

Article

Regularized Multi-Backbone Ensembles for Video-Level Deepfake Detection on Celeb-DF v2: Accuracy–Efficiency and Generalization Limits

Mohammed Alshalfi ¹, Abdulrahman Alshehri ¹, Qazi Emad Ul Haq ^{1,2}  and Tariq M. Khan ^{1,2,*} 

¹ Center of Artificial Intelligence for Security, Naif Arab University for Security Sciences, Riyadh 14812, Saudi Arabia; 202400344@student.nauss.edu.sa (M.A.); 202400070@student.nauss.edu.sa (A.A.); qabdulrab@nauss.edu.sa (Q.E.U.H.)

² Department of Cybersecurity and Digital Forensics, College of Forensic & Investigative Sciences, Naif Arab University for Security Sciences, Riyadh 14812, Saudi Arabia

* Correspondence: tariq045@gmail.com

Abstract

The increasing realism of deepfake videos has intensified the need for reliable video-level detection systems, but benchmark performance must be interpreted together with generalization, temporal-modeling, and efficiency limits. This study presents a reproducible multi-backbone framework for detecting manipulated videos on the official Celeb-DF v2 benchmark. Five pretrained image-classification architectures—ResNet-50, EfficientNet-B4, ConvNeXt-Small, ViT-Base, and Swin-Base—are fine-tuned under a unified protocol using uniform frame sampling, class-balanced training, RandAugment, MixUp, CutMix, random erasing, label smoothing, AdamW optimization, cosine learning-rate scheduling, and exponential moving average weights. During inference, frame-level fake probabilities are stabilized using horizontal-flip test-time augmentation and aggregated into video-level predictions by mean probability pooling. This aggregation is a fixed probability-pooling rule rather than an explicit temporal model. A probability-level ensemble of ResNet-50, ConvNeXt-Small, and Swin-Base combines convolutional and attention-based representations. On the official 518-video Celeb-DF v2 test set, the top-three ensemble achieves a video-level AUC of 99.967% and an average precision of 99.983%, while ResNet-50 provides the strongest single-model accuracy–efficiency trade-off. Additional analyses examine frame-to-video aggregation, ROC and precision–recall behavior, probability distributions, frame-probability stability over sampled frames, a proof-of-concept temporal-splice sensitivity test, and inference efficiency. The results demonstrate highly competitive in-dataset performance on Celeb-DF v2 only. Because no external benchmark testing, learned temporal baseline, confidence-gated cascade, or broad partial-manipulation benchmark is included, the results should not be interpreted as evidence of cross-dataset robustness, in-the-wild deployment readiness, learned temporal reasoning, or general localization capability. Cross-dataset evaluation on FaceForensics++, DFDC, WildDeepfake, and related benchmarks, probability calibration, false-positive control, explicit temporal modeling, cascade-based inference, and expanded localization evaluation are identified as priority future-work directions.



Academic Editor: Amit Kumar Mishra

Received: 3 June 2026

Revised: 15 July 2026

Accepted: 17 July 2026

Published: 21 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\)](#) license.

Keywords: deepfake video detection; video-level classification; multi-backbone ensemble; Celeb-DF v2; vision transformer

1. Introduction

The rapid expansion of digital communication has made images and videos central to news reporting, social interaction, criminal investigation, corporate communication, education, and identity verification. However, the same digital ecosystem has also become increasingly vulnerable to synthetic and manipulated media. Deepfakes are AI-generated or AI-manipulated media in which a person's face, expression, voice, or identity-related visual appearance is altered in a way that can appear realistic to both human observers and automated verification systems. Their social risk is significant because digital media is no longer limited to entertainment; they are also used as evidence, authentication material, political communication, and reputational information. As a result, manipulated media can affect public communication, personal identity, institutional trust, and the credibility of digital evidence. These concerns show that deepfake detection is not just a computer vision task but also a broader issue of digital trust, media authentication, and security.

The technical difficulty of detecting deepfakes has increased in tandem with the advancement of generative models. Visible artifacts, including color disparity, low-quality blending, unnatural winking, discontinuous head pose, and facial boundary defects, were frequently generated by early face manipulation pipelines. As a result, early forensic systems concentrated on handcrafted or semantically interpretable signals, such as irregular eye-blinking [1], inconsistent head pose [2], optical-flow aberrations [3], and visible face-manipulation artifacts [4]. Despite the significant early progress that these methods made, their assumptions have become less dependable as deepfake generation methods have improved in spatial resolution, color correction, temporal smoothness, and post-processing. Modern deepfakes may be compressed or resized by online platforms, may contain fewer apparent anomalies, and may maintain plausible facial behavior over brief video segments. As a result, recent research on deepfake detection has transitioned from the use of hand-crafted forensic rules to learning-based methods that capture subtle spatial, frequency, temporal, and multimodal evidence [5]. Nevertheless, high performance on one benchmark does not by itself establish real-world authentication reliability, because detectors can overfit to dataset-specific compression, preprocessing, face-alignment, and identity-distribution cues.

As a result, deep learning has become the primary method for identifying manipulated facial media. Convolutional neural networks (CNNs) have the capacity to acquire local spatial features that are associated with blending boundaries, texture discontinuities, sensor-noise inconsistencies, and compression-sensitive anomalies. As image-level detectors or feature extractors for video-level classification, architectures such as MesoNet [6], Xception [7], ResNet [8] and EfficientNet [9] have been widely used. FaceForensics++, DeepFakeTIMIT, and Celeb-DF v2 are among the benchmark datasets on which CNN-based methods have exhibited exceptional performance [10–12]. In contrast, numerous detectors continue to exhibit inadequate generalization when the training and testing distributions diverge, as they may overfit to manipulation traces, compression statistics, camera pipelines, or face-alignment anomalies that are unique to the dataset.

This generalization problem has motivated research on alternative forensic cues and stronger model architectures. Frequency-domain methods attempt to detect spectral traces that may not be visible in the spatial domain [13–17]. Biological and physiological approaches examine cues such as pulse-related signals and facial behavior [18]. Lip movement and face expression methods focus on mouth dynamics, visual speech consistency, and self-supervised facial motion representations [19,20]. Self-blended and consistency-based methods generate or exploit generic manipulation cues to improve robustness to unseen forgeries [21–23]. Multimodal approaches also exploit audio-visual inconsistencies, al-

though they require reliable audio and can be affected by dubbing, noise, compression, or missing modalities [24,25].

More recently, attention- and transformer-based architectures have been introduced into computer vision and deepfake detection because attention can model long-range relationships across image regions [26]. The vision transformer (ViT) introduced image classification using non-overlapping patch tokens and global self-attention [27], while the Swin Transformer reduced the computational cost of attention through hierarchical shifted windows [28]. Deepfake-specific attention and temporal-transformer models also show the value of attending to manipulated facial regions and modeling frame-to-frame inconsistency [29,30]. At the same time, transformers are often data-hungry and may not always outperform CNNs on constrained forensic benchmarks. Modern convolutional architectures such as ConvNeXt revisit CNN design by using transformer-inspired components while retaining useful convolutional inductive bias [31]. This makes it important to compare classical CNNs, efficient CNNs, modern CNNs, vision transformers, and hierarchical transformers under the same dataset split, sampling protocol, optimization strategy, and video-level evaluation procedure.

Benchmark design is also critical for fair deepfake detection research. Many reported results are difficult to compare because they depend on the dataset, split strategy, face-cropping pipeline, frame sampling rate, compression setting, training data, and evaluation level. FaceForensics++ is widely used because it contains multiple manipulation methods and compression levels [11]. The Deepfake Detection Challenge (DFDC) dataset increased scale and diversity and encouraged stronger training and ensembling strategies [32]. DeeperForensics-1.0 introduced large-scale real-world perturbation analysis [33], while WildDeepfake emphasized the difficulty of in-the-wild manipulated videos [34]. Celeb-DF v2 was proposed as a more challenging benchmark with higher-quality face-swap videos and an official test split designed to reduce identity leakage [12]. Celeb-DF v2 is therefore employed in this study because it provides a recognized benchmark for high-quality face-forgery detection and supports fair video-level evaluation through its official test list.

Although there have been recent advances, several research gaps remain. First, many studies report results using only one backbone or one model family, which makes it difficult to determine whether improvements are caused by the choice of architecture, training strategy, or evaluation protocol. Second, frame-level reporting may be misleading because adjacent frames from the same video are highly correlated; therefore, video-level decisions are necessary for practical deployment. Third, high AUC values can obscure threshold-dependent behavior, such as false positives on real videos. Consequently, AUC, average precision, accuracy, recall, precision, F1-score, confusion matrices, probability distributions, and efficiency measurements should all be used to interpret detector behavior. Fourth, the computational cost of deploying highly accurate detectors may be prohibitive, particularly when multi-model ensembles or large transformer backbones are used. Fifth, deepfake datasets are often imbalanced, so training protocols must address data leakage, overconfidence, and class imbalance.

This article introduces a controlled and reproducible video-level deepfake detection framework for the Celeb-DF v2 benchmark. The proposed framework assesses five representative pretrained backbones: ResNet-50, EfficientNet-B4, ConvNeXt-Small, ViT-Base, and Swin-Base. These models are trained and evaluated within a unified Celeb-DF v2 pipeline. The pipeline uses uniform frame sampling, consistent preprocessing, class-balanced sampling, MixUp and CutMix regularization, label smoothing, AdamW optimization, cosine learning-rate scheduling, exponential moving average of weights, horizontal-flip test-time augmentation, and video-level probability aggregation. These

choices are motivated by established training and regularization techniques, including label smoothing, RandAugment, MixUp, AdamW, and Random Erasing [35–39].

Before preprocessing the frames to a shared spatial resolution, the framework samples a fixed number of frames from each video. Each sampled frame is classified as either genuine or manipulated, and the resulting frame-level fake probabilities are aggregated into video-level probabilities. Probability-level ensembling is then used to combine complementary models. In this work, video-level aggregation means fixed averaging of independently predicted frame probabilities; it is not a learned temporal model and does not explicitly encode motion dynamics or frame-to-frame dependencies. Previous work using recurrent video models, temporal coherence modeling, thumbnail-based temporal layouts, and ensemble-based systems supports the broader value of aggregating evidence across frames and models for more reliable deepfake detection [40–43]. Recent strong systems also show that performance can improve by uncovering complementary forgery clues or using temporally aware training strategies [44,45].

This article offers diagnostic analyses that provide a more comprehensive understanding of the model's behavior than the standard Celeb-DF v2 evaluation. These analyses encompass confidence-stratified frame examples, frame-probability stability across sampled video frames, a limited proof-of-concept temporal-splice sensitivity scenario, inference latency, throughput, and comparisons with the related literature. In addition to high accuracy, practical deepfake detection requires interpretable probability behavior, stable video-level decisions, and reasonable computational efficiency. Consequently, this broader diagnostic evaluation is reported together with explicit limitations on generalization and deployment claims.

The following are the primary contributions of this article. First, it introduces a unified video-level deepfake detection framework that evaluates CNN and transformer backbones under the same Celeb-DF v2 protocol. Second, it provides a controlled comparison of five representative architectures, allowing the influence of architecture choice to be analyzed without changing preprocessing, optimization, or inference settings. Third, it studies a top-three probability ensemble that combines ResNet-50, ConvNeXt-Small, and Swin-Base for threshold-independent video-level ranking, while explicitly reporting the associated computational overhead. Fourth, it conducts a diagnostic analysis that goes beyond a single accuracy value by reporting frame-level metrics, video-level metrics, aggregation effects, confusion matrices, ROC and precision–recall curves, test-time augmentation, frame-probability stability, a proof-of-concept temporal-splice sensitivity test, and inference efficiency. All contributions are demonstrated on the official Celeb-DF v2 benchmark under a single-dataset, in-dataset evaluation protocol; transfer to other datasets, manipulation methods, compression settings, or in-the-wild content is not established by the present study and is addressed as future work in Section 5.8.

Scope of the Present Study

This study should be read as a controlled Celeb-DF v2 architecture and aggregation analysis rather than as a complete real-world authentication system. Four boundaries are therefore made explicit. First, no cross-dataset testing is performed in the present study; consequently, the near-ceiling Celeb-DF v2 AUC values do not establish robustness to FaceForensics++, DFDC, WildDeepfake, DeeperForensics-1.0, social-media videos, or unseen manipulation pipelines. Second, the video-level score is produced by mean pooling independently predicted frame probabilities; it is not a CNN-LSTM, 3D CNN, temporal-transformer, or other learned temporal model. Third, the reported top-three ensemble is an offline maximum-ranking configuration, not a deployment-optimized cascade; a ResNet-50-first confidence-gated routing strategy is proposed only as future work. Fourth,

the temporal-splice experiment is a limited diagnostic example based on a constructed source-video pair and should not be interpreted as a general temporal or spatial localization benchmark. These boundaries are reflected throughout the Results, Discussion, Limitations, and Future Work Sections to avoid overextending the claims beyond the conducted experiments.

The remainder of the article is organized as follows. The literature on deepfake detection is comprehensively reviewed in Section 2. The proposed methodology is explained in Section 3, which encompasses dataset preparation, leakage prevention, frame sampling, preprocessing, backbone architectures, training strategy, video-level inference, ensembling, and evaluation metrics. In Section 4, the statistical and qualitative findings of Celeb-DF v2 are presented. Section 5 addresses future work, limitations, and implications. The article concludes with Section 6.

2. Related Work

Deepfake detection has evolved from early artifact-based forensic methods to modern learning-based systems that utilize CNNs, transformers, frequency-domain features, temporal modeling, multimodal learning, and ensemble strategies. Early methods were predicated on the notion that face-swap and synthesis pipelines introduce visible or measurable irregularities. Li et al. employed eye-blink patterns to identify phony videos, as the initial-generated face videos frequently failed to replicate natural blink behavior [1]. Yang et al. examined inconsistent head poses, Amerini et al. employed optical-flow analysis to identify anomalous motion across consecutive frames, and Matern et al. concentrated on visual artifacts, including color mismatch, missing reflections, boundary defects, and local blending errors [2–4]. These methods yielded interpretable evidence; however, they are contingent upon the ongoing presence of particular artifacts and may become less dependable as the manipulation's quality improves.

Other studies examined biological, physiological, behavioral, and identity-related cues. FakeCatcher used biological signals derived from subtle skin-color variations caused by blood flow [18]. LipForensics focused on the dynamics of the mouth-region and visual speech representations for the detection of facial counterfeits, while RealForensics used self-supervision of real talking faces to improve robustness [19,20]. Such methods are valuable because they use interpretable signals beyond simple texture artifacts. However, they can depend on stable face tracking, sufficient video quality, reliable illumination, and the availability of the relevant facial or speech cues.

CNN-based approaches replaced manually designed cues with learned spatial representations. MesoNet was one of the first compact CNN models designed to capture mesoscopic facial forgery features [6]. Xception later became a widely used deepfake detection baseline after its adoption in FaceForensics++ benchmarking [7,11]. ResNet introduced residual learning and skip connections, enabling deep hierarchical feature extraction, while EfficientNet introduced compound scaling and became influential due to its high precision-to-parameter efficiency [8,9]. CNNs remain attractive because they are efficient, benefit from large-scale image pretraining, and provide useful local inductive bias for detecting spatial artifacts. However, CNN-based detectors can overfit to dataset-specific traces, compression statistics, camera pipelines, or face-alignment artifacts, limiting cross-dataset generalization [5].

Several spatial methods have been proposed to improve generalization beyond standard CNN classification. Face X-Ray models blending boundaries to detect face forgery in a more general way [21]. Learning self-consistency encourages detectors to identify inconsistencies within manipulated facial regions [23]. Modern CNNs such as ConvNeXt revisit convolutional design by incorporating transformer-inspired components, including

larger depthwise kernels, LayerNorm, GELU activations, and inverted bottlenecks while preserving convolutional locality [31]. This makes ConvNeXt a relevant backbone for comparing modern convolutional representations with both classical CNNs and transformer-based models.

