applied sciences

MDPI

Article

AT-HSTNet: An Efficient Hierarchical Action-Transformer Framework for Deepfake Video Detection

Sameena Javaid [1], Marwa Chendeb El Rai [2], Abeer Elkhouly [3,*], Obada Al-Khatib [3], Aicha Beya Far [4] and May El Barachi [5]

[1] School of Information Technology, Murdoch University, Dubai 500700, United Arab Emirates; sameena.javaid@murdoch.edu.au
[2] Mathematics Department, School of Art and Science, American University in Dubai, Dubai 28282, United Arab Emirates; melrai@aud.edu
[3] School of Engineering, University of Wollongong in Dubai, Dubai 20183, United Arab Emirates
[4] School of Engineering, American University in Dubai, Dubai 28282, United Arab Emirates
[5] School of Computer Science, University of Wollongong in Dubai, Dubai 20183, United Arab Emirates; maielbarachi@uowdubai.ac.ae
* Correspondence: abeerelkhouly@uowdubai.ac.ae

Abstract

The rapid advancement of deepfake generation technologies presents significant challenges to the verification of digital video authenticity. These time-dependent artifacts are difficult to detect using conventional frame-based detection approaches. This paper introduces AT-HSTNet, an Action-Transformer-based Hierarchical Spatiotemporal Network designed for robust and computationally efficient deepfake video detection. The proposed framework adopts a multi-stage hierarchical architecture in which frame-level visual features are extracted using an EfficientNet-B0 backbone, short- and medium-range temporal patterns are modeled through Bidirectional Long Short-Term Memory (BiLSTM) networks, and long-range temporal dependencies are captured using an action-aware Transformer operating on temporally aggregated representations. Unlike conventional video transformers that apply self-attention directly to raw frame-level features, the proposed action-aware attention mechanism reduces redundant computation and improves stability in temporal reasoning. Extensive experiments on the balanced FFIW-10K dataset demonstrate that AT-HSTNet achieves an accuracy of 98.7%, with 98.0% precision, 96.0% recall, and a 96.9% F1-score, outperforming representative CNN–BiLSTM and CNN–Transformer baseline architectures. In addition, AT-HSTNet is highly efficient, requiring only 0.45 GFLOPs and achieving an inference speed of approximately 30 FPS on consumer-grade GPU hardware. As a result of this study, we found hierarchical temporal modeling more effective when combined with action-aware attention for any deepfake video detection.

Keywords: deepfake detection; video forensics; spatiotemporal modeling; action transformer; hierarchical temporal learning; efficient deep learning

Academic Editor: Thomas Lindner

Received: 2 March 2026
Revised: 28 March 2026
Accepted: 29 March 2026
Published: 2 April 2026

1. Introduction

Digital Media Creation is significantly transformed by advanced deep learning and generative artificial intelligence models for the creation of substantially realistic images and videos [1]. With these developments, deepfakes have emerged as a potential threat to media credibility and authenticity, as manipulated facial identities, expressions, and voices are extremely convincing [2].

Recent advancements in deepfake generation techniques have achieved an exceptional extent of visual realism, presenting significant threats to information integrity, public trust, and digital security. This rapid advancement is mainly directed by intensive research on generative artificial intelligence, its architectures, and the foundation of generative adversarial networks [3].

Previous research on deepfake detection focused on spatial discrepancies between frames, such as incorrect blending, texture anomalies, or unusual facial boundaries [4]. While these strategies were effective for the first generations of deepfake datasets, quick advances in generative models have considerably decreased visible artifacts in more recent datasets [5]. As a result, many recent deepfakes appear realistic in individual frames but exhibit flaws over time. These issues include unnatural facial motions, inconsistent expressions, and shifts in identity across frames [6]. While these flaws may not be visible in a single frame, they become clear when the video is studied as a continuous series.

Recent deepfake detection research has integrated both spatial and temporal clues to alleviate these constraints. Convolutional Neural Networks (CNNs) are commonly used to extract discriminative spatial features from individual video frames, while recurrent architectures like LSTM networks model short-term temporal inconsistencies between consecutive frames to improve deepfake video detection performance [7–9]. Researchers have also employed Transformer-based attention mechanisms, which have lately shown enormous potential for capturing long-range temporal dependencies by allowing direct interactions across non-neighboring frames [10]. However, there are several limitations to such approaches. Methodologies based on frame-level feature extraction often result in increased memory consumption, higher complexity in computation, and unstable training because of redundant representation across adjacent frames. Similarly, temporal models may also suffer from gradient instability and sensitivity to noise, which can affect detection accuracy and efficiency, as well.

A further notable limitation of current deepfake detection algorithms is their computational demand. A number of current models require deep backbones or extensive temporal attention mechanisms and hence cannot be applied in resource-constrained settings [11,12]. Deepfake datasets are often unbalanced and contain videos with different compression levels, resolutions, and other processing artifacts. These differences make it difficult to train a model that performs well on new data.

AT-HSTNet, an Action-Transformer-based hierarchical spatiotemporal network, is proposed to address these challenges through a simple and structured approach for temporal modeling. The overall framework follows a hierarchical design in which spatial features are first extracted from each video frame using a lightweight convolutional backbone. Then, short- and medium-range time dependencies are modeled using bidirectional recurrent encoders to produce aggregated sequence representations. Based on these representations, an action-aware Transformer module performs high-level sequence reasoning to identify long-range anomalies in the video. This reduces redundant computation, as time-abstracted tokens are used instead of raw frame-level features, with the additional benefit of improving sensitivity to small changes over time.

