

excluded from available material, and releasing the partition manifests would be the most direct way to address it.

Absence of ablation, repeated runs and statistical testing. No component-wise ablation was performed, so no module's independent contribution is known. The reported results derive from a single training run; the random seed is not documented, no repeated-seed trials were conducted, and stability across random initialisations was therefore not assessed. No statistical hypothesis test was performed, and no claim of statistical significance is made anywhere in this paper. The Wilson interval reported in Table 4 describes the sampling variability of one observed proportion only and carries no information about run-to-run variability.

Literature-derived rather than reproduced baselines. No baseline architecture was retrained under the present protocol. The comparison figures in Table 1 are author-reported values from other studies on other datasets, and the 92.45 % attributed to [18] is a literature-derived reference point rather than a reproduction (Section 4.5). The study therefore provides no matched comparative evidence, and no claim of superiority over any prior method is made.

Single-dataset evaluation. Evaluation is confined to a subset of one benchmark. No cross-dataset transfer was evaluated, and no robustness sweep over compression, resolution or noise was conducted. Generalisation to unseen identities, unseen manipulation methods, other codecs and other distribution platforms is not established. Because a subset rather than the full benchmark was used, comparability even with published FFIW-10K results is reduced. Additional external validation is required before any claim of general capability could be made.

Estimated rather than measured efficiency. The per-video figure of approximately 20.25 GFLOPs is an analytical estimate of the spatial path alone (Section 3.6). No end-to-end profiling was performed, no inference latency or throughput was measured, and peak memory usage was not recorded. The effect of the memory-efficiency measures was consequently not quantified, and no claim of real-time or edge-device capability is made.

Incomplete implementation documentation. The architecture's topology, tensor dimensions and training regime are reported in full, but three groups of detail material to exact reproduction are not. In the spatial path, the internal form of the attention operator $\mathcal{A}(\cdot)$ in Equation (5), the specific backbone stages supplying F_1, F_2 and F_3, the resolution-alignment operator, and the kernel configuration of the fusion convolution in Equation (7) are not specified. In the temporal and global stages, the attention projection dimension d_a in Equation (12) and the form of the positional encoding in Equation (15) are not specified. In the training configuration, the normalisation constants, the exact MixUp mixing parameter within the stated 0.1–0.2 range, the colour-jitter magnitudes, whether random erasing is applied per frame or per clip, the temporal-jitter magnitude and whether it remains active at evaluation, the mixed-precision data type, the chunk size, the random seed, the set of frozen

[OFFICIAL]

parameters, whether evaluation used the exponential-moving-average parameters, and the early-stopping and checkpoint-selection criteria are not specified. An independent reimplementation would therefore reproduce the architecture and its data flow but not necessarily every operator or hyperparameter, and could not be expected to reproduce the reported figures exactly.

Task scope. The model produces a single video-level binary decision. It does not identify which face in a multi-person frame was manipulated, does not localise manipulation in time, and does not identify the manipulation method, so it does not support forensic use cases requiring any of these.

5.3 Practical implications and future work

Subject to the limitations above, the study has two practical implications. The first concerns training economy. The complete architecture a convolutional backbone applied to 45 frames, a two-layer bidirectional recurrent network and a Transformer encoder was trained to convergence within 30 epochs on a single 12 GB consumer graphics processing unit, using gradient accumulation, mixed-precision arithmetic, chunked frame processing and exponential moving averaging (Section 3.4). This matters in a field where detection models degrade as generation methods evolve, so that the ability to retrain economically against newly emerging generators is a functional requirement rather than a convenience. The measures used are architecture-agnostic and transfer to any video model built on a per-frame backbone.

The second concerns how efficiency should be reported. Separating per-frame from per-video cost (Section 3.6) changes the apparent complexity of this architecture family by more than an order of magnitude from 0.45 to approximately 20.25 GFLOPs. Because six of the ten studies in Table 1 report no complexity figure and the remainder use incompatible conventions, the accuracy–efficiency trade-offs claimed across this literature are not currently comparable. Adopting an explicit per-frame and per-video convention would be a low-cost improvement to reporting practice in this area, independent of any architectural contribution.

No deployment claim is made. The evaluation hardware is a desktop graphics processing unit rather than an edge platform, no end-to-end latency or throughput was measured, and the results are limited to the reported evaluation setting.

Future work should address the evidential gaps in order of their weight. The most consequential is a shortcut-learning control: comparing face-cropped, whole-frame and face-masked inputs, supported by saliency analysis and by inspection of the learned temporal attention weights, would establish what the model actually responds to and would resolve the principal threat to construct validity identified in Section 5.2. A component-wise ablation over the multi-scale path, the spatial attention refinement, the temporal attention and the Transformer stage, repeated across random seeds and reported with paired statistical tests, would establish which architectural elements contribute and would convert the interpretations of Section 5.1 into findings. Retraining representative baselines under an identical

[OFFICIAL]

protocol would supply the matched comparison this study lacks. Cross-dataset evaluation on FaceForensics++, Celeb-DF v2 and DFDC without fine-tuning, together with a robustness sweep over compression and resolution, would characterise generalisation. End-to-end profiling of latency, throughput and peak memory would replace the analytical cost estimate of Equation (26) with a measured system cost. Finally, extending the video-level decision to per-face and per-frame localisation would address the multi-person setting that FFIW-10K is specifically constructed to pose, and would move the architecture from classification toward the forensic capability that deployment ultimately requires.