Attention-based methods were introduced to improve global reasoning and the localization of manipulated regions. The general attention mechanism was popularized by transformer architectures [26]. Multi-attention approaches use attention maps to focus on discriminative facial regions and reduce reliance on irrelevant background information [29]. More generally, attention mechanisms are useful because manipulation evidence may be distributed across different parts of the face rather than being confined to a single local region. However, attention-based models can be computationally expensive and may require careful training to avoid overfitting on limited forensic datasets.

Transformer architectures have also influenced deepfake detection research. The vision transformer processes images as patch sequences using global self-attention, allowing long-range relationships between image regions to be modeled [27]. The Swin Transformer reduces the computational cost of attention through hierarchical shifted windows and preserves stronger locality than plain global self-attention [28]. Interpretable spatial-temporal transformer methods further extend attention to video-level modeling by capturing both spatial and temporal evidence [30]. These architectures motivate the inclusion of transformer-based backbones in deepfake detection, where both local artifacts and broader facial-region relationships may be relevant. At the same time, transformer models can be data-hungry and may not always outperform CNNs on constrained forensic benchmarks.

Temporal modeling is another important direction because deepfake videos contain frame-to-frame information that cannot be fully captured by isolated image classifiers. CNN-LSTM and recurrent video models attempt to capture temporal inconsistencies between frames rather than relying only on single-frame evidence [40]. Temporal coherence methods explicitly model frame-to-frame consistency for more general video face-forgery detection [41]. TALL converts temporal information into thumbnail layouts to capture long-range video evidence efficiently [42], while AltFreezing aims to improve generalization by alternating spatial and temporal training behavior [45]. Temporal reasoning is useful because manipulated videos can contain unstable facial motion, inconsistent expressions, or frame-level artifacts that become clearer when predictions are aggregated over time. Nevertheless, temporal models can be sensitive to clip length, frame rate, face-tracking quality, and temporal smoothing.

Frequency-domain approaches address a different weakness of spatial classifiers. These methods are based on the observation that generative models and face-blending pipelines may leave spectral artifacts that are not obvious in RGB space. Durall et al. showed that synthetic images can exhibit abnormal frequency distributions, motivating the use of frequency information to detect generated or manipulated visual content [13]. F3-Net mines frequency-aware forgery clues to improve face-forgery detection [14]. High-frequency feature methods and bilateral high-pass filtering further demonstrate that frequency-sensitive cues can improve robustness under distribution changes [15,16]. More recent work such as FreqBlender uses frequency knowledge to strengthen deepfake detection performance [17]. Frequency-based cues can complement spatial features, but may be weakened by compression, resizing, denoising, and platform transcoding.

To improve robustness, self-blending and augmentation-based training strategies have also been proposed. Self-Blended Images generate pseudo-fake examples from real images to teach detectors more generic blending artifacts [22]. Standard augmentation and regularization methods, including RandAugment, MixUp, Random Erasing, and label smoothing, can further reduce overfitting and improve training stability [35–37,39].

AdamW is also widely used to stabilize training with decoupled weight decay [38]. These strategies are useful because deepfake detectors often suffer from dataset bias and may otherwise learn manipulation-specific or dataset-specific shortcuts.

Multimodal and ensemble methods provide additional ways to improve robustness. Multimodal approaches exploit audio-visual inconsistencies or affective cues, but they require reliable audio and can be affected by dubbing, noise, compression, or missing modalities [24,25]. Ensemble learning is a practical alternative because models with different architectures can capture complementary evidence and reduce error variance. Large-scale benchmark efforts such as the Deepfake Detection Challenge encouraged the use of stronger training strategies and model ensembling for deepfake detection [32]. DeepfakeStack also showed the value of ensemble-based learning in combining complementary detection evidence [43].

Benchmark datasets and evaluation protocols strongly influence the reported performance. Early datasets such as DeepFakeTIMIT supported initial evaluation, but were limited in scale and realism [10]. FaceForensics++ became a standard benchmark because it includes multiple manipulation methods and compression settings [11]. The Deepfake Detection Challenge dataset increased scale and diversity [32]. DeeperForensics-1.0 introduced real-world perturbation analysis [33], and WildDeepfake emphasized the difficulty of detecting videos manipulated in-the-wild [34]. Celeb-DF v2 is especially important because it contains higher-quality face-swap videos with improved synthesis quality, color correction, and temporal smoothness, as well as an official test list designed to reduce leakage between training and evaluation [12]. These datasets show that benchmark choice, preprocessing, split design, compression level, and evaluation protocol strongly affect the interpretation of reported results.

Recent high-performance methods further show the importance of combining various cues. Exposing the Deception aims to uncover additional forgery clues for stronger deepfake detection [44]. Frequency-aware systems such as F3-Net and FreqBlender emphasize spectral evidence [14,17]. Temporal systems such as ISTVT, TALL, and AltFreezing emphasize video-level structure and frame-to-frame consistency [30,42,45]. These developments indicate that robust detection increasingly depends on the integration of complementary spatial, temporal, frequency, and model-level evidence rather than relying on a single artifact type.

The present study is at the intersection of CNN-based detection, transformer-based detection, video-level aggregation, and ensemble learning. Unlike methods that rely on a single artifact type or a single model family, it compares ResNet-50, EfficientNet-B4, ConvNeXt-Small, ViT-Base, and Swin-Base under the same Celeb-DF v2 protocol. It also uses video-level probability aggregation and a top-three probability ensemble to improve ranking reliability while reporting frame-level metrics, video-level metrics, aggregation effects, confusion matrices, probability distributions, frame-probability stability, proof-of-concept temporal-splice sensitivity, and inference efficiency. This design is intended to provide a controlled evaluation of architecture choice, aggregation behavior, and deployment-relevant trade-offs for video-level deepfake detection.

3. Proposed Methodology

The proposed framework is developed for video-level deepfake detection on Celeb-DF v2. The goal is to determine whether an input video is genuine or manipulated while keeping all evaluated backbones under the same training and inference protocol. Unlike image-level detection, the proposed evaluation treats the video as the primary decision unit rather than treating every frame as an independent test sample. Frames extracted from the same video are temporally correlated and often share the same identity, background,

illumination, compression characteristics, and manipulation artifacts. Therefore, reporting only frame-level performance may overstate practical detection ability. The framework first estimates frame-level fake probabilities and then aggregates them into a single video-level probability. This aggregation is a fixed probability-pooling operation, not a learned temporal-reasoning module.

Let the complete video dataset be denoted by

$$\mathcal{D} = \{(v_i, y_i)\}_{i=1}^N, \quad (1)$$

where v_i represents the i -th video, $y_i \in \{0, 1\}$ is its binary label, $y_i = 0$ denotes an authentic video, $y_i = 1$ denotes a fake video, and N is the total number of videos. The goal of the detector is to learn a mapping function

$$\Phi : v_i \rightarrow \hat{p}_i, \quad (2)$$

where $\hat{p}_i \in [0, 1]$ is the estimated probability that video v_i belongs to the fake class. The final predicted label is obtained by applying a decision threshold τ :

$$\hat{y}_i = \begin{cases} 1, & \hat{p}_i \geq \tau, \\ 0, & \hat{p}_i < \tau. \end{cases} \quad (3)$$

In this study, $\tau = 0.5$ is used as the default decision threshold unless otherwise stated.

The overall methodology is illustrated in Figure 1. The framework consists of four main stages: (i) video-level data partitioning and leakage prevention; (ii) uniform frame sampling and preprocessing; (iii) frame-level classification using multiple pretrained backbone architectures; and (iv) video-level probability aggregation followed by ensemble prediction. The end-to-end operational pipeline is shown in Figure 2.

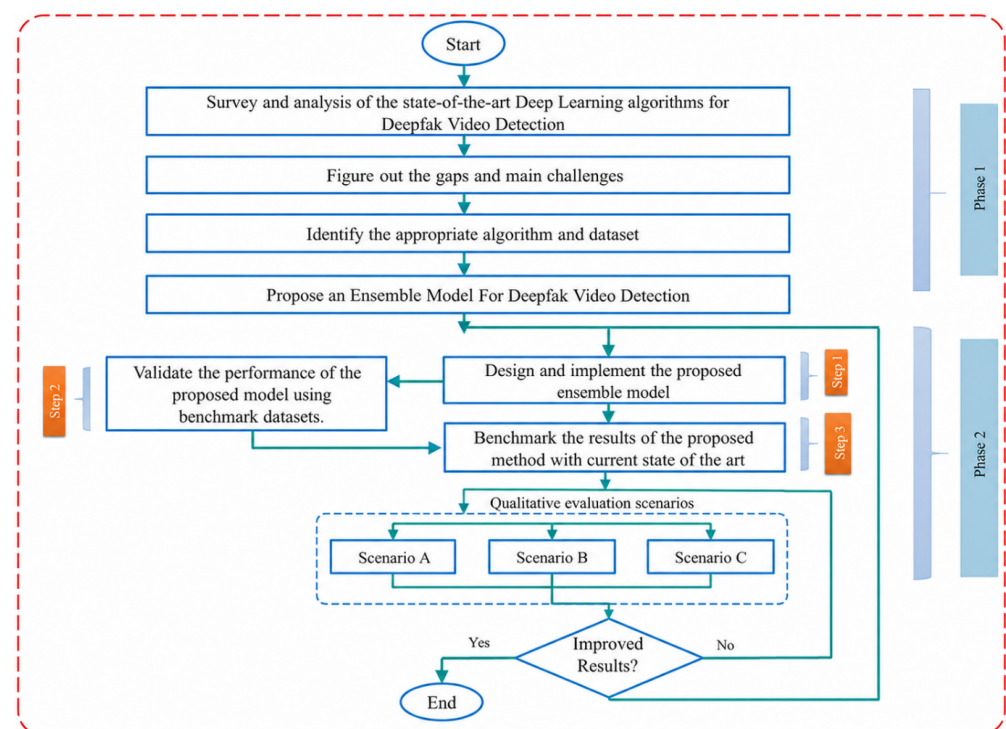


Figure 1. Proposed research methodology for video-level deepfake detection. The diagram summarizes the controlled Celeb-DF v2 protocol, including video-level partitioning, frame sampling, backbone training, and evaluation.

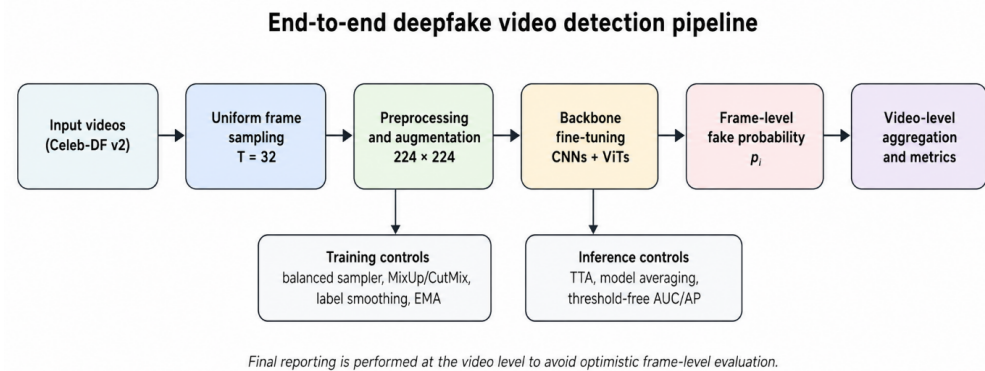


Figure 2. End-to-end deepfake video detection pipeline, including uniform frame sampling, preprocessing, CNN/transformer backbone fine-tuning, frame-level fake probability estimation, test-time augmentation, video-level mean pooling, and optional top-3 probability ensembling.

3.1. Video-Level Detection Formulation

Each input video v_i is first represented as an ordered sequence of decoded frames:

$$v_i = \{r_{i,1}, r_{i,2}, \dots, r_{i,N_i}\}, \quad (4)$$

where $r_{i,j}$ denotes the j -th raw frame of video v_i , and N_i is the total number of frames in that video. Since videos may have different durations and frame rates, a fixed number of frames is sampled from each video to produce a uniform input representation. Let T denote the number of frames sampled per video. The set of sampled frames for video v_i is defined as

$$\mathcal{F}_i = \{x_{i,1}, x_{i,2}, \dots, x_{i,T}\}, \quad (5)$$

where $x_{i,t} \in \mathbb{R}^{3 \times H \times W}$ is the preprocessed RGB frame t -th, and $H = W = 224$ in the proposed implementation.

The detection task is therefore transformed from direct video classification into frame-level probability estimation followed by video-level aggregation:

$$v_i \rightarrow \mathcal{F}_i \rightarrow \{p_{i,1}, p_{i,2}, \dots, p_{i,T}\} \rightarrow \hat{p}_i. \quad (6)$$

Here, $p_{i,t}$ denotes the fake probability assigned to the t -th sampled frame of the video v_i , and $\hat{p}_i$ denotes the final fake probability at the video-level.

3.2. Uniform Frame Sampling

To ensure that every video contributes the same number of frames to the model, uniform temporal sampling is used. Uniform sampling avoids biasing the detector towards the beginning, middle, or end of a video. It also ensures that the sampled frames cover the full temporal span of the clip rather than being concentrated in a short segment.

For a video v_i containing N_i decoded frames, the index of the t -th sampled frame is computed as

$$s_{i,t} = \text{round}\left(\frac{t(N_i - 1)}{T - 1}\right), \quad t = 0, 1, \dots, T - 1. \quad (7)$$

The sampled frame is then given by

$$x_{i,t} = \mathcal{P}(r_{i,s_{i,t}}), \quad (8)$$

where $\mathcal{P}(\cdot)$ denotes the preprocessing and normalization operation applied to the raw frame. In this study, $T = 32$, meaning that each video is represented by 32 uniformly

sampled frames. This fixed-length representation allows all backbones to be evaluated under the same input protocol and enables fair comparison across model families.

3.3. Frame Preprocessing and Normalization

Each sampled frame is resized and transformed into a normalized tensor before being passed to the backbone network. Let $x_{i,t}^{raw}$ denote a sampled RGB frame before normalization. The normalized frame $x_{i,t}$ is computed channel-wise as

$$x_{i,t}^{(c,h,w)} = \frac{x_{i,t}^{raw(c,h,w)} - \mu_c}{\sigma_c}, \quad c \in \{R, G, B\}, \quad (9)$$

where μ_c and σ_c denote the ImageNet channel-wise mean and standard deviation, respectively. The values used are

$$\mu = (0.485, 0.456, 0.406), \quad \sigma = (0.229, 0.224, 0.225). \quad (10)$$

During training, stochastic transformations are applied to improve generalization and reduce dependence on low-level artifacts specific to the dataset. These include random resized cropping, horizontal flipping, color jittering, RandAugment, and random erasing. During validation and testing, deterministic preprocessing is used to ensure repeatable evaluation. The same preprocessing protocol is applied to all backbone architectures so that the comparison between models is not affected by inconsistent input transformations.

3.4. Frame-Level Feature Extraction

The proposed framework evaluates multiple pretrained image-classification backbones under the same training and inference protocol. The five selected backbones are intended to represent complementary model families rather than minor variants of a single architecture: ResNet-50 represents a classical residual CNN, EfficientNet-B4 represents compound-scaled efficient CNN design, ConvNeXt-Small represents a modern CNN with transformer-inspired design choices, ViT-Base represents global patch-based self-attention, and Swin-Base represents hierarchical window-based transformer modeling. This selection allows a controlled comparison across classical CNNs, efficient CNNs, modern CNNs, global vision transformers, and hierarchical transformers while keeping data partitioning, augmentation, optimization, frame sampling, and video-level aggregation fixed. Let $g_{\theta_k}(\cdot)$ denote the k -th backbone network, where $k \in \{1, 2, \dots, K\}$, and K is the number of evaluated backbones. For a preprocessed frame $x_{i,t}$, the backbone extracts a feature vector:

$$f_{i,t}^{(k)} = g_{\theta_k}(x_{i,t}), \quad f_{i,t}^{(k)} \in \mathbb{R}^{D_k}, \quad (11)$$

where D_k is the feature dimension of the k -th backbone.