To facilitate efficient training on long video sequences, the model includes several memory-efficient optimizations, such as gradient accumulation, mixed-precision training, and exponential moving average-based parameter smoothing. Robust data augmentation approaches are used to improve generalization across many deepfake generating methods and real-world video distortions. AT-HSTNet achieves a compromise between detection accuracy, computational efficiency, and training robustness by capturing spatial artifacts and time-based inconsistencies on several levels. This makes the proposed model ideal for actual deepfake detection tasks that demand consistent and efficient performance.

2. Related Work

The rapid rise of deepfake generation technology has prompted extensive research into deepfake detection. Early research mostly focused on detecting visual artifacts in single frames via classical machine learning and CNN-based approaches [13]. These approaches relied mostly on low-level spatial indicators such as texture pattern anomalies, blending errors, and face boundary discrepancies. While CNN-based detectors performed well on early datasets, their effectiveness has decreased as more recent generation models provide visually consistent frames with little spatial artifacts [14].

To address these limitations, current research has centered on simulating changes across time. This is based on the observation that deepfake videos frequently exhibit minor inconsistencies across frames rather than significant faults inside of a single frame. RNN-based models, such as LSTM networks, were developed to detect frame-to-frame dependencies and motion anomalies [15]. Although these methods enhanced detection robustness, they frequently struggled to describe long-term time dependencies and were sensitive to noise and redundancy in frame-level representations. Furthermore, simple CNN-RNN pipelines frequently required substantial feature engineering and demonstrated poor ability to represent complicated temporal correlations across full video sequences [16,17].

The introduction of attention mechanisms and Transformer architecture marked the beginning of an important phase shift in deepfake detection. Transformers capture direct interactions among distant temporal elements through self-attention and are therefore suitable for learning long-range dependencies. Preliminary Transformer-based approaches applied vision Transformers on frame-level features, often using CNN backbones [18]. More sophisticated Vision Transformer models have used CNN-based patch embeddings and knowledge distillation strategies to further improve detection accuracy on large-scale datasets, such as Deepfake Detection Challenge (DFDC) and Celebrity DeepFake (Celeb-DF) [19]. These models are powerful, but they demand high computational cost due to deep transformer layers and a large number of parameters.

A few more studies have addressed deepfake detection using Transformer-based methods. These approaches use multi-stream attention and patch selection within end-to-end vision transformers to improve performance on real-world, unconstrained datasets [20]. Although they achieve high accuracy, most rely on dense attention across all patches rather than frame-level features, leading to high memory usage and slow runtime.

In particular, hybrid CNN–Transformer architectures have recently emerged as an effective compromise between computational efficiency and representation ability for video understanding tasks. In these frameworks, convolutional networks are commonly used to extract spatial features from individual frames, while Transformer encoders are employed to perform higher-level reasoning and capture complex dependencies across the video sequence. Key frame selection and re-attention methods reduce temporal redundancy by assigning higher importance to informative and temporally salient frames [18,21]. Similarly, hybrid CNN–Vision Transformer architectures integrate Xception-based feature extractors with transformer modules to improve robustness under different video compression levels [22]. While these approaches improved generalization, their time modeling was often shallow and relied on heuristic frame sampling instead of clear and structured time abstraction.

More complex transformer variants approached video modeling more explicitly in both the spatial and temporal dimensions. Interpretable spatial–temporal video transformers decompose attention into spatial and temporal components, allowing for richer video-level reasoning [23]. Other studies further utilize auxiliary representations, such as UV texture maps, and adopt incremental learning strategies to enhance the analysis of temporal coherence [24].

While effective, these typically demand complex pre-processing and involve very high computational overheads. More recently, even deeper hybrid architectures that feature CNNs, recurrent networks, or transformers in a single framework have come under active investigation. Identity-aware deepfake detection methods have combined CNNs, LSTMs, and transformers with 3D morphable models, yielding impressive accuracy by enforcing temporal identity coherence [16]. However, these systems often need reference identity information and require costly preparation, which limits their practical use. As a result, there is a growing need to investigate fully Transformer-based architectures that jointly model visual appearance and temporal changes in video data. Some recent approaches use window-based attention, such as Swin Transformers, for spatial feature extraction, along with transformer encoders to model relationships over time [25]. While these models achieve state-of-the-art performance on several benchmark datasets, they still require very high computational resources due to the use of Transformers at multiple stages of the processing pipeline. Overall, the literature shows a slow but steady shift from purely spatial CNN-based detectors to more advanced Transformer-based designs that explicitly model temporal information. Although Transformer-based models significantly improve long-range time reasoning, many recent approaches focus only on frame-level features or rely on fully time-based Transformers, both of which can lead to redundant computation and reduced efficiency.

Table 1 summarizes representative Transformer-based and hybrid deepfake detection methods published between 2021 and 2025. The table compares these approaches in terms of their core techniques, evaluated datasets, reported performance metrics, and computational efficiency. Overall, while many methods achieve strong detection performance across benchmark datasets, most suffer from low to medium efficiency due to complex architectures and attention-heavy designs.

Table 1. Comparative summary of representative deepfake video detection methods.