6. Conclusion

This study described and evaluated an integrated architecture for video-level deepfake classification, combining an EfficientNet-B0 spatial backbone with a multi-scale attention-refined feature path, a two-layer bidirectional long short-term memory network with additive temporal attention over frames, and a single-layer Transformer encoder over the weighted sequence, followed by temporal averaging and binary classification. Training used gradient accumulation, mixed-precision arithmetic, chunked frame processing, exponential moving averaging and differential learning rates, allowing 45-frame sequences to be trained on a single 12 GB consumer graphics processing unit.

Evaluated on a subset of the FFIW-10K multi-person benchmark, the framework attained an accuracy of 96.18 % on a held-out evaluation set of 1,987 instances, with a 95 % Wilson score interval of 95.24–96.93 % for the observed proportion. Precision, recall and F1-score were 0.96 for both classes, specificity was 96.19 %, the Matthews correlation coefficient was 0.9235, and the area under the receiver operating characteristic curve was 0.991. Errors were numerically balanced at 38 false positives and 38 false negatives, corresponding to false-positive and false-negative rates of 3.81 % and 3.84 % and an overall error rate of 3.82 %. The spatial backbone requires approximately 0.45 GFLOPs per frame, giving an estimated 20.25 GFLOPs per video for the spatial path at 45 frames.

These findings are limited to the reported evaluation setting. The study reports a single training configuration evaluated once. It includes no component-wise ablation, so the independent contribution of the multi-scale path, the temporal attention and the Transformer stage is not established; no repeated-seed trials, so stability across initialisations is unknown; and no statistical hypothesis test, so no claim of statistical significance is made. No baseline was retrained under a matched protocol, and the earlier result of 92.45 % on this benchmark is cited as a literature-based reference point rather than a reproduction. Most consequentially, the pipeline processes whole frames without face localisation, so it cannot be established that the observed separation is driven by facial manipulation evidence rather than by scene-level correlates of the label.

Future work should therefore begin with a shortcut-learning control and a component-wise ablation with repeated seeds, followed by matched baseline reproduction, cross-dataset validation and end-to-

[OFFICIAL]

end efficiency profiling. Extending the video-level decision to per-face localisation would address the multi-person condition that motivates the FFIW-10K benchmark.

Ethical Statement

This study involved secondary analysis of a publicly released benchmark dataset and did not involve the collection of new data from human participants; ethical approval was therefore not sought. The FFIW-10K dataset [13] was obtained and used under its published terms of use, and no dataset content is redistributed beyond the single frame reproduced in Figure 2 for the purpose of illustrating the preprocessing pipeline.

The architecture described here is intended to support media-authentication research. Its output is a probability and should not be treated as dispositive forensic evidence. As set out in Section 5.2, this study does not demonstrate that the classifier responds to manipulation artefacts rather than to dataset-specific correlates, does not establish generalisation beyond the evaluated benchmark, and provides no evidence of performance on manipulation methods absent from the training distribution. A negative output does not establish authenticity.

Author Contributions:

Conceptualization, S.J., F.Z.; data curation, S.J., F.Z. and S.S.; formal analysis, S.J., F.Z., S.S., S.R., M.M.A., and M.M.S.; investigation, F.Z., S.S., and S.R.; methodology, S.J., F.Z., and M.M.A; project administration, F.Z., S.R., M.M.A., and M.M.S.; visualization, S.J. and F.Z.; writing—original draft, S.J., F.Z., and S.S.; writing—review and editing, S.R., M.M.A., and M.M.S.;

All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement:

The dataset used in this study, Face Forensics in the Wild (FFIW-10K), is a publicly documented third-party benchmark maintained by its original developers. The dataset is available to researchers upon request and approval by the dataset owners, subject to the applicable Terms of Use. Interested researchers can request access through the official GitHub repository (https://github.com/tfzhou/FFIW) by completing the required Terms of Use form and submitting it to the dataset custodians at ffiwdataset@gmail.com. Upon approval, the dataset developers provide access to the dataset. The authors obtained the dataset through this official procedure. As the dataset is owned by a third party, redistribution of the original data is subject to the dataset owners' permission.

Conflicts of Interest: The authors declare no conflicts of interest.

[OFFICIAL]

References

[1] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," *ACM Computing Surveys*, vol. 54, no. 1, art. 7, pp. 1–41, 2021, doi: 10.1145/3425780.

[2] 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, doi: 10.1016/j.inffus.2020.06.014.

[3] S. Sharma and A. Selwal, "Improved deepfake detection with optimized preprocessing for low-quality images," *Procedia Computer Science*, vol. 258, pp. 507–516, 2025.

[4] D. Guera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *Proc. 15th IEEE Int. Conf. Advanced Video and Signal Based Surveillance (AVSS)*, 2018, pp. 1–6, doi: 10.1109/AVSS.2018.8639163.