A binary classification head $h_{\psi_k}(\cdot)$ maps the extracted feature vector to two logits corresponding to the real and fake classes:

$$o_{i,t}^{(k)} = h_{\psi_k}(f_{i,t}^{(k)}) = W_k f_{i,t}^{(k)} + b_k, \quad (12)$$

where $W_k \in \mathbb{R}^{2 \times D_k}$ and $b_k \in \mathbb{R}^2$ are the learnable parameters of the classification head.

All parameter counts reported in this article follow a single implementation-level convention. The original pretraining classification layer of each backbone was replaced by the two-output real/fake head in Equation (12), and the reported count includes the complete backbone and this binary head. Counts were obtained from the instantiated two-class models and are rounded to two decimal places in millions. For the top-3 ensemble, the exact unrounded counts of ResNet-50, ConvNeXt-Small, and Swin-Base were summed

before rounding. Consequently, these values differ from nominal architecture counts that include the original 1000-class pretraining heads.

The logits are converted into class probabilities using the softmax function:

$$P_k(y = c \mid x_{i,t}) = \frac{\exp(o_{i,t,c}^{(k)})}{\sum_{j=0}^1 \exp(o_{i,t,j}^{(k)})}, \quad c \in \{0, 1\}. \quad (13)$$

The fake probability predicted by the k -th model for frame $x_{i,t}$ is therefore

$$p_{i,t}^{(k)} = P_k(y = 1 \mid x_{i,t}). \quad (14)$$

This design allows the proposed detector to use strong pretrained spatial representations while keeping the final decision at the video level. The frame classifier learns local and global visual evidence of manipulation, whereas the aggregation stage reduces the influence of noisy or ambiguous individual frames.

3.5. Training Objective

Each backbone is trained as a binary classifier using the frame labels inherited from the corresponding video labels. Since every sampled frame $x_{i,t}$ originates from video v_i , the frame-level label is assigned as

$$y_{i,t} = y_i. \quad (15)$$

For standard training samples, label-smoothed cross-entropy is used. Let $C = 2$ be the number of classes and let ε be the label-smoothing parameter. The smoothed target distribution q_c is defined as

$$q_c = \begin{cases} 1 - \varepsilon, & c = y_{i,t}, \\ \frac{\varepsilon}{C - 1}, & c \neq y_{i,t}. \end{cases} \quad (16)$$

The label-smoothed cross-entropy loss for a frame is

$$\mathcal{L}_{LS} = - \sum_{c=0}^{C-1} q_c \log P_k(y = c \mid x_{i,t}). \quad (17)$$

When MixUp or CutMix is applied, the model receives a mixed sample and a soft target. For two frames x_a and x_b , with labels y_a and y_b , MixUp constructs

$$\tilde{x} = \lambda x_a + (1 - \lambda) x_b, \quad (18)$$

$$\tilde{y} = \lambda y_a + (1 - \lambda) y_b, \quad (19)$$

where

$$\lambda \sim \text{Beta}(\alpha, \alpha). \quad (20)$$

For CutMix, a rectangular region from one image is inserted into another image, and the effective mixing coefficient is adjusted according to the patch area:

$$\lambda = 1 - \frac{A_{\text{patch}}}{A_{\text{image}}}. \quad (21)$$

The soft-target cross-entropy loss is then used:

$$\mathcal{L}_{\text{soft}} = - \sum_{c=0}^{C-1} \tilde{y}_c \log P_k(y = c \mid \tilde{x}). \quad (22)$$

The overall training objective for model k is expressed as

$$\theta_k^*, \psi_k^* = \arg \min_{\theta_k, \psi_k} \frac{1}{|\mathcal{B}|} \sum_{(x,y) \in \mathcal{B}} \mathcal{L}(x, y; \theta_k, \psi_k), \quad (23)$$

where $\mathcal{B}$ denotes a mini-batch and $\mathcal{L}$ is either $\mathcal{L}_{LS}$ or $\mathcal{L}_{soft}$, depending on whether sample mixing is applied.

3.6. Class-Balanced Mini-Batch Sampling

The Celeb-DF v2 training set contains more fake frames than real frames. If standard random sampling is used, the model may become biased toward the majority class. To reduce this effect, a class-balanced weighted sampler is used during mini-batch construction.

Let n_c denote the number of training frames belonging to class c . For a training frame x_j with label y_j , the sampling weight is defined as

$$w_j = \frac{1}{n_{y_j}}. \quad (24)$$

The probability of selecting frame x_j is then

$$P(x_j) = \frac{w_j}{\sum_{m=1}^M w_m}, \quad (25)$$

where M is the total number of training frames. This strategy increases the probability of sampling minority-class examples and produces approximately balanced mini-batches without discarding majority-class samples.

3.7. Video-Level Probability Aggregation

After a backbone predicts fake probabilities for all sampled frames, frame-level predictions are aggregated into a video-level score. For model k , the fake probability at the video-level is computed as the arithmetic mean of the frame probabilities:

$$\hat{p}_i^{(k)} = \frac{1}{T} \sum_{t=1}^T p_{i,t}^{(k)}. \quad (26)$$

Mean aggregation is used because it is simple, parameter-free, reproducible, and does not require training an additional temporal classifier. It also reduces the influence of isolated noisy frames by averaging evidence over the sampled span of the video. Importantly, mean pooling should not be interpreted as explicit temporal modeling: it averages independently computed per-frame probabilities and does not learn motion dynamics, temporal ordering, or frame-to-frame dependencies. References to video-level behavior in this article therefore describe probability aggregation across sampled frames rather than a learned temporal architecture.

The predicted video label for model k is then obtained as

$$\hat{y}_i^{(k)} = \mathbb{I}(\hat{p}_i^{(k)} \geq \tau), \quad (27)$$

where $\mathbb{I}(\cdot)$ is the indicator function and $\tau = 0.5$ is the default threshold.

Although the proposed backbones operate on individual frames, the final decision is explicitly video-level. This distinction is central to the methodology because practical deepfake detection systems are normally required to classify complete videos rather than isolated frames.

3.8. Test-Time Augmentation

To improve the stability of frame-level predictions during inference, horizontal-flip test-time augmentation is applied. For each frame $x_{i,t}$, the model predicts the fake probability for the original frame and for its horizontally flipped version. Let $\mathcal{H}(\cdot)$ denote the horizontal flip operation. The original and flipped probabilities are

$$p_{i,t,orig}^{(k)} = P_k(y = 1 | x_{i,t}), \quad (28)$$

$$p_{i,t,flip}^{(k)} = P_k(y = 1 | \mathcal{H}(x_{i,t})). \quad (29)$$

The test-time-augmented frame probability is computed as

$$p_{i,t,TTA}^{(k)} = \frac{1}{2} (p_{i,t,orig}^{(k)} + p_{i,t,flip}^{(k)}). \quad (30)$$

The corresponding video-level probability with TTA is

$$\hat{p}_{i,TTA}^{(k)} = \frac{1}{T} \sum_{t=1}^T p_{i,t,TTA}^{(k)}. \quad (31)$$

This inference strategy does not introduce additional trainable parameters. It only averages predictions under a simple spatial transformation, thereby reducing sensitivity to left-right orientation and small spatial variations.

3.9. Multi-Backbone Ensemble Prediction

The proposed framework evaluates multiple backbone architectures and combines the strongest complementary models using probability-level ensembling. Let $\mathcal{K}_{ens}$ denote the selected set of ensemble models, and let $K_{ens} = |\mathcal{K}_{ens}|$ be the number of models in the ensemble. For each frame, the ensemble fake probability is computed as

$$p_{i,t}^{ens} = \frac{1}{K_{ens}} \sum_{k \in \mathcal{K}_{ens}} p_{i,t,TTA}^{(k)}. \quad (32)$$

The final video-level ensemble probability is then obtained by averaging over the sampled frames:

$$\hat{p}_i^{ens} = \frac{1}{T} \sum_{t=1}^T p_{i,t}^{ens}. \quad (33)$$

Equivalently, the ensemble prediction can be written as an average of model-level video probabilities:

$$\hat{p}_i^{ens} = \frac{1}{K_{ens}} \sum_{k \in \mathcal{K}_{ens}} \hat{p}_{i,TTA}^{(k)}. \quad (34)$$

The ensemble decision is obtained as

$$\hat{y}_i^{ens} = \mathbb{I}(\hat{p}_i^{ens} \geq \tau). \quad (35)$$

Probability averaging is chosen over stacking or meta-classification because it does not introduce an additional trainable decision layer. This mitigates the risk of overfitting an additional ensemble classifier and keeps the inference protocol transparent and reproducible. The reported top-3 ensemble combines the strongest and most complementary in-dataset backbones: ResNet-50 contributes residual local convolutional features, ConvNeXt-Small contributes modern convolutional features with larger-kernel and transformer-inspired design choices, and Swin-Base contributes hierarchical window-based attention. The en-

semble is therefore intended to combine complementary decision evidence from different representation families. Because this combination increases inference cost, it is evaluated as an offline maximum-ranking configuration rather than as the default choice for all deployment settings. No confidence-gated cascade or exhaustive search over all possible ensemble subsets is implemented in the present study; these deployment-oriented variants are reserved for future work.

3.10. Leakage-Controlled Model Selection

To ensure that the official test set is used only for final evaluation, model selection is performed on the development validation split. Let $\mathcal{V}_{train}$, $\mathcal{V}_{val}$, and $\mathcal{V}_{test}$ denote the training, validation, and official test video sets, respectively. The splits satisfy

$$\mathcal{V}_{train} \cap \mathcal{V}_{val} = \emptyset, \quad \mathcal{V}_{train} \cap \mathcal{V}_{test} = \emptyset, \quad \mathcal{V}_{val} \cap \mathcal{V}_{test} = \emptyset. \quad (36)$$

All frames inherit the split assignment of their source video. Therefore, no frame from a video in the official test set is used during training, validation, checkpoint selection, or hyperparameter tuning.

For each backbone, the checkpoint is selected according to the video-level AUC of validation:

$$e^* = \arg \max_{e \in \{1, \dots, E\}} AUC_{val}(\theta_{k,e}^{EMA}), \quad (37)$$

where E is the total number of training epochs and $\theta_{k,e}^{EMA}$ denotes the exponential moving average parameters of model k at epoch e . The final selected parameters are

$$\theta_k^* = \theta_{k,e^*}^{EMA}. \quad (38)$$

After selecting the checkpoint, the selected model is evaluated once on the official Celeb-DF v2 test set.

3.11. Overall Algorithm

The complete inference procedure of the proposed framework is summarized in Algorithm 1. The algorithm receives an input video, samples a fixed number of frames, applies preprocessing, obtains frame-level probabilities from the selected backbones, applies test-time augmentation, aggregates the probabilities at the video level, and produces the final predicted label.

Algorithm 1 Proposed video-level deepfake detection framework

Require: Input video v_i , sampled frame number T , selected models $\mathcal{K}_{ens}$, threshold τ

Ensure: Video-level fake probability $\hat{p}_i^{ens}$ and predicted label $\hat{y}_i^{ens}$

- 1: Decode video v_i into raw frames $\{r_{i,1}, r_{i,2}, \dots, r_{i,N_i}\}$
- 2: **for** $t = 0$ to $T - 1$ **do**
- 3: Compute sampled frame index:

$$s_{i,t} = \text{round}\left(\frac{t(N_i - 1)}{T - 1}\right)$$

- 4: Preprocess sampled frame:

$$x_{i,t} = \mathcal{P}(r_{i,s_{i,t}})$$

- 5: **end for**
-

Algorithm 1 *Cont.*

```

6: for each model  $k \in \mathcal{K}_{ens}$  do
7:   for each sampled frame  $x_{i,t}$  do
8:     Predict original-frame fake probability:

```

$$p_{i,t,orig}^{(k)} = P_k(y = 1 \mid x_{i,t})$$

```

9:     Predict flipped-frame fake probability:

```

$$p_{i,t,flip}^{(k)} = P_k(y = 1 \mid \mathcal{H}(x_{i,t}))$$

```

10:    Compute TTA probability:

```

$$p_{i,t,TTA}^{(k)} = \frac{1}{2} (p_{i,t,orig}^{(k)} + p_{i,t,flip}^{(k)})$$

```

11:  end for

```

```

12:  Compute model-level video probability:

```

$$\hat{p}_{i,TTA}^{(k)} = \frac{1}{T} \sum_{t=1}^T p_{i,t,TTA}^{(k)}$$

```

13: end for

```

```

14: Compute ensemble video-level fake probability:

```

$$\hat{p}_i^{ens} = \frac{1}{|\mathcal{K}_{ens}|} \sum_{k \in \mathcal{K}_{ens}} \hat{p}_{i,TTA}^{(k)}$$

```

15: Compute final video label:

```

$$\hat{y}_i^{ens} = \mathbb{I}(\hat{p}_i^{ens} \geq \tau)$$

```

16: return  $\hat{p}_i^{ens}, \hat{y}_i^{ens}$ 

```

4. Results

This section reports the experimental performance of the proposed deepfake detection framework on the official Celeb-DF v2 test set. The evaluation is organized into six parts. First, the frame-level and video-level performance of the five individual backbones and the top-three ensemble are reported. Second, the effect of video-level aggregation is analyzed to determine how much performance is gained by aggregating frame predictions across each video. Third, the ensemble configuration is examined through confusion matrices, ROC curves, precision–recall curves, and probability distributions. Fourth, test-time augmentation is evaluated to quantify its contribution to final video-level performance. Fifth, qualitative and diagnostic experiments are presented to examine confidence behavior, frame-probability stability, and proof-of-concept temporal-splice sensitivity. Finally, inference efficiency and comparison with related studies are reported to evaluate the practical relevance of the proposed approach.

All primary metrics are computed at the video level, because the official benchmark unit in Celeb-DF v2 is the video, rather than the individual frame. Frame-level metrics are included only as a diagnostic analysis to evaluate the discriminative strength of individual sampled frames before temporal aggregation. The main performance indicators are accuracy, precision, recall, F1-score, area under the ROC curve (AUC), and average precision (AP). Among these, video-level AUC is treated as the primary metric because it is threshold independent and directly measures the ability of a detector to rank fake videos above real videos across all possible thresholds.

4.1. Per-Backbone Performance

This subsection compares the five evaluated backbone architectures: ResNet-50, EfficientNet-B4, ConvNeXt-Small, ViT-Base and Swin-Base. The comparison is conducted under the same dataset split, frame sampling strategy, input resolution, optimization setting, and video-level aggregation protocol. This controlled setting allows the observed differences to be attributed primarily to architectural behavior rather than to changes in the preprocessing or training protocol.

4.1.1. Frame-Level Performance

Table 1 reports frame-level performance on the 16,576 sampled test frames. At this level, each frame is treated as an independent image and classified using a fixed threshold of 0.5. Although this evaluation is not the main benchmark criterion, it provides useful insight into how much discriminative evidence each architecture can extract from a single frame before video-level aggregation.

Table 1. Frame-level performance of the five backbone classifiers and the top-3 ensemble on the official Celeb-DF v2 test set [12]. Frame-level evaluation treats the 16,576 sampled test frames, obtained from 518×32 video-frame samples, as independent samples. Parameter counts correspond to the instantiated two-class models and are rounded from exact counts. Reported performance values are measured in this work; citations identify the original backbone architectures.