Year	Technique	Dataset(s)	Metrics	Efficiency
2021 [24]	Video Transformer with UV texture alignment and incremental learning (Xception + Video Transformer)	FaceForensics++, Deep Fake Detection Challenge (DFDC)	FF++: ACC 99.52%, AUC 99.64% DFDC: ACC 91.69%	High computational cost
2023 [19]	Improved Vision Transformer with CNN–patch feature fusion and knowledge distillation (EfficientNet-B7 + ViT + DeiT)	Deep Fake Detection Challenge (DFDC), Celeb-DF	DFDC: AUC 97.8%, F1 91.9% Celeb-DF: AUC 99.3%, F1 97.8%	Very high computational cost (440 M parameters, ~8–10× higher than CNN-based models)
2022 [20]	DFDT: End-to-End Vision Transformer with multi-stream re-attention and patch selection	FaceForensics++, Celeb-DF, WildDeepfake	FF++: ACC 99.41%, AUC 99.94% Celeb-DF: ACC 99.31%, AUC 99.26%	High computational cost (multi-stream Transformer architecture, requires multi-GPU training)
2025 [26]	Hybrid Transformer Network with early feature fusion (XceptionNet + EfficientNet-B4 + ViT)	FaceForensics++, Deep Fake Detection Challenge (DFDC)	FF++: ACC 97.00% DFDC: ACC 98.24%	Moderate computational cost (hybrid CNN–Transformer with Time Sformer)

Table 1. *Cont.*

Year	Technique	Dataset(s)	Metrics	Efficiency
2025 [18]	Deep Convolutional Pooling Transformer with key frame selection and re-attention mechanism	FaceForensics++, Deep Fake Detection Challenge (DFDC), Celeb-DF, DeeperForensics	FF++: ACC 92.11%, AUC 97.66% DFDC: ACC 65.76%	High computational cost (hybrid Xception + ViT with spatiotemporal attention and preprocessing overhead)
2023 [23]	ISTVT: Interpretable Spatial–Temporal Video Transformer with decomposed attention and self-subtraction	FaceForensics++, FaceShifter, DeeperForensics, Celeb-DF, Deep Fake Detection Challenge (DFDC)	FF++: ACC 99.6%, AUC 99.6% Celeb-DF: AUC 99.8%	High computational cost (spatiotemporal Transformer with decomposed self-attention)
2023 [21]	Deep Convolutional Pooling Transformer with key frame selection and re-attention mechanism	FaceForensics++, Deep Fake Detection Challenge (DFDC), Celeb-DF. DeeperForensics	FF++: ACC 92.11%, AUC 97.66% DFDC: ACC 65.76%, AUC 73.68% Celeb-DF: ACC 63.27%, AUC 72.43%	High computational cost (deep CNN + 24-layer Transformer with re-attention and preprocessing overhead)
2022 [22]	HCiT: Hybrid CNN–Vision Transformer for spatiotemporal deepfake detection (Xception + ViT)	FaceForensics++, Deep Fake Detection Challenge (DFDC), Celeb-DF	FF++: ACC 96.0%, F1 93.86% DFDC-p: ACC 97.82%	High computational cost (dual CNN feature extractors + Transformer with feature fusion and preprocessing overhead)
2024 [25]	SFormer: End-to-end spatiotemporal Transformer using Swin Transformer for spatial modeling and Transformer encoder for temporal reasoning	FaceForensics++, DFD, Celeb-DF, Deep Fake Detection Challenge (DFDC), Deeper-Forensics	FF++: 100% DFD: 97.81% Celeb-DF: 99.1% DFDC: 93.67%	Moderate to high computational cost (end-to-end spatiotemporal Transformer with Swin backbone)
2025 [16]	Hybrid CNN–LSTM–Transformer with 3D Morphable Models for identity-aware deepfake detection	VoxCeleb2 (train), DFD, Celeb-DF, FF++ (test)	DFD: AUC ≈ 97% Celeb-DF: AUC ≈ 86% FF++: AUC ≈ 99%	Moderate computational cost with improved inference efficiency (hybrid CNN–LSTM–Transformer; moderately reduced inference time)

Table 1 further highlights the importance of modeling changes over time in deepfake detection while also revealing a trade-off between detection accuracy and computational cost. Existing methods either rely on large time-aware Transformers with high computational demands or apply attention at the frame level, leading to redundant computation and unstable training on long videos. These findings motivate the need for an efficient hierarchical approach. Accordingly, we propose a framework that combines visual feature extraction, recurrent sequence encoding, and action-aware, Transformer-based reasoning, which is detailed in the next Section 3.

3. Proposed AT-HSTNet Framework and Methodology

The Hierarchical Action-Transformer design is presented in the current architecture of AT-HSTNet. This deepfake detection model analyzes both visual artifacts in frames and inconsistencies that appear in videos over time. At the frame level, first, the visual

features are extracted and then modeled across frames when there are short- and medium-term changes. Furthermore, using an action-aware Transformer, long-range pattern are captured. Stable training, video-level reasoning, and efficient computation are maintained by this design.

Figure 1 provides an overview of the AT-HSTNet architecture. Here, EfficientNet-based spatial feature extraction, temporal modeling by BiLSTM, and Action-Transformer based long-range reasoning constitute the robust hybrid spatiotemporal deepfake detection system.