[5] S. AlMuhaideb, H. Alshaya, L. Almutairi, D. Alomran, and S. T. Alhamed, "LightFakeDetect: A lightweight model for deepfake detection in videos that focuses on facial regions," *Mathematics*, vol. 13, no. 19, art. 3088, 2025, doi: 10.3390/math13193088.

[6] P. Ghosh, R. Paul, S. Koley, P. Kumar, D. Sarkar, R. Chowdhury, and T. Goto, "DEVision: A lightweight model for deepfake video detection," SSRN preprint 5472931, 2025. [Online]. Available: https://ssrn.com/abstract=5472931

[7] S. Maheswari, V. D. Kumar, D. A. Kumar, R. Jahnavi, and C. Laharee, "Real-time deepfake detection using a hybrid MobileNet-LSTM model for enhanced media integrity," in *Proc. Int. Conf. Intelligent Systems and Digital Transformation (ICISD 2025)*, Chennai, India, 5–6 May 2025, Atlantis Press, pp. 980–991, doi: 10.2991/978-94-6463-866-0_79.

[8] V. S. Anandhasivam, A. K. Anusri, M. Logeshwar, and R. Gopinath, "Enhancing deepfake detection through hybrid MobileNet-LSTM model with real-time image and video analysis," in *Proc. 4th Int. Conf. Ubiquitous Computing and Intelligent Information Systems (ICUIS)*, IEEE, 2024, pp. 1989–1995, doi: 10.1109/ICUIS64676.2024.10867159.

[9] I. L. O. Bastos, A. R. Rocha, and W. R. Schwartz, "VideoFakeD: Video-based deepfake detection," in *Proc. 38th SIBGRAPI Conf. Graphics, Patterns and Images*, 2025, pp. 164–169.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017, pp. 5998–6008.

[11] H. Liu, C. Gong, M. He, J. Li, K. Han, and S. Luo, "When deepfake detection meets graph neural network: A unified and lightweight framework," in *Proc. 32nd ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD 2026)*, vol. 1, ACM, 2026, doi: 10.1145/3770854.3780219.

[OFFICIAL]

[12] M. Amen and M. L. Ranam, "Lightweight deepfake detection on mobile devices using attention-enhanced MobileNet and frequency domain analysis," *Journal of Technology Informatics and Engineering*, vol. 4, no. 1, pp. 95–114, 2025.

[13] T. Zhou, W. Wang, Z. Liang, and J. Shen, "Face Forensics in the Wild," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 5778–5788.

[14] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2117–2125, doi: 10.1109/CVPR.2017.106.

[15] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141, doi: 10.1109/CVPR.2018.00745.

[16] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. European Conf. Computer Vision (ECCV)*, LNCS vol. 11211, Springer, 2018, pp. 3–19, doi: 10.1007/978-3-030-01234-2_1.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[18] F. Zafar, T. A. Khan, S. Akbar, M. T. Ubaid, S. Javaid, and K. A. Kadir, "A hybrid deep learning framework for deepfake detection using temporal and spatial features," *IEEE Access*, vol. 13, pp. 79560–79570, 2025, doi: 10.1109/ACCESS.2025.3566008.

[19] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016, doi: 10.1109/LSP.2016.2603342.

[20] 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, doi: 10.1109/CVPR42600.2020.00327.

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

[22] A. Rossler, 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, doi: 10.1109/ICCV.2019.00009.

[23] P. Aryan, R. M. R. Yanamala, A. Pallakonda, R. D. A. Raj, and K. K. Prakasha, "Lightweight end-to-end patch-based self-attention network for robust image forgery detection," *IEEE Access*, vol. 13, 2025, doi: 10.1109/ACCESS.2025.3607206.

[OFFICIAL]

[24] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2018, arXiv:1710.09412.

[25] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI Conf. Artificial Intelligence*, vol. 34, no. 7, 2020, pp. 13001–13008, doi: 10.1609/aaai.v34i07.7000.

[26] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Machine Learning (ICML)*, PMLR vol. 97, 2019, pp. 6105–6114.

[27] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997, doi: 10.1109/78.650093.

[28] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2015, arXiv:1409.0473.

[29] B. T. Polyak and A. B. Juditsky, "Acceleration of stochastic approximation by averaging," *SIAM Journal on Control and Optimization*, vol. 30, no. 4, pp. 838–855, 1992, doi: 10.1137/0330046.

[30] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, and H. Wu, "Mixed precision training," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2018, arXiv:1710.03740.

[31] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2019, arXiv:1711.05101.

[32] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2017, arXiv:1608.03983.

[33] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019, pp. 8024–8035.

[34] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta (BBA) Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975, doi: 10.1016/0005-2795(75)90109-9.

[35] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006, doi: 10.1016/j.patrec.2005.10.010.

[36] E. B. Wilson, "Probable inference, the law of succession, and statistical inference," *Journal of the American Statistical Association*, vol. 22, no. 158, pp. 209–212, 1927, doi: 10.1080/01621459.1927.10502953.

[OFFICIAL]