Model	Params (M)	Accuracy (%)	F1 (%)	AUC (%)	AP (%)
ResNet-50 [8]	23.51	94.673	96.077	98.074	97.644
ConvNeXt-Small [31]	49.46	95.150	96.414	99.297	99.436
EfficientNet-B4 [9]	17.55	93.732	95.324	98.337	98.624
Swin-Base [28]	86.75	92.930	94.873	95.586	95.048
ViT-Base [27]	85.80	83.283	88.396	89.359	92.038
Ensemble (top-3, this work)	159.71	94.739	96.137	99.372	99.459

Note: Bold values indicate the best result in each metric column; ties are bolded jointly.

ConvNeXt-Small achieves the strongest individual frame-level performance, with an AUC of 99.297%, an accuracy of 95.150%, and an F1-score of 96.414%. EfficientNet-B4 and ResNet-50 also perform strongly, reaching frame-level AUC values of 98.337% and 98.074%, respectively. These results indicate that convolutional and modern convolutional architectures are highly effective in detecting frame-level artifacts in Celeb-DF v2.

The transformer-based models show a more mixed pattern. Swin-Base achieves a frame-level AUC of 95.586%, while ViT-Base achieves only 89.359%. This implies that single-frame forensic classification in this context is less effective when the manipulation evidence is nuanced, local, and texture-based, as global self-attention alone is used. In contrast, ConvNeXt-Small capitalizes on convolutional locality and integrates contemporary design strategies, including transformer-inspired normalization and large-kernel depthwise convolutions, which may account for its exceptional single-frame performance.

The top-three ensemble achieves the highest frame-level AUC and AP, with values of 99.372% and 99.459%, respectively. On the other hand, its frame-level accuracy is marginally inferior to that of ConvNeXt-Small. Accuracy is indicative of performance at the fixed 0.5 threshold, while AUC evaluates ranking quality across all thresholds. This distinction is crucial. As a result, the ensemble has already enhanced threshold-independent separability at the frame level, despite the fact that ConvNeXt-Small remains the most powerful individual model at the default operating threshold.

4.1.2. Video-Level Performance

Table 2 presents the primary video-level results on the 518 official Celeb-DF v2 test videos. For each video, the fake probabilities of the sampled frames are averaged to

obtain a single video-level score. Horizontal-flip test-time augmentation is applied before aggregation. A video is classified as fake if its final probability is greater than or equal to 0.5.

Table 2. Video-level performance of the five backbone classifiers and the top-three ensemble on the official Celeb-DF v2 test set [12]. Video-level AUC is used as the primary metric. Reported values are measured in this work; citations identify the original backbone architectures.

Model	Acc (%)	Prec (%)	Rec (%)	F1 (%)	AUC (%)	AP (%)	Δ AUC (pp)
ResNet-50 [8]	94.981	92.896	100.000	96.317	99.937	99.967	+1.863
ConvNeXt-Small [31]	96.139	94.444	100.000	97.143	99.845	99.918	+0.548
EfficientNet-B4 [9]	95.753	94.167	99.706	96.857	99.430	99.698	+1.093
Swin-Base [28]	93.436	90.909	100.000	95.238	99.868	99.930	+4.282
ViT-Base [27]	84.556	81.250	99.412	89.418	93.923	95.765	+4.564
Ensemble (top-3, this work)	95.174	93.151	100.000	96.454	99.967	99.983	+0.595

Note: Bold values indicate the best result in each metric column; ties are bolded jointly.

The video-level results show a substantial improvement over frame-level evaluation for most architectures. ResNet-50 achieves the highest AUC at the individual video-level of 99.937%, followed by Swin-Base with 99.868% and ConvNeXt-Small with 99.845%. EfficientNet-B4 also performs strongly, with an AUC of 99.430%. ViT-Base remains the weakest model, with a video-level AUC of 93.923%, but still benefits from aggregation compared to its frame-level result.

The top-three ensemble, consisting of ResNet-50, Swin-Base and ConvNeXt-Small, achieves the best overall ranking performance, with a video-level AUC of 99.967% and an AP of 99.983%. This confirms that the ensemble improves the threshold-independent separability between real and fake videos. However, the ensemble does not achieve the highest accuracy or precision at the fixed threshold of 0.5. ConvNeXt-Small achieves the highest video-level accuracy of 96.139% and the highest precision of 94.444%. This result shows that the model with the best ranking performance is not necessarily the model with the best fixed-threshold classification behavior.

A consistent pattern across the results is that the main residual error is the false-positive classification of real videos rather than the false-negative classification of fake videos. ResNet-50, ConvNeXt-Small, Swin-Base, and the ensemble all achieve a 100.000% recall, which means that none of the fake videos is missed at the 0.5 threshold. In contrast, precision varies substantially, from 81.250% for ViT-Base to 94.444% for ConvNeXt-Small. This indicates that the main challenge in this evaluation is not detecting fake videos but reducing the number of real videos incorrectly flagged as fake.

The difference between AUC and accuracy is especially important for interpreting the ensemble. The ensemble achieves the highest AUC, indicating the best overall ranking of real and fake videos across thresholds. However, its accuracy is lower than that of ConvNeXt-Small at the fixed 0.5 threshold. This occurs because probability averaging can slightly increase the predicted fake probability of some borderline real videos, causing a small number of real videos to cross the decision threshold. Therefore, the ensemble is most appropriate when ranking quality or threshold-independent detection is the main objective, whereas ConvNeXt-Small is preferable when the default threshold and false-positive reduction are more important.

4.1.3. Cross-Architecture Ranking

Table 3 ranks all models by video-level AUC and includes the standardized two-class parameter count, pretraining corpus, frame-level AUC, video-level AUC, and the gain obtained from frame-to-video aggregation. The ranking highlights three important findings.

Table 3. Rank comparison of all five backbones and the ensemble in terms of video-level AUC on the Celeb-DF v2 test set. Parameter counts correspond to the instantiated two-class models and are rounded from exact counts. Reported AUC values are measured in this work; citations identify the original backbone architectures.

Rank	Model	Params (M)	Pretraining Corpus	Frame AUC (%)	Video AUC (%)	Δ AUC (pp)
1	Ensemble (top-3, this work)	159.71	N/A	99.372	99.967	+0.595
2	ResNet-50 [8]	23.51	ImageNet-1k	98.074	99.937	+1.863
3	Swin-Base [28]	86.75	ImageNet-22k	95.586	99.868	+4.282
4	ConvNeXt-Small [31]	49.46	ImageNet-22k	99.297	99.845	+0.548
5	EfficientNet-B4 [9]	17.55	JFT-300M (NS)	98.337	99.430	+1.093
6	ViT-Base [27]	85.80	ImageNet-21k	89.359	93.923	+4.564

Note: Δ AUC represents the improvement from frame-level AUC to video-level AUC in percentage points. Boldface identifies the highest-ranked configuration.

At first, ResNet-50 generates the most robust individual video-level AUC, despite its significantly smaller size than Swin-Base and ViT-Base. This implies that a convolutional model with a well-regularized structure may be more effective than a larger transformer-based architecture in the Celeb-DF v2 setting of the dataset. A plausible explanation is that the manipulation traces in Celeb-DF v2 are powerfully represented by local spatial anomalies, including texture inconsistency, boundary blending, and local color disparity. The convolutional receptive fields are well-suited to these stimuli.

Second, ConvNeXt-Small attains the maximum frame-level AUC and video-level accuracy, despite ranking below ResNet-50 and Swin-Base in video-level AUC. This implies that ConvNeXt-Small is the most stable individual model at the default operating threshold, while ResNet-50 and Swin-Base exhibit slightly superior ranking behavior following video aggregation.

Third, ViT-Base performs the worst among the evaluated backbones. Its low frame-level AUC suggests that the standard ViT architecture is less effective at extracting the local forensic traces required for Celeb-DF v2 when trained under the present protocol. Although video-level aggregation improves its AUC by 4.564 percentage points, the final performance remains below the other backbones.

4.2. Effect of Video-Level Aggregation

Table 4 isolates the contribution of video-level mean aggregation by comparing frame-level and video-level AUC for each model. This analysis addresses whether the final performance is mainly due to strong frame-level classification or to the stabilizing effect of aggregating multiple frames.

All models benefit from video-level aggregation. The gain is smallest for ConvNeXt-Small, which improves by only +0.548 percentage points because its frame-level predictions are already very strong. In contrast, ViT-Base and Swin-Base experience considerable improvements as a result of aggregation, with a +4.564 and +4.282 percentage point increase, respectively. This implies that transformer-based models may generate noisier single-frame predictions, but they become more competitive when independently predicted frame probabilities are averaged across multiple sampled frames.

The average AUC gain across the five individual backbones is approximately +2.47 percentage points. This supports the view that video-level aggregation is not merely a reporting option, but a critical element of the detection pipeline because it improves the reliability of the final video-level decision by reducing the influence of isolated equivocal frames. Nevertheless, the result also suggests that robust frame-level performance

remains crucial: the most effective aggregated systems are still those whose individual frame predictions are already highly discriminative.

Table 4. Effect of video-level mean-pool aggregation on AUC and accuracy. Frame-level metrics treat each sampled frame independently, whereas video-level metrics aggregate predictions across 32 sampled frames per video. Reported values are measured in this work; citations identify the original backbone architectures.

Model	Frame AUC (%)	Video AUC (%)	Frame Acc (%)	Video Acc (%)	Δ AUC (pp)
ResNet-50 [8]	98.074	99.937	94.673	94.981	+1.863
ConvNeXt-Small [31]	99.297	99.845	95.150	96.139	+0.548
EfficientNet-B4 [9]	98.337	99.430	93.732	95.753	+1.093
Swin-Base [28]	95.586	99.868	92.930	93.436	+4.282
ViT-Base [27]	89.359	93.923	83.283	84.556	+4.564
Ensemble (top-3, this work)	99.372	99.967	94.739	95.174	+0.595

Note: Boldface identifies the proposed ensemble and its best AUC values.

Two practical implications arise from these findings. First, evaluation of deepfake video detectors using a single frame is insufficient because models with weaker frame-level performance can improve substantially after video-level probability aggregation. Second, the most suitable model depends on the application. ConvNeXt-Small and ResNet-50 are more suitable options for forensic screening at the single-image or thumbnail level. ResNet-50, Swin-Base, and the top-three ensemble become more competitive for complete-video analysis because their predictions are stabilized through aggregation.

4.3. Ensemble Analysis

Probability averaging is employed to combine ResNet-50, Swin-Base, and ConvNeXt-Small in the top-three ensemble. This subsection examines the ensemble from three perspectives: threshold-independent ranking behavior, threshold-dependent classification behavior, and model composition. Because the AUC gain over ResNet-50 is very small, the ensemble is analyzed together with its inference overhead rather than treated as the default deployment configuration.

4.3.1. Ensemble Composition

The ensemble members are analyzed according to individual video-level ranking performance and architectural complementarity within the Celeb-DF v2 protocol. ResNet-50, Swin-Base, and ConvNeXt-Small are included in the reported top-three ensemble because they are the strongest individual models in video-level AUC among the evaluated backbones and provide complementary convolutional and attention-based representations. EfficientNet-B4 and ViT-Base are not included in this top-three configuration because their video-level AUCs are lower. This should be interpreted as an in-dataset model-composition analysis for Celeb-DF v2, not as evidence that the same ensemble composition is optimal under cross-dataset distribution shift or under a deployment-specific efficiency constraint.

The ensemble that was chosen is also architecturally diverse. With its robust local feature extraction, ResNet-50 offers a classical convolutional representation. The ConvNeXt-Small model offers a contemporary convolutional representation characterized by transformer-inspired design choices and larger spatial kernels. Hierarchical window-based self-attention, which Swin-Base contributes, is capable of capturing broader spatial relationships across facial regions. The ensemble is able to incorporate complementary local and global evidence as a result of this combination. Table 5 summarizes the individual ranking performance, parameter count, pretraining source, and inclusion decision for all five evaluated backbones.

Table 5. Composition of the top-3 ensemble used for the in-dataset Celeb-DF v2 analysis. The table reports final individual video-level test AUC values after the unified training and checkpoint-selection protocol. Parameter counts correspond to the instantiated two-class models and are rounded from exact counts. The inclusion decision should be interpreted as a Celeb-DF v2 model-composition analysis based on ranking performance and architectural complementarity, not as cross-dataset or deployment-optimized model selection. Citations identify the original backbone architectures.

Model	Video AUC (%)	Params (M)	Pretraining	Included
ResNet-50 [8]	99.937	23.51	ImageNet-1k	Yes
Swin-Base [28]	99.868	86.75	ImageNet-22k	Yes
ConvNeXt-Small [31]	99.845	49.46	ImageNet-22k	Yes
EfficientNet-B4 [9]	99.430	17.55	JFT-300M (NS)	No
ViT-Base [27]	93.923	85.80	ImageNet-21k	No

EfficientNet-B4 is not included in the reported top-three analysis despite its high accuracy because its AUC is lower than those of the three selected models. ViT-Base is not included because its substantially weaker AUC and high false-positive tendency would reduce the reliability of the ensemble average. Therefore, the selected ensemble is not simply a combination of the largest models; it is a combination of the strongest and most complementary models under the in-dataset Celeb-DF v2 video-level benchmark. A broader comparison of all pairwise, top-three, all-model, and cascade variants is not performed here and is identified as future work.

4.3.2. Confusion Matrix Analysis

Figure 3 and Table 6 report the threshold-dependent confusion matrix behavior at the 0.5 decision threshold. The test set contains 178 real videos and 340 fake videos.

The most important observation is that false positives dominate the remaining error cases. ResNet-50, ConvNeXt-Small, Swin-Base, and the ensemble all correctly detect all 340 fake videos, producing zero false negatives. EfficientNet-B4 misses one fake video, while ViT-Base misses two. Therefore, all models except EfficientNet-B4 and ViT-Base achieve complete recall at the default threshold.

However, the false-positive counts differ substantially. ConvNeXt-Small produces the fewest false positives, misclassifying 20 of the 178 real videos as fake. EfficientNet-B4 follows with 21 false positives. The ensemble produces 25 false positives, which is fewer than ResNet-50 and Swin-Base but greater than ConvNeXt-Small and EfficientNet-B4. ViT-Base produces the weakest false-positive behavior, incorrectly classifying 78 real videos as fake.

Table 6. Confusion matrix counts and derived metrics for all five individual backbones and the ensemble on the Celeb-DF v2 official test set. Reported values are measured in this work; citations identify the original backbone architectures.

Model	TP	FP	FN	TN	FPR (%)	AUC	AP
ResNet-50 [8]	340	26	0	152	14.6	0.9994	0.9997
EfficientNet-B4 [9]	339	21	1	157	11.8	0.9943	0.9970
ConvNeXt-Small [31]	340	20	0	158	11.2	0.9984	0.9992
ViT-Base [27]	338	78	2	100	43.8	0.9392	0.9577
Swin-Base [28]	340	34	0	144	19.1	0.9987	0.9993
Ensemble (top-3, this work)	340	25	0	153	14.0	0.9997	0.9998

TP = fake videos correctly detected as fake; FP = real videos incorrectly classified as fake; FN = fake videos incorrectly classified as real; TN = real videos correctly classified as real. FPR is computed over the 178 real videos. Bold values indicate the best result in each metric column; ties are bolded jointly.