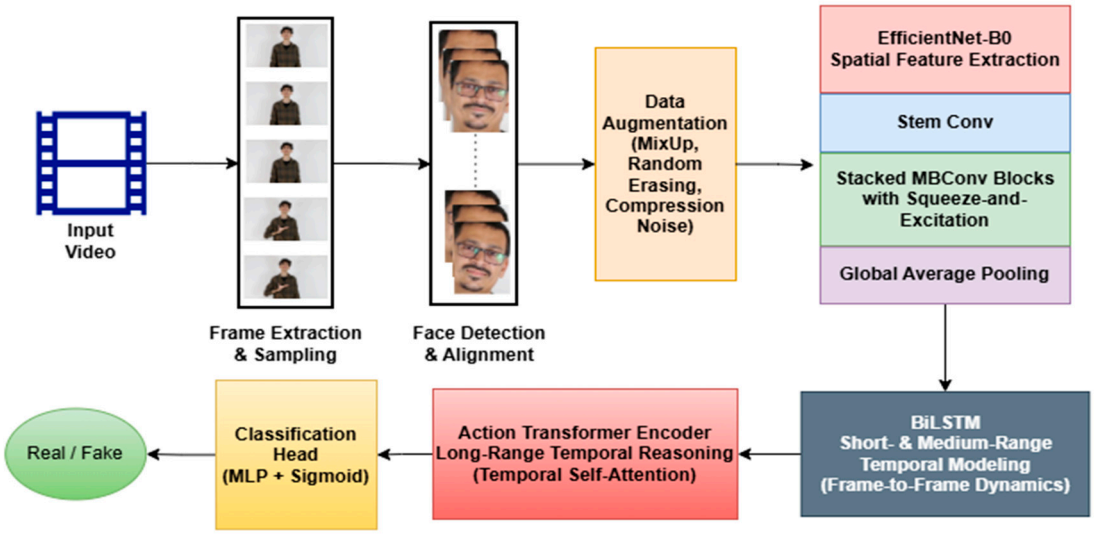

Figure 1. Architecture of AT-HSTNet (Action-Transformer-Based Hierarchical Spatiotemporal Network).

To further improve clarity, Algorithm 1 summarizes the model depicted in Figure 1.

Algorithm 1: AT-HSTNet for Deepfake Video Detection.

Input: Video sequence V
Output: Prediction label $y \in \{\text{Real}, \text{Fake}\}$
Extract frame sequence $F = \{f_1, f_2, \ldots, f_n\}$ from input video V
For each frame $f_i \in F$:

a. Detect facial region R_i
b. Perform face alignment and normalization to obtain A_i
c. Construct aligned face sequence $S = \{A_1, A_2, \ldots, A_n\}$
d. Spatial Feature Extraction:
   i. For each $A_i \in S$, extract feature vector x_i using EfficientNet-B0
   ii. Obtain feature sequence $X = \{x_1, x_2, \ldots, x_n\}$
e. Short-Term Temporal Modeling:
   Pass X through BiLSTM to obtain temporal features H
f. Long-Term Temporal Modeling:
   Feed H into the Action Transformer to obtain Z
g. Feature Aggregation and Classification:
   Aggregate features Z and pass through fully connected layers
h. Compute final predictions using SoftMax:
   $$y = \text{Softmax}(\text{FC}(Z))$$

The AT-HSTNet architecture is described in detail in the following sections. Specifically, Section 3.1 describes the data and its preprocessing pipeline. Spatial feature extraction is explained in Section 3.2 via BiLSTM, and the Action-Transformer is further described in Section 3.3 for long-term temporal reasoning. Finally, the classification and prediction mechanism is elaborated in Section 3.4.

3.1. Dataset and Pre-Processing

In the current work, the Face Forensics in Wild-10K (FFIW-10K) dataset is employed both for training and testing. In total, 10,000 videos were distributed into two genuine and manipulated classes, with 5000 videos each [27]. Having the same amount in both classes prevents issue of training bias. FFIW-10K holds a wide range of facial expressions, movements, and manipulation techniques, which makes it suitable for the evaluation of spatiotemporal deepfake analysis and detection models.

To enhance generalization, data augmentation is employed before training. At the video level, MixUp is applied through interpolation of sample pairs and their corresponding labels, which allows the model to develop smoother decision boundaries and reduces its sensitivity to ambiguous instances [28]. Further, Random Erasing is implemented at the art frame level, including randomly sized regions to reduce reliance on localized artifacts and improve learning [29]. Frames are extracted through uniform temporal sampling to ensure temporal coherence across videos of different lengths. When a video has an insufficient number of frames, deterministic frame repetition is used to maintain a stable temporal structure. All frames are resized to 224 × 224 pixels and normalized based on ImageNet statistics to match the EfficientNetB0 backbone. All data in FFIW-10K were split into training, validation, and test sets according to a video-level partitioning scheme. Specifically, 70% of the videos were assigned to the training set, 15% to the validation set, and the remaining 15% to the test set.

3.2. AT-HSTNet Hybrid Spatiotemporal Feature Extraction Architecture

AT-HSTNet's core combines efficient spatial representation learning and hierarchical temporal modeling to detect static manipulation artifacts as well as dynamic temporal inconsistencies.

3.2.1. Spatial Feature Extraction

Each preprocessed frame S_i' is passed through an EfficientNet-B0 backbone to extract compact and discriminative spatial features called $F_{spatial}^{(i)}$; the extraction of spatial features from video frames is defined by Equation (1) [30]:

$$F_{spatial}^{(i)} = EfficientNet(S_i') \quad (1)$$

EfficientNet-B0 utilizes Mobile Inverted Bottleneck Convolutions together with squeeze-and-excitation techniques that emphasize informative channel features with low computational complexity. Hence, every frame is represented by a 1280-dimensional feature vector embedding fine-grained facial texture and appearance imperfections that are crucial for deepfake detection.