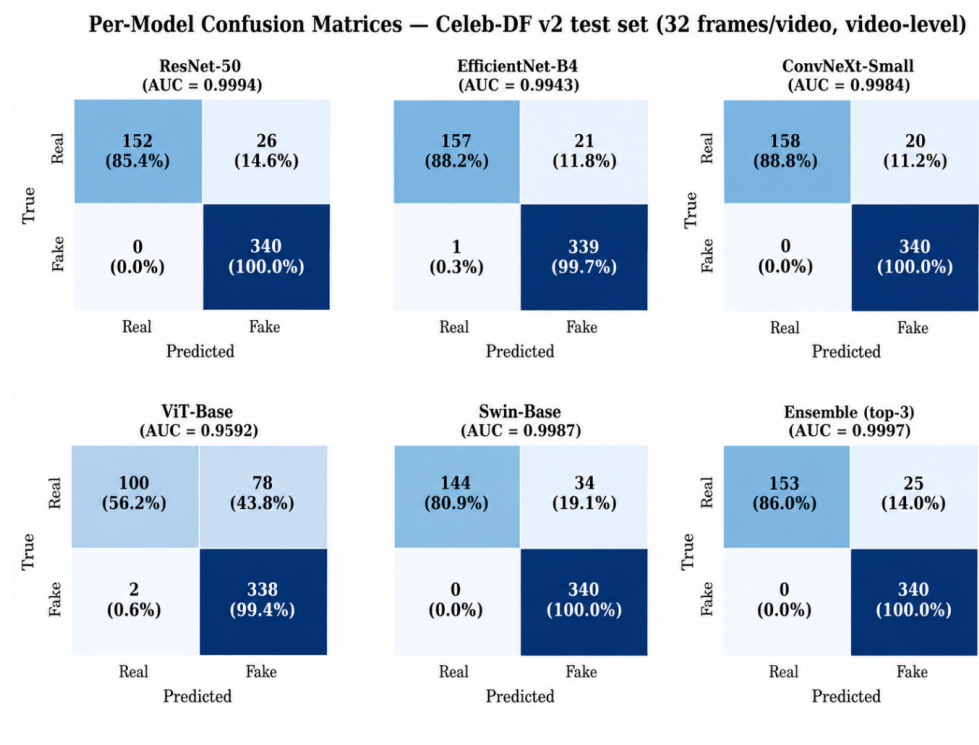


Figure 3. Per-model confusion matrices on the Celeb-DF v2 official test set. The test set contains 178 real videos and 340 fake videos.

The confusion matrix analysis clarifies the difference between ranking quality and fixed-threshold classification. The ensemble has the best AUC and AP, but ConvNeXt-Small has the lowest false-positive rate at the default threshold. Therefore, the ensemble is preferable for threshold-independent ranking, while ConvNeXt-Small is preferable when the operating threshold is fixed at 0.5 and false positives are especially costly.

4.3.3. ROC and Precision–Recall Analysis

Figure 4 shows the ROC and precision–recall curves for the five backbones and the ensemble. Because several models operate near the top-left region of the ROC space, their curves are visually close in the full plot. The zoomed-in ROC curve in Figure 5 provides a clearer view of the differences in the low-FPR region.

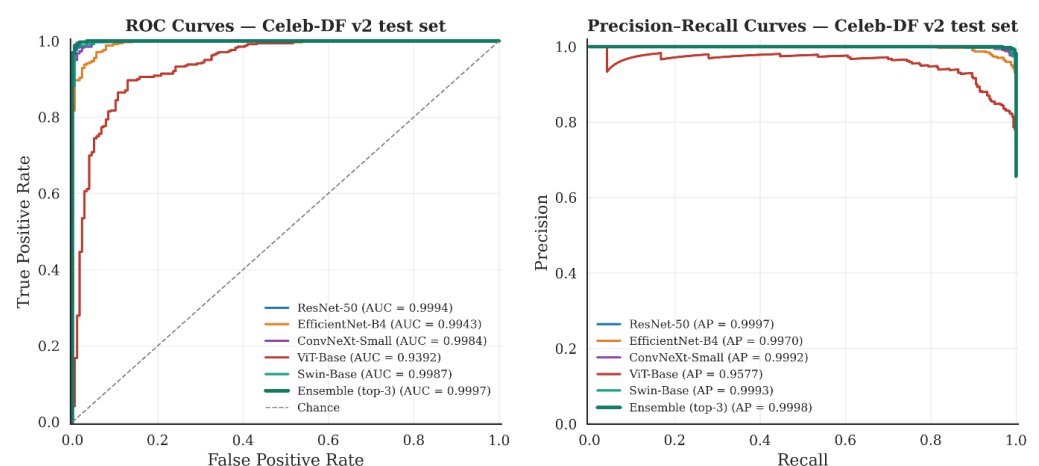


Figure 4. ROC curves and precision–recall curves for the five backbones and the top-3 ensemble at the video level.

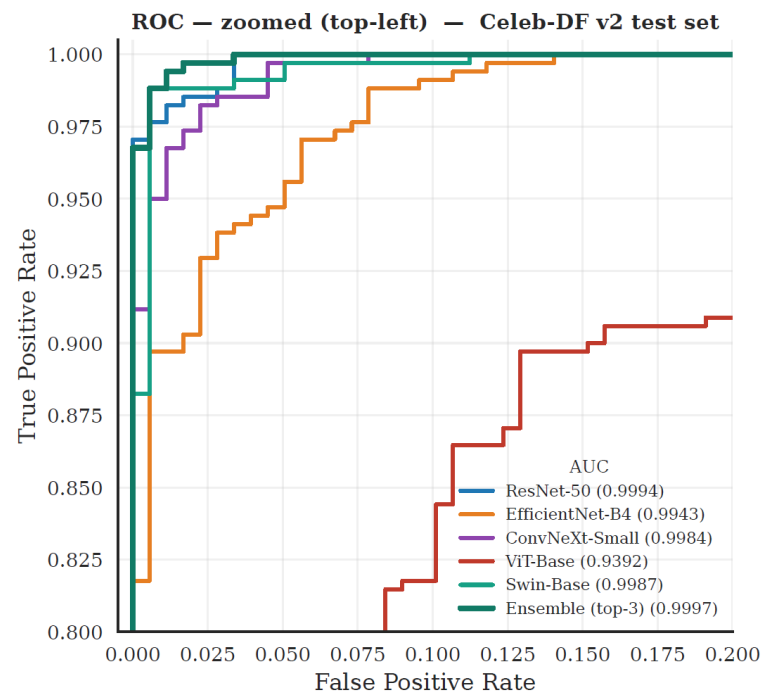


Figure 5. Zoomed ROC curves for FPR between 0 and 0.20 and TPR between 0.80 and 1.00.

The ensemble achieves the highest ROC-AUC and AP, indicating the best ranking performance on all thresholds. ResNet-50, Swin-Base, and ConvNeXt-Small also maintain near-ceiling ranking performance. EfficientNet-B4 remains competitive, but shows a slightly less favorable ROC trajectory. ViT-Base is clearly separated from the other models, particularly in the low-FPR region, which is consistent with its high false-positive count in Table 6.

The precision–recall curves are particularly relevant because the test set is class-imbalanced, with more fake than real videos. The ensemble achieves the highest AP, confirming that its probability scores preserve a strong fake-video ranking across recall levels. However, the fixed-threshold results show that a practical deployment would still require threshold selection based on the desired false-positive tolerance. In applications where false accusations of manipulation are costly, the operating threshold should be selected on a validation set rather than fixed automatically at 0.5.

4.3.4. Video-Level Probability Distribution

Figure 6 shows the fake-probability distributions at the video-level for real and fake videos. The fake video distributions are strongly concentrated near high fake probabilities for most models, indicating that fake videos are generally detected with high confidence. The main difference between models appears in the distribution of real-video probabilities.

ConvNeXt-Small and EfficientNet-B4 show comparatively better separation around the decision threshold for real videos, which is consistent with their lower false-positive counts. ResNet-50 and the ensemble show strong separation overall but retain a tail of real videos above the threshold. Swin-Base exhibits a wider real-video distribution, leading to more false positives. ViT-Base shows the weakest separation, with many real videos receiving elevated fake probabilities. This explains why ViT-Base has both low accuracy and a high false-positive rate despite achieving high recall.

These probability distributions reinforce the need to report more than AUC alone. A detector may rank fake videos above real videos very well while still assigning some real video probabilities above the default decision threshold. Therefore, probability distribution analysis is essential to understand how the detector behaves at practical operating points.

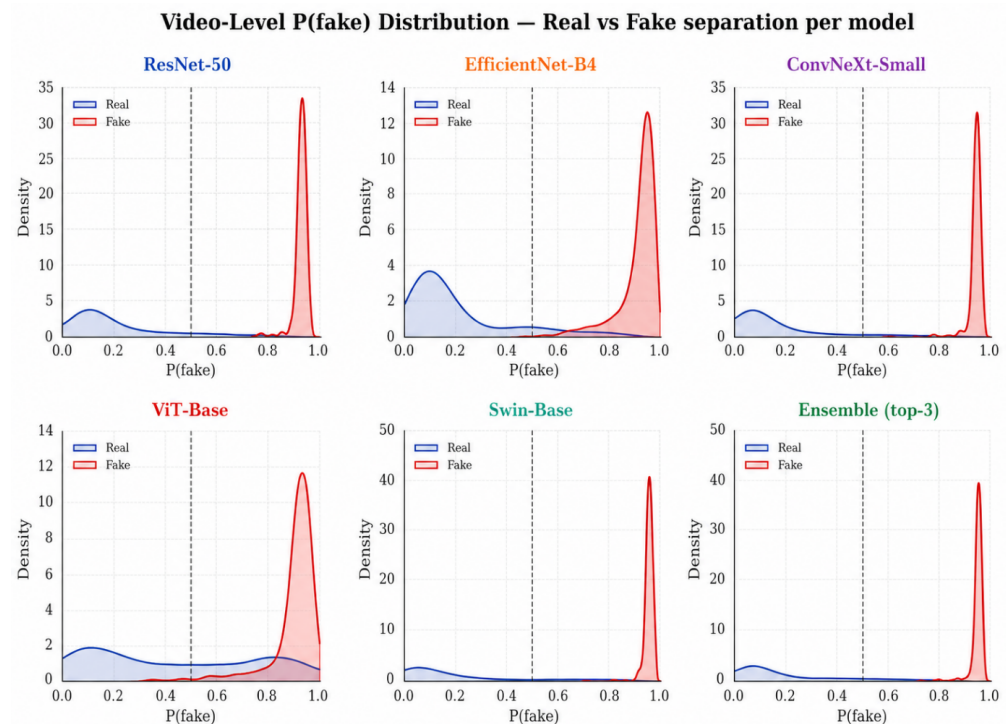


Figure 6. Video-level $p(\text{fake})$ distributions for real and fake videos across all models and the proposed ensemble. The dashed vertical line indicates the 0.5 decision threshold.

4.4. Effect of Test-Time Augmentation

Table 7 reports the effect of the horizontal re-expansion of the test time on the video-level AUC. TTA is applied by averaging the fake probability of each original frame and its horizontally flipped counterpart before video-level aggregation.

Table 7. Effect of horizontal-flip test-time augmentation on video-level AUC for each backbone. AUC values are measured in this work; citations identify the original backbone architectures.

Model	Without TTA (AUC, %)	With TTA (AUC, %)	ΔAUC (pp)
ResNet-50 [8]	99.900	99.937	+0.037
Swin-Base [28]	99.850	99.868	+0.018
ConvNeXt-Small [31]	99.830	99.845	+0.015
EfficientNet-B4 [9]	99.520	99.430	−0.090
ViT-Base [27]	93.690	93.923	+0.233

Note: Boldface indicates the largest positive TTA-related AUC change.

TTA improves four of the five backbones. ViT-Base benefits the most, gaining +0.233 percentage points. This suggests that the weaker single-frame stability of ViT-Base is partly improved by averaging predictions over the original and flipped views. ResNet-50, ConvNeXt-Small, and Swin-Base show smaller gains of +0.037, +0.015, and +0.018 percentage points, respectively. These smaller gains are expected because these models already achieve high video-level AUC without TTA.

EfficientNet-B4 shows a small decrease of −0.090 percentage points after TTA. This does not necessarily imply that TTA is harmful in general. Instead, it suggests that, for this checkpoint and this dataset, the flipped-view probabilities slightly altered the ranking of a small number of borderline videos. Since the change is small, the result should be interpreted as a model-specific sensitivity rather than a systematic failure of TTA.

Overall, TTA provides a modest improvement for most models, but the gains are small compared to the gains obtained from video-level aggregation. Therefore, TTA should

be viewed as a minor inference-time stabilization step rather than as the main source of performance improvement.

4.5. Diagnostic and Proof-of-Concept Scenario Analysis

Beyond aggregate metrics, three diagnostic scenarios are evaluated to better understand how the ensemble behaves at the frame and sampled-video levels. These analyses are not intended to replace the benchmark evaluation or cross-dataset testing. Instead, they provide interpretability-oriented evidence about confidence behavior, probability stability across sampled frames, and a limited proof-of-concept response to a synthetic temporal splice.

4.5.1. Scenario A: Frame-Level Confidence Stratification

Scenario A examines the ensemble's frame-level predictions across different confidence ranges. The frames are grouped according to their predicted $p(\text{fake})$ values into high-confidence correct predictions, lower-confidence correct predictions, and misclassified predictions. This analysis helps identify which visual conditions lead to confident or uncertain predictions.

High-confidence real frames receive consistently low fake probabilities, as illustrated in Figure 7, while high-confidence fake frames receive consistently high fake probabilities. The ensemble probabilities of the selected real frames are less than 0.10, while the ensemble probabilities of the selected fake frames are approximately 0.95–0.96. This suggests that the ensemble generates frame-level scores that are both stable and well-separated in visually evident cases.

The borderline and misclassified images are the most informative. These typically involve one or more of the following conditions: nonstandard illumination, atypical head poses, motion distortion, strong compression anomalies, or background and facial appearance patterns that deviate from typical training examples. Some of these instances indicate that residual errors may be partially caused by non-manipulation factors that influence the visual statistics of the frame. This discovery underscores the necessity of conducting additional robustness experiments in the domains of blur, compression, resolution degradation, and face-cropping variations.

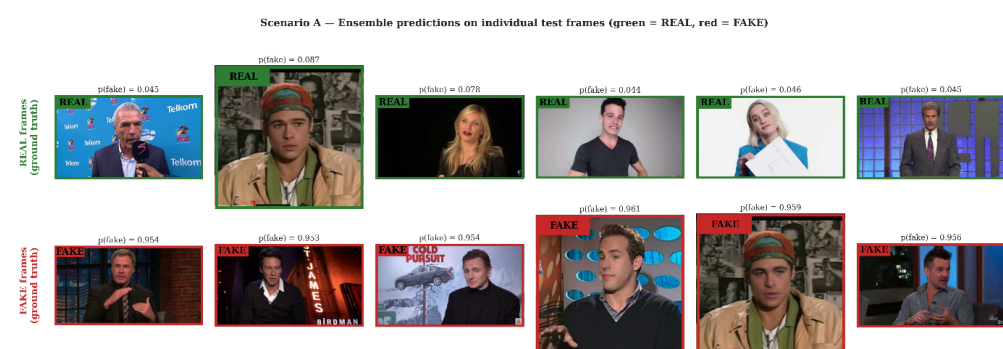


Figure 7. Scenario A: frame-level confidence gallery on the Celeb-DF v2 test set. The top row shows high-confidence real frames, and the bottom row shows high-confidence fake frames, with ensemble $p(\text{fake})$ values shown above each frame.

4.5.2. Scenario B: Frame-Probability Stability Across Sampled Video Frames

Scenario B assesses whether the ensemble generates consistent frame-level probabilities across sampled frames from correctly classified videos. Two genuine videos and two fabricated videos are selected from the correctly classified pool. Each video contains eight uniformly distributed frames, each annotated with the ensemble's fake probability. Figures 8 and 9 present the selected real and fake examples, respectively. This analysis

describes probability stability across selected sampled frames; it should not be interpreted as evidence of learned temporal reasoning.

Scenario B – videos classified entirely as REAL (8 uniformly-sampled frames per video)

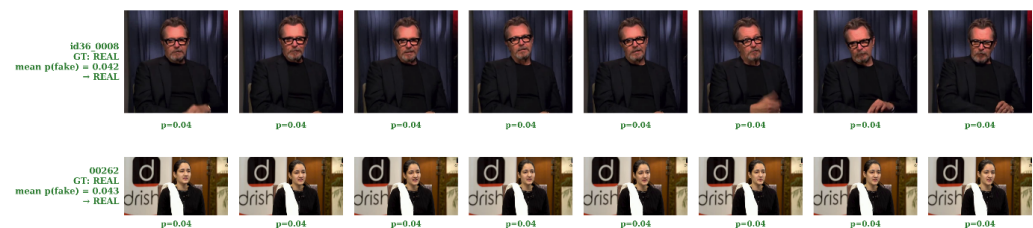


Figure 8. Scenario B: two correctly classified real test videos sampled from Celeb-DF v2, using eight uniformly sampled frames per video.