3.2.2. Temporal Sequence Modeling

In this section, for Equation (2), h_{t-1} and c_{t-1} come from the previous recurrent step of the BiLSTM itself. Specifically, these are the hidden state and the cell state, respectively. These are generated with the time step of t − 1 when the previous spatial feature vector $F_{t-1}^{spatial}$ is processed by the network. While sequential modeling LSTM computes h_1, c_1 initially and further for each subsequent step, the new computed states are forwarded

recursively. Thus, h_{t-1} and c_{t-1} are internally generated memory representations, allowing the model to retain and generate temporal information.

Therefore, the sequence of frame-level spatial features ($F_{spatial}$), the previous hidden state (h_{t-1}), and the previous cell state (c_{t-1}) is processed using a Bidirectional Long Short-Term Memory (BiLSTM) network to encode short- and medium-range temporal dependencies, defined by Equation (2) [31]:

$$h_t, c_t = LSTM\left(F_{spatial}, h_{t-1}, c_{t-1}\right) \quad (2)$$

AT-HSTNet is able to incorporate both previous and upcoming contextual information due to bidirectional formulation, which aids the detection of irregular temporal transitions and localized inconsistencies over frames.

3.3. Action-Transformer-Based Contextual Reasoning

In this AT-HSTNet model, the Transformer module serves as a high-level action-aware temporal reasoning head network rather than a general temporal encoder. Unlike previous CNN-Transformer pipelines, which apply self-attention directly to frame-level features, AT-HSTNet applies the Transformer to temporally abstracted BiLSTM representations.

3.3.1. Input Representation

The input to the Transformer encoder consists of the BiLSTM output sequence, as expressed by Equation (3):

$$H_{lstm} = \{h_1, h_2, \ldots, h_N\} \quad (3)$$

Here, in Equation (3), each of the hidden states h_t captures the localized temporal dynamics from the frames in the neighborhood, which represent not only motion continuity and expression transitions but even temporary inconsistencies. These states can be regarded as tokens summarizing temporal information, like action proposals used in video understanding frameworks based on Action-Transformers.

3.3.2. Self-Attention Mechanism

Each temporal token is projected into query (Q), key (K), and value (V) spaces, as represented by Equation (4):

$$Q = H_{lstm}W_Q, K = H_{lstm}W_K, V = H_{lstm}W_V \quad (4)$$

Self-attention is computed by using Equation (5):

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (5)$$

This technique allows the tokens that are far apart temporally to communicate directly with each other and enables AT-HSTNet to identify long-range irregularities in videos, such as unbalanced facial movements, rapid shifts in identities, and incoherent expressions for an instance [32].

The proposed architecture enables temporal reasoning inside of an action-aware framework by considering each BiLSTM hidden state as a temporal event instead of an isolated frame description. In this way, the Action-Transformer is capable of selectively attending to time-informative regions while down-weighting time-invariant or less discriminative regions, which is particularly beneficial in deepfake detection scenarios where the manipulation signals are not continuously present throughout the video. The overall approach to handling the temporal dimension is hierarchical; here, BiLSTM captures short- to medium-

range temporally changing aspects, and it also includes continuity of local movements, whereas the Action-Transformer models long-range temporal dependances using global contextual reasoning. The problem of low-level temporal forms is avoided by hierarchical pre-arrangements. It reduces the computational overhead. This alignment is quite consistent with the nature of modern deepfake generation algorithms, which tend to retain high spatial realism without sustained coherent temporal dynamics over long periods of time. By integrating information across distant temporal segments, the proposed approach enhances sensitivity to subtle long-range discrepancies, thereby improving robustness against sophisticated manipulation techniques that are difficult to detect using purely spatial or short-term temporal cues. Figure 2 provides an overview of the Action-Transformer module in the AT-HSTNet architecture.

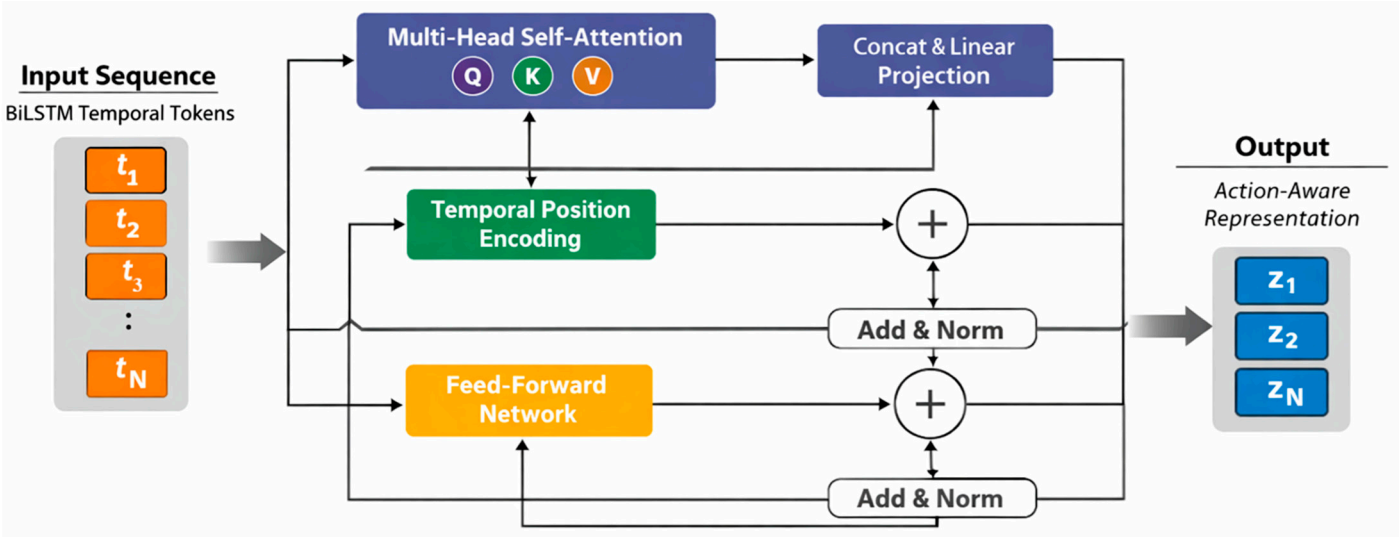

Figure 2. Action-Transformer module for AT-HSTNet architecture.

3.4. Feature Aggregation and Classification

Global Average Pooling (GAP) is used to aggregate the Transformer's output throughout the temporal dimension, producing a compact video-level representation given by Equation (6) [33]:

$$F_{video} = GAP\left(H_{transformer}\right) \tag{6}$$

This representation is processed through a fully connected classification head with dropout regularization to predict whether the input video is real or manipulated.

3.5. Training Strategy and Experimental Configuration

3.5.1. Memory Optimization and Training Strategy

Several memory optimization approaches are used to allow AT-HSTNet to train more efficiently. Gradient accumulation simulates larger batch sizes by accumulating gradients across numerous iterations, as expressed in (7), where ∇_{eff} denotes the effective accumulated gradient used for the model parameter update after gradient accumulation, N represents the number of accumulation steps (i.e., the number of mini-batches over which gradients are accumulated), ∇_i is the gradient computed from the loss of the ith mini-batch during backpropagation, and i indexes each mini-batch within the accumulation window such that averaging simulates a larger effective batch size without increasing GPU memory usage.

$$\nabla_{eff} = \frac{1}{N}\sum_{i=1}^{N}\nabla_i \tag{7}$$

Videos are processed in temporal chunks to reduce peak memory utilization while maintaining temporal continuity [34].

3.5.2. Training Stabilization and Optimization

To promote stability and generalization, an exponential moving average (EMA) of model parameters is kept as

$$\theta_{EMA} = \beta\theta_{EMA} + (1 - \beta)\theta \tag{8}$$

where $\beta = 0.999$. EMA parameters are used during validation and inference. Also, in Equation (8), θ_{EMA} denotes the exponential moving average of model parameters and model parameters are learnable weights and biases of AT-HSTNet architecture.

3.5.3. Experimental Setup and Implementation Details

Experiments were conducted on a workstation configured with an 8th-generation Intel Core i7 processor, 32 GB of RAM, and an NVIDIA GeForce GTX 1650 GPU, representing typical consumer-grade hardware. In both convolutional and temporal layers, a Leaky Rectified Linear Unit (Leaky ReLU) activation function was used to permit gradient flow for negative activations and promote stable optimization. Furthermore, for optimization, an AdamW optimizer with a learning rate of 1×10^{-6} was used for a maximum of 50 epochs; momentum parameter and weight decay were 0.93 and 0.001 to avoid overfitting. The chosen batch size was 64. For stable convergence and improved generalization, a learning rate scheduling strategy was retained where if the validation loss was not approved for 5 successive epochs the learning rate was reduced by 0.1. Moreover, if no remarkable improvement was observed for 10 successive epochs, an early stopping mechanism was applied.

EfficientNet-B0 was pre-trained with ImageNet weights, while the newly added temporal and classification layers were set with Xavier initialization. This was done to ensure that the variance is consistent between each layer and that the gradient flow is not disrupted. The binary cross-entropy function was used as the training function. This is because it is an appropriate function for binary classification of deepfake images, as it calculates the divergence between the predicted probabilities and actual values.

3.5.4. Training Convergence and Performance Metrics

Standard metrics of classification performance, precision, recall, F1-score, and accuracy [35] are therefore used to provide a more all-inclusive evaluation of deepfake detector performance. Precision quantifies the ratio of correctly classified fake videos among all samples identified as fake and is defined by Equations (9)–(11):

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

Recall refers to how well the model can detect manipulated videos and is calculated based on the total number of genuine and fake samples:

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

The F1-score is the harmonic means of precision and recall and provides a balanced view of classification performance when both false positives and false negatives are significant:

$$F1\text{-}Score = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \tag{11}$$

4. Experimental Results and Analysis

This section evaluates the proposed AT-HSTNet model and compares it with two baseline models trained experimentally. These baselines have similar architectural components but vary in their parameter settings and temporal modeling approaches defined in last part of this section. The following analysis assesses the effectiveness of the proposed hierarchical spatiotemporal design in a controlled optimization environment.

4.1. Training and Validation Performance Analysis

Overall classification accuracy is used to measure correct real and fake predictions and to compare AT-HSTNet with baseline models. Figure 3 shows that the training and validation accuracy increase steadily and remain close, indicating stable training and good generalization, and 99.2% and 98.7% accuracies are achieved for training and validation, respectively.