Scenario B – two REAL test videos (randomly sampled from 295 correctly-classified) (8 uniformly-sampled frames per video; random sample, not cherry-picked)



Figure 9. Scenario B: two correctly classified fake test videos sampled from Celeb-DF v2, using eight uniformly sampled frames per video.

The per-frame fake probabilities for the selected real videos remain consistently close to 0.04 for all displayed frames. The per-frame fake probabilities for the selected fake videos are nearly 0.96. This suggests that the video-level decision is not the result of a limited number of isolated frames. Rather, the model generates consistent evidence across the sampled frames.

This frame-probability stability is useful because a detector that depends on only a few frames may be unstable in practical deployment. The observed stability suggests that, for these correctly classified examples, the ensemble detects a persistent visual signal across the sampled frames. However, because this analysis uses only selected correctly classified videos and does not learn frame-to-frame dependencies, it should be interpreted as diagnostic evidence rather than proof of universal temporal robustness.

4.5.3. Scenario C: Proof-of-Concept Temporal-Splice Sensitivity

Scenario C evaluates whether the ensemble's frame-level probabilities respond to a synthetic temporal splice within a partially manipulated sequence. Since Celeb-DF v2 videos are labeled entirely real or entirely fake, partial sequences are created by splicing frames from a fake video into a real video and vice versa. This analysis is included only as a proof-of-concept diagnostic of probability response. It is not a comprehensive temporal-localization benchmark, not a learned temporal-modeling experiment, and not a spatial localization experiment. Therefore, the perfect values reported below should not be read as evidence of general partial-manipulation detection ability.

Two types of sequence are constructed. In the REAL→FAKE→REAL sequence, the beginning and ending frames are taken from a real video, while the middle segment is replaced with frames from a fake video. In the FAKE→REAL→FAKE sequence, the opposite arrangement is used. The manipulated or inserted segment occupies frames 6 to 25,

corresponding to 20 frames. Table 8 summarizes the frame-level AUC and detection-delay results for both sequence types and all smoothing-window settings.

Table 8. Proof-of-concept temporal-splice sensitivity results for two synthetic partial-manipulation sequences across different smoothing window sizes. The result is based on one real/fake source-video pair and should not be interpreted as a general localization benchmark.

Sequence Type	Smoothing Window	Frame-Level AUC	Detection Delay
REAL→FAKE→REAL	1	1.0000	0
REAL→FAKE→REAL	3	1.0000	0
REAL→FAKE→REAL	5	1.0000	0
REAL→FAKE→REAL	7	1.0000	0
FAKE→REAL→FAKE	1	1.0000	0
FAKE→REAL→FAKE	3	1.0000	0
FAKE→REAL→FAKE	5	1.0000	0
FAKE→REAL→FAKE	7	1.0000	0

Note: Boldface denotes the identical best AUC obtained by all tested sequence and smoothing-window configurations.

For this single real/fake source-video pair, the ensemble achieves a frame-level AUC of 1.0000 for both sequence types under all smoothing windows. The detection delay is zero frames, meaning that the smoothed probability crosses the 0.5 threshold immediately after the class transition. This value is deliberately reported as a diagnostic outcome rather than as a benchmark claim. It shows that the ensemble probabilities respond promptly to this constructed source change, but it does not establish general temporal-splice detection performance across identities, manipulation qualities, compression levels, splice positions, splice durations, or lower-confidence manipulations. The probability trajectories in Figures 10 and 11 show a sharp transition at the splice boundaries for this controlled example only.

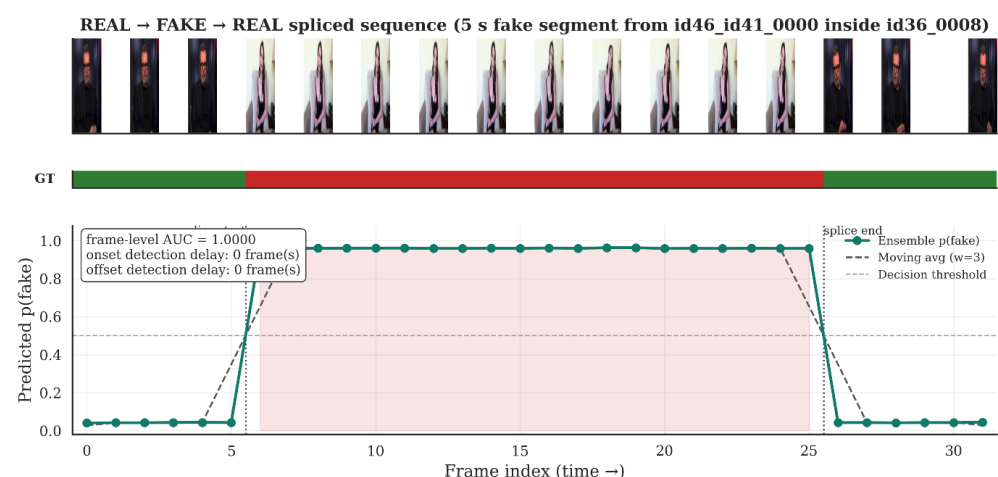


Figure 10. Scenario C: REAL-to-FAKE-to-REAL probability trajectory in the proof-of-concept temporal-splice sensitivity test. The ensemble $p(\text{fake})$ trajectory rises at the inserted fake segment and returns to the real range after the splice offset for this constructed example.

The result demonstrates that the ensemble can respond to the constructed temporal splice in this controlled setting. However, this experiment has an important limitation: it uses one real source video and one fake source video that are already classified with high confidence. Therefore, the perfect result should not be generalized to all possible partial manipulations. More difficult cases, such as low-confidence fake segments, heavy compression, identity-preserving edits, shorter tampered intervals, or more subtle within-frame manipulations, may produce smaller probability jumps and longer detection delays.

The experiment should therefore be interpreted as a proof-of-concept diagnostic analysis rather than a comprehensive temporal-splice benchmark or evidence of pixel-level or region-level localization.

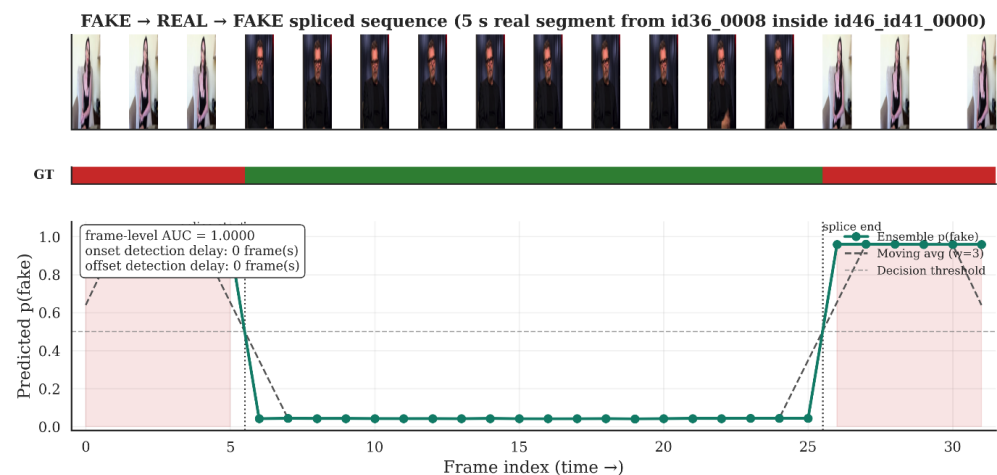


Figure 11. Scenario C: FAKE-to-REAL-to-FAKE probability trajectory in the proof-of-concept temporal-splice sensitivity test. The ensemble probability drops during the inserted real segment and rises again after the fake segment resumes for this constructed example.

4.6. Inference Efficiency and Model Complexity

Predictive performance alone is insufficient to evaluate a deepfake detector intended for practical deployment. A model with a high AUC may not be suitable for real-time or large-scale screening if its latency is high or its throughput is low. Therefore, the evaluated backbones are compared in terms of parameter count, batched throughput, and single-frame latency on an NVIDIA H100 GPU.

4.6.1. Latency and Throughput

Table 9 and Figure 12 summarize the results of the inference efficiency. ResNet-50 achieves the highest batched throughput, processing 11,575 frames/s, with a single-frame latency of 3.14 ms. EfficientNet-B4 has the smallest parameter count, but its throughput and latency are weaker than those of ResNet-50. ConvNeXt-Small provides strong accuracy, but requires higher computation than ResNet-50. ViT-Base achieves the lowest single-frame latency, but its predictive performance is substantially weaker. Swin-Base is the slowest individual model, with 10.49 ms/frame latency and 1993 frames/s throughput.

Table 9. Inference efficiency of the evaluated models. Parameter counts use the same instantiated two-class model convention as all performance tables; ensemble parameters are summed from exact member counts before rounding. Throughput and latency are measured in this work; citations identify the original backbone architectures.

Model	Parameters (M)	Throughput (Frames/s)	Latency (ms/Frame)
ResNet-50 [8]	23.51	11,575	3.14
EfficientNet-B4 [9]	17.55	5986	8.50
ConvNeXt-Small [31]	49.46	4249	6.36
ViT-Base [27]	85.80	7386	2.94
Swin-Base [28]	86.75	1993	10.49
Ensemble (top-3, this work)	159.71	1214	19.99

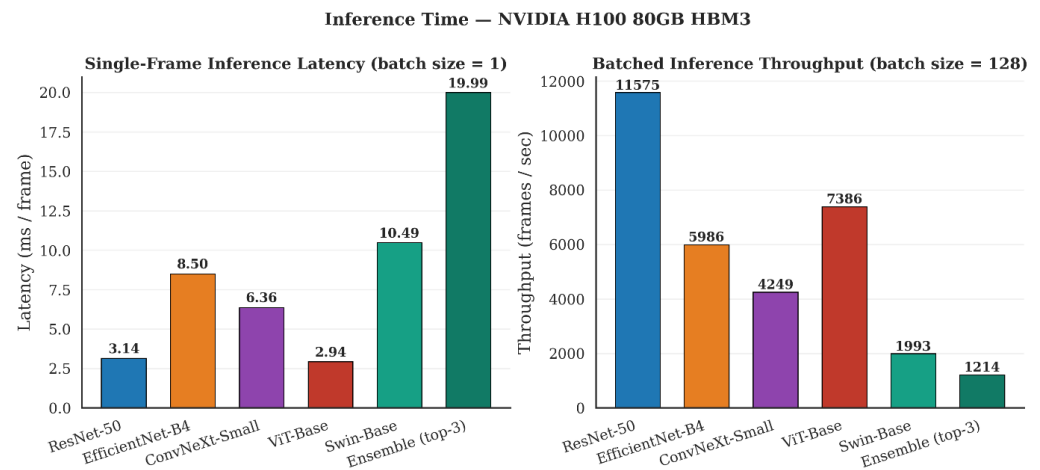


Figure 12. Inference efficiency comparison on an NVIDIA H100 GPU. Single-frame latency is measured at batch size 1, and batched throughput is measured at batch size 128.

These results show that parameter count is not the only determinant of inference speed. EfficientNet-B4 has fewer parameters than ResNet-50 but is slower in both latency and throughput. Similarly, ViT-Base has many more parameters than ResNet-50 but has slightly lower single-frame latency. This indicates that architecture type, operator efficiency, memory access pattern, and GPU kernel implementation also strongly affect inference speed.

4.6.2. Accuracy–Speed Trade-Off

Figure 13 compares video-level AUC with single-frame latency. The preferred region is the upper-left area of the plot, corresponding to a high AUC and low latency.

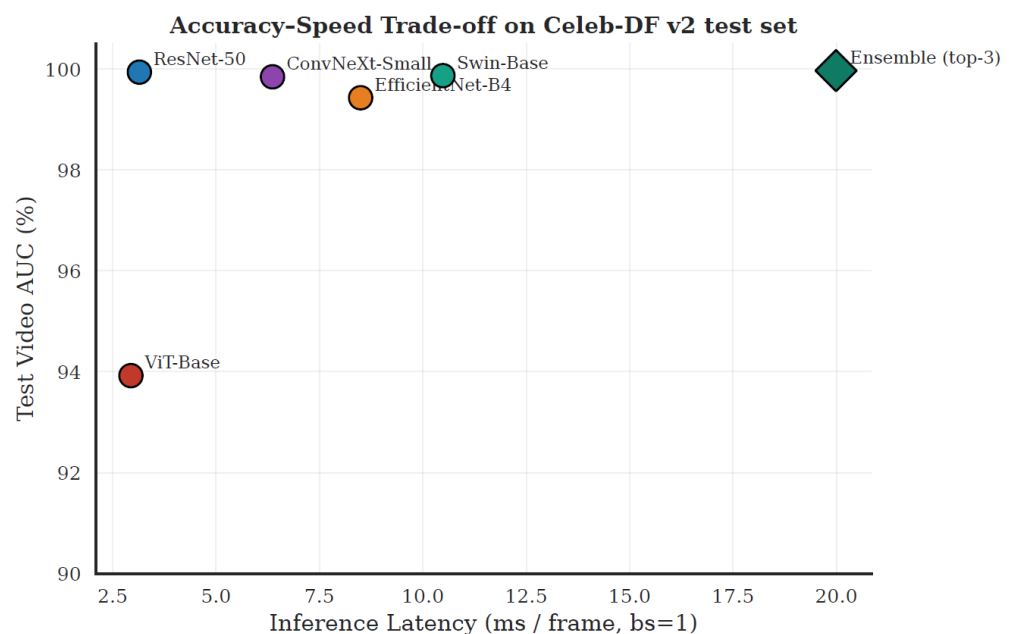


Figure 13. Accuracy–speed trade-off of the evaluated models on the Celeb-DF v2 test set.

ResNet-50 provides the strongest practical trade-off. It achieves near-ceiling video-level AUC while maintaining low latency and the highest throughput. ConvNeXt-Small provides the best fixed-threshold accuracy but is slower than ResNet-50. Swin-Base achieves high AUC but has substantially higher latency. The ensemble achieves the highest AUC, but its latency increases to 19.99 ms/frame and its throughput drops to 1214 frames/s.

Compared with ResNet-50, this corresponds to a $6.4\times$ latency increase and an 89.5% throughput reduction for only a 0.030 percentage-point video-level AUC gain. This makes the ensemble difficult to justify for real-time deployment, although it may still be useful for offline or resource-rich forensic ranking.

Therefore, the preferred model depends on the deployment constraints. ResNet-50 is the best choice when high accuracy and computational efficiency are required. ConvNeXt-Small is attractive when fixed-threshold precision and accuracy are prioritized. The ensemble is most appropriate for offline forensic analysis or resource-rich settings where maximum ranking performance is more important than speed.

4.7. Comparison with Related Literature

Table 10 compares the proposed models with the representative published approaches evaluated on Celeb-DF v2. The video-level AUC is used as the main comparison metric because it is the most commonly reported metric for this benchmark. However, such comparisons should be interpreted carefully because published methods may differ in training data, preprocessing, face-cropping strategy, frame sampling, augmentation, compression setting, model selection, and whether the evaluation is in-dataset or cross-dataset.

The proposed top-three ensemble reaches a video-level AUC of 99.97%, while the proposed ResNet-50 achieves 99.94%. These results are competitive with the strongest Celeb-DF v2 results reported. Importantly, the ResNet-50 result suggests that near-saturated in-dataset performance can be achieved without relying on a very large transformer or a heavy ensemble. This supports the earlier finding that local convolutional representations remain highly effective for Celeb-DF v2.

Table 10. Comparison with the related literature on the Celeb-DF v2 test set. Reported values should be interpreted with caution because training data, preprocessing, sampling strategy, and evaluation settings may differ across studies.

Rank	Method/Model	Reference	Reported Setting	Video-Level AUC (%)	Accuracy (%)
1	Proposed ensemble (top-3)	This work	Official Celeb-DF v2 test set; video-level	99.97	95.17
2	Proposed ResNet-50	This work	Official Celeb-DF v2 test set; video-level	99.94	94.98
3	Exposing the Deception	[44]	Celeb-DF v2 test set; video-level AUC	99.90	NR
4	Proposed Swin-Base	This work	Official Celeb-DF v2 test set; video-level	99.87	93.44
5	Proposed ConvNeXt-Small	This work	Official Celeb-DF v2 test set; video-level	99.84	96.14
6	Proposed EfficientNet-B4	This work	Official Celeb-DF v2 test set; video-level	99.43	95.75
7	FreqBlender	[17]	Celeb-DF v2 test set; video-level AUC	99.30	NR
8	Xception/FaceForensics++ baseline	[7,11]	Celeb-DF v2 test set; video-level AUC	98.50	NR
9	Proposed ViT-Base	This work	Official Celeb-DF v2 test set; video-level	93.92	84.56
10	Self-Blended Images (SBI)	[22]	Celeb-DF v2 test set; video-level AUC	93.70	NR
11	TALL	[42]	Celeb-DF v2 test set; video-level AUC	90.80	NR
12	AltFreezing	[45]	Celeb-DF v2 test set; video-level AUC	89.50	NR
13	FTCN/Temporal Coherence	[41]	Celeb-DF v2 test set; video-level AUC	86.90	NR
14	RealForensics	[20]	Celeb-DF v2 test set; video-level AUC	85.70	NR
15	LipForensics	[19]	Celeb-DF v2 test set; video-level AUC	82.40	NR
16	F3-Net	[14]	Celeb-DF v2 test set; video-level AUC	78.90	NR
17	Meso4 / MesoNet	[6]	Celeb-DF v2 test set; video-level AUC	66.20	NR

Note: NR means not reported. Boldface identifies results obtained in this work; the highest reported accuracy is also bolded.

The comparison should not be interpreted as a definitive classification of all methods, despite the fact that the proposed results are highly competitive, unless all models are evaluated under a unified protocol. Differences in preprocessing, training data, test-time sampling, face extraction, compression level, and model selection can significantly affect Celeb-DF v2 results. Consequently, the comparison is most effectively interpreted as evidence that the proposed framework performs competitively within the Celeb-DF v2

reported literature, whereas a completely controlled comparison would necessitate more standardized benchmarking.

5. Discussion

The experimental results offer a number of critical insights into video-level deepfake detection on Celeb-DF v2. The top-three ensemble obtains the maximum video-level AUC and AP among the evaluated configurations, demonstrating very high in-dataset performance under the official Celeb-DF v2 protocol. However, the results should be interpreted within the scope of the experiments actually performed. A high in-dataset AUC alone is insufficient to establish cross-dataset robustness, learned temporal reasoning, localization capability, or deployment readiness. Differences between frame-level and video-level performance, threshold-independent and threshold-dependent metrics, false-positive and false-negative errors, and predictive performance versus computational cost provide complementary information. This section examines these findings in the context of architecture selection, video-level aggregation, ensemble design, deployment feasibility, and remaining constraints.

5.1. Effectiveness of Video-Level Aggregation

Video-level aggregation significantly improves detection performance in comparison to isolated frame-level classification, which is one of the most apparent findings. All evaluated backbones demonstrate a higher AUC after aggregating predictions across sampled frames. This supports treating deepfake detection as a video-level decision problem rather than as a collection of independent image-level decisions. Occlusion, illumination variation, motion distortion, compression, extreme pose, or weak manipulation traces may affect individual frames. A more stable decision is achieved by averaging predictions across multiple frames, thereby mitigating the influence of such isolated uncertainty. However, this improvement should not be interpreted as evidence of learned temporal modeling, because the aggregation rule does not learn motion dynamics or frame-to-frame dependencies.

Across architectures, the extent of the aggregation advantage is inconsistent. ConvNeXt-Small demonstrates the least growth in AUC from frame-level to video-level, indicating that its single-frame predictions are already highly discriminative. The advantages of aggregation are significantly greater for Swin-Base and ViT-Base, in contrast. This suggests that transformer-based models may produce less stable frame-level predictions in this context; however, their predictions become more reliable when temporal evidence is accumulate across multiple frames. Subsequently, the results suggest that video-level aggregation is not merely a post-processing procedure; rather, it is a critical element of the detection framework that substantially impacts the final benchmark performance.

This observation also has practical implications. Models with robust frame-level performance, such as ConvNeXt-Small and ResNet-50, are more appropriate for applications that require the classification of still images, thumbnails, or extremely brief snippets. Nevertheless, models that capitalize on aggregation can become extremely competitive in the context of full-video verification. Consequently, the detector that is most suitable for the objective task is contingent upon whether it is frame-level screening, short-clip verification, or complete video-level authentication.

5.2. Architecture-Specific Behavior

The findings indicate that the efficacy of deepfake detection is not necessarily enhanced by larger or more recent architectures. ResNet-50 obtains the highest individual video-level AUC, despite having significantly fewer parameters than Swin-Base and ViT-Base. ConvNeXt-Small attains the highest frame-level AUC, the highest video-level accuracy,

and the highest video-level precision. Conversely, ViT-Base is the least effective of the backbones that were assessed.

This outcome suggests that architectural inductive bias remains highly important for deepfake detection. Celeb-DF v2 manipulations often leave local spatial evidence, including texture irregularities, blending artifacts, color inconsistencies, and boundary-level artifacts around manipulated facial regions. Convolutional architectures are naturally well-suited to such local cues because they learn hierarchical spatial filters with locality and translation-related bias. ResNet-50 and ConvNeXt-Small therefore appear to capture the dominant forensic signals in the dataset more effectively than the standard ViT architecture.

The weaker performance of ViT-Base does not necessarily imply that transformer models are unsuitable for deepfake detection. Rather, it suggests that a plain global self-attention architecture may be less effective when the available training data are limited and the discriminative evidence is local and fine-grained. Swin-Base performs substantially better than ViT-Base, likely because its shifted-window mechanism and hierarchical structure preserve more locality than standard ViT. This indicates that transformer architectures with built-in locality or multi-scale representation may be more appropriate for face-forgery detection than fully global patch-based attention alone.

The results also suggest that the pretraining scale alone is not sufficient to explain the deepfake detection performance. EfficientNet-B4 and ViT-Base benefit from large-scale pretraining, but neither outperforms ResNet-50 in this benchmark. This indicates that the compatibility between the model architecture and the forensic artifact structure may be more important than the size of the corpus before training alone. For in-dataset Celeb-DF v2 evaluation, a well-regularized convolutional model can remain highly competitive against larger transformer-based alternatives.

5.3. AUC, Accuracy, and False-Positive Behavior

A central finding is that the best model depends on the metric and operating condition. The top-three ensemble achieves the highest video-level AUC and AP, which means that it provides the strongest threshold-independent ranking of real and fake videos. However, ConvNeXt-Small achieves the highest fixed-threshold accuracy and precision at $\tau = 0.5$. This difference highlights the importance of distinguishing ranking performance from deployment performance.

The extent to which a detector prioritizes fake videos over real videos across all practicable thresholds is quantified by the AUC. However, recall, precision, accuracy, and F1-score are contingent on a specific decision threshold. Although a model may achieve an exceptional AUC, it may still produce a higher number of false positives at a specific threshold if the probabilities of specific actual videos exceed the decision boundary. Despite its superior overall classification, the ensemble fails to produce the lowest false-positive count at the default threshold, as evidenced by the ensemble.

The false-positive classification of real videos as fake is the primary error mode that persists. This results in zero false negatives, as multiple models are able to identify all fraudulent videos at the default threshold. In spite of this, it is not uncommon for real recordings to be inaccurately assigned high levels of fakery. In the context of practical media authentication, false positives can have serious consequences, which is why this is so important. A detrimental effect on trust, reputation, or evidential value can result from the misidentification of an authentic video as manipulated. Consequently, future research should not exclusively focus on maximizing AUC; it should also optimize false-positive control, calibration, and threshold selection.

Based on the application, the deployment threshold should be determined. A security screening system may prioritize high recall and tolerate a greater number of false positives.

Even if recall is marginally reduced, a forensic verification system may require extremely low false-positive rates. Consequently, the evaluation would be further supported by reporting metrics such as FPR at fixed TPR, TPR at fixed FPR, expected calibration error, and reliability diagrams.

5.4. Role and Cost of the Ensemble

The top-three ensemble improves video-level AUC and AP compared to all individual backbones. This confirms that the combination of architecturally different models can improve the ranking performance. The selected ensemble includes ResNet-50, ConvNeXt-Small, and Swin-Base, which represent complementary modeling biases. ResNet-50 provides strong classical convolutional local feature extraction, ConvNeXt-Small provides modern convolutional representation with larger spatial kernels, and Swin-Base contributes hierarchical window-based attention. Probability averaging allows the final score to combine local texture evidence and a broader spatial context.

Still, the advantage of the ensemble over the most powerful individual model is negligible. The ensemble only marginally enhances ResNet-50's near-ceiling video-level AUC. Similarly, the ensemble significantly increases the inference cost by necessitating the execution of all three models. According to the measured latency and throughput results, the ensemble is the slowest configuration, while ResNet-50 offers the most robust accuracy–efficiency trade-off.

In the context of deployment, this establishes a critical distinction. The ensemble is best justified for offline or resource-rich forensic analysis in which threshold-independent ranking quality is the primary objective and inference latency is not a binding constraint. It should not be interpreted as the recommended configuration for all practical deployments. When inference speed and resource efficiency are critical, ResNet-50 is more suitable for high-throughput or real-time environments. When fixed-threshold accuracy and precision are prioritized, ConvNeXt-Small is an appealing alternative. Therefore, the results do not support a single universal deployment option; they instead suggest task-dependent operating modes: ResNet-50 for efficiency, ConvNeXt-Small for fixed-threshold precision/accuracy, and the ensemble only for maximum ranking performance in latency-insensitive settings.

A reliability-aware or cascade-based ensemble is therefore a suitable future direction. Such a system could first evaluate a video using a lightweight model such as ResNet-50 and activate ConvNeXt-Small and Swin-Base only when the initial prediction is uncertain. This could preserve much of the ranking benefit of the ensemble while reducing average inference cost. The strong single-backbone performance observed in this study makes this direction particularly relevant, but it is not implemented here and should be evaluated in future work with explicit uncertainty thresholds and average-cost reporting.

5.5. Interpretation of Diagnostic Scenarios

Diagnostic information beyond aggregate metrics is provided by the diagnostic experiments. In Scenario A, the ensemble effectively differentiates between the genuine and deceptive high-confidence frames. Real frames that exhibit apparent authentic characteristics receive low fake probabilities, while fake frames that contain visible manipulation evidence receive high fake probabilities. Despite this, borderline and misclassified cases serve as evidence that the detector can be affected by compression, atypical visual style, variation of illumination, motion distortion, and an unusual pose. Both of these observations are consistent with the quantitative results, which indicate false-positive behavior.

Consistent frame-level probabilities across sampled frames are examined in Scenario B to determine whether accurate video-level predictions are supported by multiple frames

rather than by a few isolated samples. The ensemble generates highly consistent frame-level predictions for selected videos that have been accurately classified as genuine or fake. This implies that the video-level score is not determined by a small number of isolated decisive frames in these examples. Instead, the model detects a visual signal that remains consistent across the sampled span. This is useful for video-level evaluation, but it should not be interpreted as learned temporal modeling because the frame probabilities are computed independently before mean pooling.

Scenario C evaluates the sensitivity of frame-level probabilities to a synthetic temporal splice. For the evaluated source-video pair, the ensemble produces a perfect frame-level AUC and zero-frame detection delay. This illustrates prompt probability response under controlled conditions. However, the experiment uses one authentic source video and one fake source video that were already classified with high confidence. Therefore, the perfect result does not validate a general temporal-splice detection capability across different videos, manipulation types, compression levels, splice positions, or splice durations, and it provides no evidence of spatial within-frame localization. A more complete evaluation would require a diverse set of source-video pairs, compressed perturbations, shorter manipulated segments, lower-confidence fakes, genuine partially manipulated videos, and, where available, manipulated-region masks.

In general, diagnostic scenarios are valuable because they reveal how the detector behaves internally. They show that the ensemble can produce stable and interpretable probability patterns, but they also expose the need for broader robustness testing.

5.6. Comparison with Existing Methods

The proposed ensemble and the strongest individual backbone achieve a highly competitive Celeb-DF v2 video-level AUC compared to previously reported methods. This suggests that the combination of controlled video-level sampling, strong regularization, pre-trained backbones, test-time enhancement, and probability aggregation is highly effective for this benchmark.

Nevertheless, comparisons with the related literature should be interpreted with caution. Deepfake detection results are highly sensitive to the experimental protocol. Differences in training data, frame sampling, face detection, cropping strategy, compression level, augmentation, model selection, thresholding, and use of external data can significantly affect reported performance. Therefore, ranking methods solely according to Celeb-DF v2 AUC reported may be misleading unless all methods are evaluated within the same framework.

The near-saturated AUC values also suggest that Celeb-DF v2 is not sufficient by itself to demonstrate general robustness. High in-dataset performance does not necessarily imply strong cross-dataset performance. A detector may learn dataset-specific artifacts, compression signatures, face-alignment patterns, or identity-related biases. Therefore, the general detection capability of the proposed framework beyond Celeb-DF v2 remains an open question rather than a demonstrated property. Future work should evaluate cross-dataset generalization on benchmarks such as FaceForensics++, DFDC, DeeperForensics-1.0, and WildDeepfake. Such testing would provide stronger evidence of robustness to unseen manipulation methods and real-world distribution shifts.

5.7. Limitations

Despite the strong results, several limitations remain. First, the evaluation is conducted only on the Celeb-DF v2 benchmark. Although the official test split is used and the protocol is leakage-controlled, this is still an in-dataset evaluation. No cross-dataset testing is performed in the present study. The near-ceiling AUC values therefore cannot estab-

lish generalization to unseen datasets, unseen generation methods, different compression pipelines, or in-the-wild social-media videos. This limitation is particularly important because deepfake detectors can exploit dataset-specific compression statistics, face-alignment patterns, preprocessing artifacts, and identity distributions.

Second, the proposed framework relies on visual information only. It does not incorporate audio, speech–lip synchronization, physiological signals, identity consistency, or metadata. In real-world deepfake detection, multimodal evidence may be necessary because some manipulations may be visually subtle but inconsistent in audio-visual timing, identity behavior, or contextual metadata.

Third, the framework operates on sampled frames and uses mean aggregation rather than explicit temporal modeling. This makes the system simple and reproducible, but it does not learn motion dynamics or frame-to-frame dependencies directly. No CNN-LSTM, 3D CNN, Video Swin Transformer, TimeSformer, or related learned temporal baseline is trained in this work. Therefore, the results cannot determine whether explicit temporal modeling would outperform or complement probability averaging.

Fourth, the ensemble increases computational cost. Although it achieves the best ranking performance, its improvement over ResNet-50 is only 0.030 percentage points in video-level AUC while requiring a $6.4\times$ latency increase. This limits its suitability for real-time deployment. The present study does not implement a ResNet-50-first confidence-gated cascade, pruning, distillation, or dynamic model selection. Therefore, the full ensemble should be interpreted as an offline maximum-ranking configuration rather than as a recommended low-cost deployment model.

Fifth, alternative ensemble schemes are not exhaustively evaluated. The reported top-three ensemble is motivated by in-dataset ranking performance and architectural complementarity, but pairwise ensembles, all-five ensembles, weighted averaging, calibration-aware fusion, and uncertainty-gated routing may produce different accuracy–efficiency trade-offs. These variants require a dedicated deployment-oriented study.