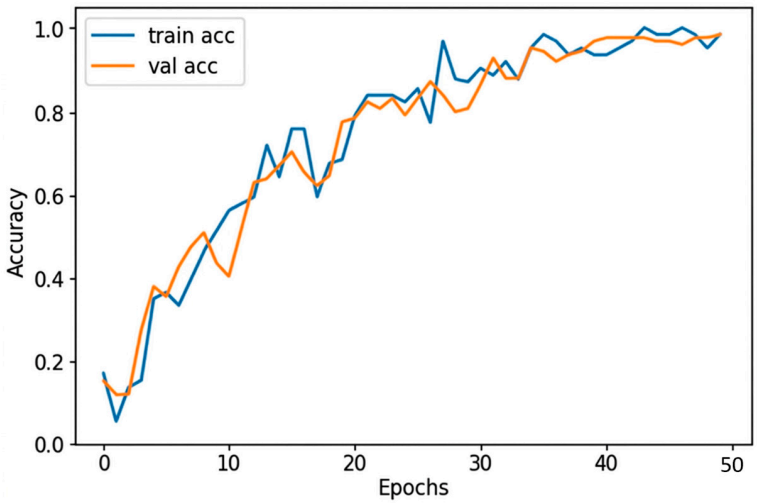

Figure 3. Training and validation accuracy curves of AT-HSTNet.

Figure 4 shows AT-HSTNet's training and validation loss curves. Both losses decrease rapidly in the initial epoch and eventually converge smoothly, indicating stable optimization dynamics. The minor and identical gap between the two curves indicates a minimal generalization error. By the last epoch, both losses stabilized at around 0.215, showing reliable and consistent model convergence.

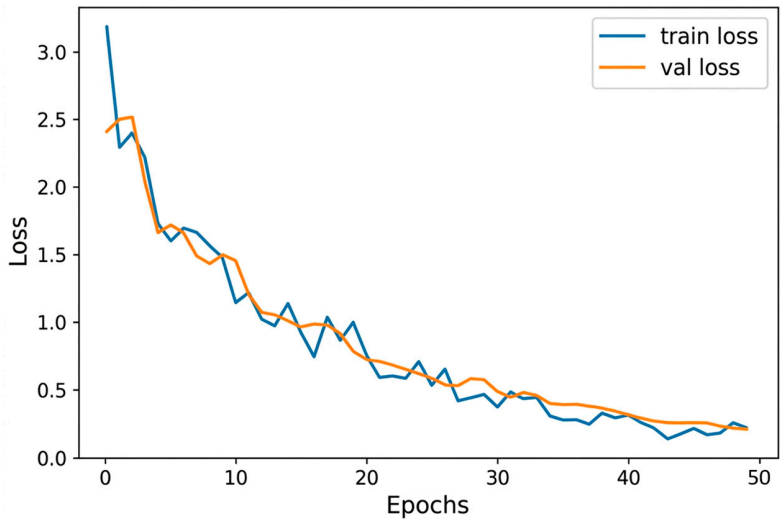

Figure 4. Training and validation loss curves of AT-HSTNet.

Figure 5 shows the training and validation recall curves of AT-HSTNet. Recall increases steadily, indicating improved sensitivity to manipulated videos. Minor fluctuations appear in the validation curve during intermediate epochs, but both curves converge toward higher values, with validation recall stabilizing at approximately 96%, demonstrating strong detection performance and good generalization.

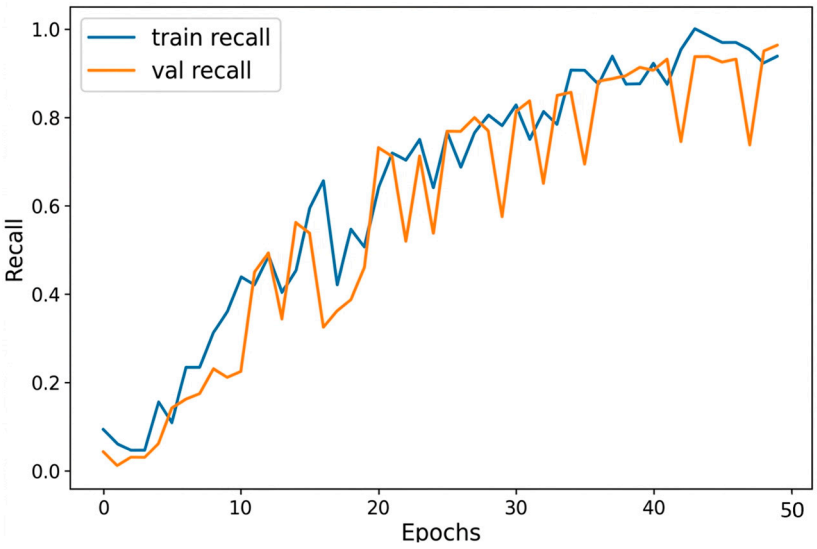

Figure 5. Training and validation recall curves of AT-HSTNet.

Figure 6 shows the training and validation precision curves of AT-HSTNet at different epochs of training. Precision increases substantially at the beginning of training and remains consistently high, reflecting a low false positive rate. The validation precision converges to around 98%, indicating reliable discrimination between real and fake videos.