Sixth, the false-positive rate remains a concern. Several real videos are incorrectly classified as fake at the default threshold. For practical digital media authentication, reducing false accusations is as important as detecting fake videos. Future work should therefore include calibration, threshold optimization, and operating-point analysis at fixed false-positive rates.

Seventh, the temporal-splice sensitivity experiment is diagnostic rather than comprehensive. It uses synthetic splicing and one real/fake source-video pair, and it evaluates frame-level probability response rather than spatial localization. It does not establish general partial-manipulation detection, temporal localization, or pixel/region-level localization. A more complete localization evaluation would require a larger set of partial manipulations, different splice lengths, multiple identities, different manipulation qualities, real-world post-processing conditions, and, for spatial localization, manipulated-region masks or supervised localization annotations.

5.8. Future Work

Several directions can extend this study. First, cross-dataset evaluation is the highest-priority extension. Future work should evaluate the Celeb-DF v2-trained checkpoints directly on additional mainstream benchmarks such as FaceForensics++, DFDC, Wild-Deepfake, and DeeperForensics-1.0 without fine-tuning to quantify zero-shot cross-domain AUC, accuracy, AP, and EER degradation. The same leakage-controlled video-level protocol should then be repeated independently on these datasets, and, where computationally feasible, leave-one-dataset-out or joint-dataset training should be performed to determine whether the top-three ensemble composition identified on Celeb-DF v2 remains optimal

under distribution shift. Until such experiments are completed, the general detection capability of the proposed framework beyond Celeb-DF v2 remains unproven.

Second, calibration should be added to the evaluation protocol. Temperature scaling, expected calibration error, Brier score, and reliability diagrams would help determine whether the predicted probabilities can be trusted as confidence estimates. This is especially important for forensic and legal applications where probability interpretation matters and false accusations of authentic content must be minimized.

Third, robustness experiments should be conducted under common real-world distortions, including JPEG compression, resizing, blur, noise, occlusion, brightness changes, face-cropping perturbations, and social-media transcoding. Such experiments would provide a more realistic assessment of deployment performance than a clean in-dataset benchmark alone.

Fourth, the ensemble should be studied under deployment-oriented alternatives rather than only as a full three-model average. Future work should compare pairwise ensembles, all-five ensembles, weighted probability averaging, calibration-aware fusion, and uncertainty-aware model routing. A reliability-aware cascade is particularly important: ResNet-50 could process all videos first, while ConvNeXt-Small and Swin-Base would be activated only for videos whose ResNet-50 confidence is ambiguous. Such a study should report not only AUC and accuracy but also the percentage of videos routed to the heavy ensemble, average latency, throughput, and the trade-off between uncertainty threshold and detection performance.

Fifth, explicit temporal modeling should be incorporated and directly compared against the mean-pooling baseline used in this study. Candidate architectures include CNN-LSTM or CNN-GRU models that consume sequences of per-frame backbone features, 3D convolutional networks operating on stacked frame clips, and temporal-transformer architectures such as Video Swin Transformer or TimeSformer. Such experiments would clarify whether simple probability averaging is competitive with models that directly learn motion dynamics and frame-to-frame dependencies.

Sixth, the proof-of-concept temporal-splice sensitivity experiment should be expanded toward genuine spatial and temporal localization. Future experiments should evaluate a broader and more challenging set of source-video pairs, multiple identities, lower-confidence examples, different splice lengths and positions, compression levels, and post-processing conditions. Where ground-truth manipulated-region masks are available, pixel-level or region-level localization heads, segmentation-based supervision, or saliency-based evaluation should be added so that spatial localization can be reported alongside frame-level temporal sensitivity.

Finally, multimodal detection should be explored. Combining visual features with audio, lip motion, speaker identity, physiological cues, and metadata could improve robustness against advanced deepfakes that minimize visual artifacts. Building on deepfake-specific multimodal, frequency-aware, and temporal research [14,17,24,25,30], future work should evaluate feature-level fusion of complementary spatial, frequency-domain, temporal, and audio-visual evidence rather than relying only on late probability averaging.

6. Conclusions

This study presented a deepfake detection framework at the video-level for the Celeb-DF v2 benchmark using uniform frame sampling, regularized backbone fine-tuning, test-time augmentation, frame-to-video probability aggregation, and probability-level ensembling. Five representative architectures were evaluated under the same experimental protocol: ResNet-50, EfficientNet-B4, ConvNeXt-Small, ViT-Base and Swin-Base. The evaluation included frame-level and video-level metrics, aggregation analysis, ensemble comparison,

probability-distribution analysis, temporal diagnostic experiments, inference-efficiency benchmarking, and comparison with related Celeb-DF v2 studies.

The results show that video-level aggregation is essential for reliable deepfake video detection. All evaluated models improved when frame-level probabilities were averaged across sampled frames, with particularly large gains for transformer-based models. This confirms that isolated frame-level evaluation is insufficient for assessing video deepfake detectors because individual frames may be affected by compression, motion blur, pose variation, or weak manipulation evidence. Video-level aggregation provides a more stable and practically meaningful decision.

Among individual backbones, ResNet-50 achieved the strongest video-level AUC while maintaining a low inference cost, making it the most attractive single-model choice when both accuracy and efficiency are important. ConvNeXt-Small achieved the highest fixed-threshold accuracy and precision, indicating stronger control of false positives at the default threshold. The top-three ensemble of ResNet-50, ConvNeXt-Small, and Swin-Base achieved the highest video-level AUC and average precision, demonstrating that combining complementary convolutional and attention-based representations can improve threshold-independent ranking performance. However, the ensemble also introduced substantially higher computational cost, and its improvement over ResNet-50 was relatively small. Therefore, the most suitable model depends on the deployment objective: ResNet-50 for efficient inference, ConvNeXt-Small for lower false-positive operation at the default threshold, and the ensemble for maximum ranking performance in offline or resource-rich forensic settings.

The analysis also indicates that the remaining errors are dominated by false positives on real videos rather than missed fake videos. This finding is important for practical digital media authentication, where incorrectly labeling authentic content as manipulated may have serious consequences. Therefore, future systems should not only be optimized for AUC, but should also consider false-positive control, threshold selection, calibration quality, and robustness under realistic deployment conditions.

Although the proposed framework achieves highly competitive in-dataset performance on Celeb-DF v2, the results should not be interpreted as complete evidence of cross-dataset or real-world robustness. The near-ceiling AUC values reported here reflect performance under a leakage-controlled, single-dataset evaluation specific to Celeb-DF v2. The study is limited by its reliance on one benchmark, visual detection alone, mean-based probability aggregation rather than learned temporal modeling, and a computationally expensive ensemble. Future work should therefore evaluate cross-dataset generalization, social-media compression, resolution degradation, blur, occlusion, and other real-world perturbations. Additional improvements may also come from uncertainty calibration, dynamic ensemble inference, model compression, explicit temporal architectures, multimodal cues such as audio-visual synchronization and identity consistency, and feature-level fusion of complementary spatial, frequency-domain, temporal, and audio-visual evidence.

Overall, the findings show that carefully trained convolutional and hybrid ensemble models can achieve very strong video-level performance on Celeb-DF v2. At the same time, the study highlights that high benchmark AUC must be interpreted together with false-positive behavior, computational efficiency, the absence of learned temporal modeling, the lack of cascade-based deployment evaluation, and the absence of cross-dataset validation. These considerations are essential before benchmark-oriented deepfake detection can be extended toward reliable digital media authentication in real-world settings.

Author Contributions: Conceptualization, M.A., A.A., Q.E.U.H. and T.M.K.; methodology, M.A. and A.A.; software, M.A. and A.A.; validation, M.A. and A.A.; formal analysis, M.A. and A.A.; writing—original draft preparation, M.A. and A.A.; writing—review and editing, M.A., A.A., Q.E.U.H. and

T.M.K.; supervision, Q.E.U.H. and T.M.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Naif Arab University for Security Sciences under grant No. NAUSS-26-R05.

Data Availability Statement: The Celeb-DF v2 dataset analyzed in this study is publicly available from the dataset authors. No new dataset was created in this study. The trained model checkpoints and derived evaluation outputs are available from the corresponding author upon reasonable request.

Acknowledgments: The authors would like to express their deep thanks to the Vice Presidency for Scientific Research at Naif Arab University for Security Sciences for their kind encouragement of this work.

Conflicts of Interest: The authors declare no conflicts of interest. The funder had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Li, Y.; Chang, M.C.; Lyu, S. In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking. In Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, China, 11–13 December 2018; pp. 1–7. [\[CrossRef\]](#)
2. Yang, X.; Li, Y.; Lyu, S. Exposing Deep Fakes Using Inconsistent Head Poses. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 8261–8265. [\[CrossRef\]](#)
3. Amerini, I.; Galteri, L.; Caldelli, R.; Del Bimbo, A. Deepfake Video Detection through Optical Flow Based CNN. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Seoul, Republic of Korea, 27–28 October 2019; pp. 1205–1207. [\[CrossRef\]](#)
4. Matern, F.; Riess, C.; Stamminger, M. Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations. In Proceedings of the IEEE Winter Applications of Computer Vision Workshops (WACVW), Waikoloa Village, HI, USA, 7–11 January 2019; pp. 83–92.
5. Tolosana, R.; Vera-Rodriguez, R.; Fierrez, J.; Morales, A.; Ortega-Garcia, J. Deepfakes and Beyond: A Survey of Face Manipulation and Fake Detection. *Inf. Fusion* **2020**, *64*, 131–148. [\[CrossRef\]](#)
6. Afchar, D.; Nozick, V.; Yamagishi, J.; Echizen, I. MesoNet: A Compact Facial Video Forgery Detection Network. In Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, China, 11–13 December 2018; pp. 1–7. [\[CrossRef\]](#)
7. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258. [\[CrossRef\]](#)
8. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [\[CrossRef\]](#)
9. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the International Conference on Machine Learning (ICML), Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
10. Korshunov, P.; Marcel, S. DeepFakes: A New Threat to Face Recognition? Assessment and Detection. *arXiv* **2018**. <http://arxiv.org/abs/1812.08685>.
11. Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M. FaceForensics++: Learning to Detect Manipulated Facial Images. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1–11.
12. Li, Y.; Yang, X.; Sun, P.; Qi, H.; Lyu, S. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 3207–3216.
13. Durall, R.; Keuper, M.; Keuper, J. Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks Are Failing to Reproduce Spectral Distributions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2020; pp. 7890–7899.
14. Qian, Y.; Yin, G.; Sheng, L.; Chen, Z.; Shao, J. Thinking in Frequency: Face Forgery Detection by Mining Frequency-Aware Clues. In *Proceedings of the Computer Vision—ECCV, Glasgow, UK, 23–28 August 2020*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 86–103.

15. Luo, Y.; Zhang, Y.; Yan, J.; Liu, W. Generalizing Face Forgery Detection with High-Frequency Features. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 16317–16326.
16. Jeong, Y.; Kim, D.; Min, S.; Joe, S.; Gwon, Y.; Choi, J. BiHPF: Bilateral High-Pass Filters for Robust Deepfake Detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2022; pp. 48–57.
17. Li, H.; Zhou, J.; Li, Y.; Wu, B.; Li, B.; Dong, J. FreqBlender: Enhancing DeepFake Detection by Blending Frequency Knowledge. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 10–15 December 2024; Volume 37.
18. Ciftci, U.A.; Demir, I.; Yin, L. FakeCatcher: Detection of Synthetic Portrait Videos Using Biological Signals. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 3451–3468. [[CrossRef](#)] [[PubMed](#)]
19. Haliassos, A.; Vougioukas, K.; Petridis, S.; Pantic, M. Lips Don't Lie: A Generalisable and Robust Approach to Face Forgery Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 5039–5049.
20. Haliassos, A.; Mira, R.; Petridis, S.; Pantic, M. Leveraging Real Talking Faces via Self-Supervision for Robust Forgery Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 14950–14962.
21. Li, L.; Bao, J.; Zhang, T.; Yang, H.; Chen, D.; Wen, F.; Guo, B. Face X-Ray for More General Face Forgery Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 5001–5010.
22. Shiohara, K.; Yamasaki, T. Detecting Deepfakes with Self-Blended Images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 18720–18729.
23. Zhao, T.; Xu, X.; Xu, M.; Ding, H.; Xiong, Y.; Xia, W. Learning Self-Consistency for Deepfake Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 15023–15033.
24. Mittal, T.; Bhattacharya, U.; Chandra, R.; Bera, A.; Manocha, D. Emotions Don't Lie: An Audio-Visual Deepfake Detection Method Using Affective Cues. In Proceedings of the ACM International Conference on Multimedia (ACM MM), Seattle, WA, USA, 12–16 October 2020; pp. 2823–2832.
25. Zhou, Y.; Lim, S.N. Joint Audio-Visual Deepfake Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 14800–14809.
26. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; Volume 30.
27. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Event, 3–7 May 2021.
28. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
29. Zhao, H.; Zhou, W.; Chen, D.; Wei, T.; Zhang, W.; Yu, N. Multi-Attentional Deepfake Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 2185–2194.
30. Zhao, C.; Wang, C.; Hu, G.; Chen, H.; Liu, C.; Tang, J. ISTVT: Interpretable Spatial-Temporal Video Transformer for Deepfake Detection. *IEEE Trans. Inf. Forensics Secur.* **2023**, *18*, 1335–1348. [[CrossRef](#)]
31. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A ConvNet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 11976–11986.
32. Dolhansky, B.; Bitton, J.; Pflaum, B.; Lu, J.; Howes, R.; Wang, M.; Canton-Ferrer, C. The Deepfake Detection Challenge Dataset. *arXiv* **2020**. <http://arxiv.org/abs/2006.07397>.
33. Jiang, L.; Li, R.; Wu, W.; Qian, C.; Loy, C.C. DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 2889–2898.
34. Zi, B.; Chang, M.; Chen, J.; Ma, X.; Jiang, Y.G. WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection. In Proceedings of the ACM International Conference on Multimedia (ACM MM), Seattle, WA, USA, 12–16 October 2020; pp. 2382–2390.
35. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826. [[CrossRef](#)]

36. Cubuk, E.D.; Zoph, B.; Shlens, J.; Le, Q.V. RandAugment: Practical Automated Data Augmentation with a Reduced Search Space. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Seattle, WA, USA, 14–19 June 2020; Volume 33, pp. 18613–18624.
37. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond Empirical Risk Minimization. In Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
38. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
39. Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; Yang, Y. Random Erasing Data Augmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 13001–13008. [[CrossRef](#)]
40. Güera, D.; Delp, E.J. Deepfake Video Detection Using Recurrent Neural Networks. In Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, 27–30 November 2018; pp. 1–6. [[CrossRef](#)]
41. Zheng, Y.; Bao, J.; Chen, D.; Zeng, M.; Wen, F. Exploring Temporal Coherence for More General Video Face Forgery Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 15044–15054.
42. Xu, Y.; Liang, J.; Jia, G.; Yang, Z.; Zhang, Y.; He, R. TALL: Thumbnail Layout for Deepfake Video Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 22658–22668.
43. Rana, M.S.; Sung, A.H. DeepfakeStack: A Deep Ensemble-Based Learning Technique for Deepfake Detection. In Proceedings of the IEEE International Conference on Cloud Computing and Intelligence Systems (CSCloud), New York, NY, USA, 1–3 August 2020; pp. 70–75.
44. Ba, Z.; Liu, Q.; Liu, Z.; Wu, S.; Lin, F.; Lu, L.; Ren, K. Exposing the Deception: Uncovering More Forgery Clues for Deepfake Detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 719–728. [[CrossRef](#)]
45. Wang, Z.; Bao, J.; Zhou, W.; Wang, W.; Li, H. AltFreezing for More General Video Face Forgery Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 4129–4138.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.