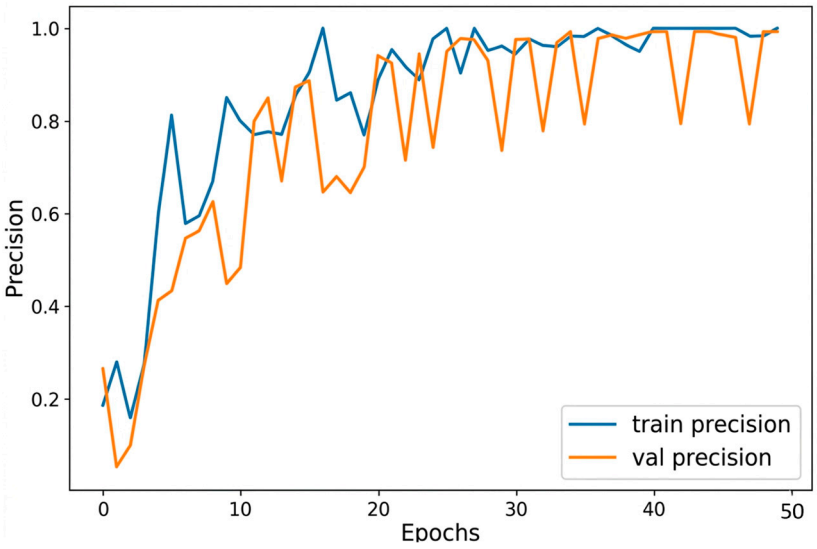

Figure 6. Training and validation precision curves of AT-HSTNet.

Figure 7 shows the training and validation F1-score trajectories of AT-HSTNet over epochs. Both curves are characterized by a steady rise corresponding to a proportional improvement in precision and recall. Minor fluctuations are noted in the validation curve during intermediate epochs; however, both curves converge to high values in the final epochs, with the validation F1-score stabilizing around 97%, indicating stable and robust classification performance.

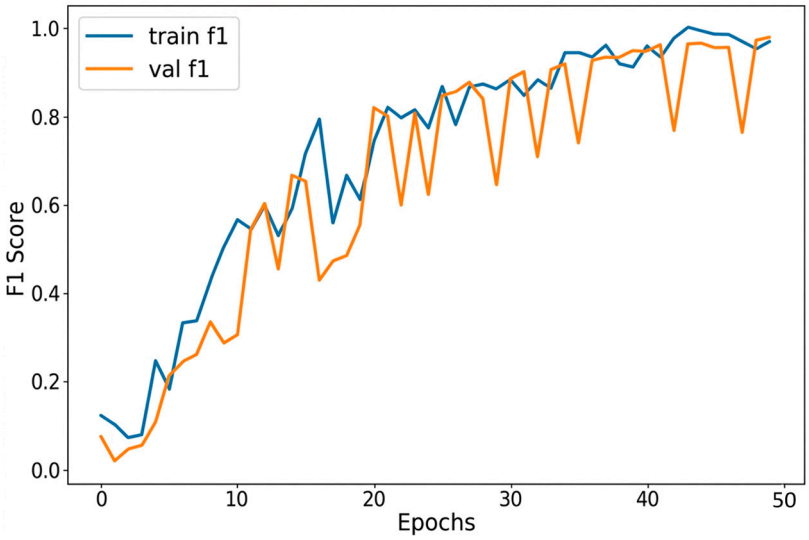

Figure 7. Training and validation F1-score curves of AT-HSTNet.

4.2. Comparative Performance Analysis

The effectiveness of the proposed AT-HSTNet framework was evaluated against two experimentally trained baseline models that share architectural similarities with the proposed approach. Model A uses a CNN–BiLSTM architecture without Transformer-based reasoning, while Model B incorporates a Transformer immediately after spatial feature extraction without performing hierarchical temporal abstraction. Table 2 presents the results obtained from the models under evaluation.

Table 2. Comparative performance and efficiency of AT-HSTNet and baseline models.

Model	Architecture Description	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	GFLOPs	FPS (Single GPU)
Model A	CNN + BiLSTM (No Transformer)	95.4	96.1	93.2	94.6	~0.90	~22
Model B	CNN + Transformer (No Hierarchical Temporal Modeling)	96.8	97.4	94.8	96.1	~1.80	~14
AT-HSTNet (Proposed)	EfficientNet-B0 + BiLSTM + Action-Transformer	98.7	98.0	96.0	96.9	0.45	~30

Table 2 shows that AT-HSTNet consistently outperforms other baseline models across accuracy, precision, recall, F1-score, and AUC while maintaining the lowest computational complexity. Model A captures short-term temporal dependencies well but is limited in modeling long-range temporal reasoning, which constrains its recall and discriminative capability.

Model B improves its performance by incorporating Transformer-based attention; however, the lack of hierarchical temporal abstraction leads to higher computational costs and reduced sensitivity to subtle long-range inconsistencies. The proposed framework operates at approximately 0.45 GFLOPs and achieves an inference throughput of approximately 30 FPS, reflecting a balanced trade-off between detection accuracy and efficiency. These results confirm that hierarchical time-based modeling information with action-aware attention enhances deepfake video detection performance.

4.3. Computational Efficiency

This work emphasizes the computational efficiency of AT-HSTNet, with a focus on deployment of consumer-grade hardware. The proposed framework requires approximately 0.45 GFLOPs, as shown in Table 3, when using EfficientNet-B0 for spatial feature extraction and hierarchical time modeling. EfficientNet-B0 employs mobile inverted bottleneck blocks based on depthwise separable convolution operations, which reduce redundant computation while maintaining strong representational capacity.

Table 3. Computational efficiency comparison of AT-HSTNet and existing methods.

Method	Representative Work	Architecture Type	GFLOPs	FPS (Single GPU)
CNN–RNN Baseline	CNN–LSTM-based detectors [16]	CNN + LSTM	~0.90	~22
CNN–ViT Hybrid	HCiT [22], Hybrid Transformer Network [22]	CNN + Vision Transformer	~1.8	~14
Spatiotemporal Transformer	ISTVT [23], Video Transformer with UV alignment [24]	Video Transformer	~3.5	~8
Swin-based Transformer	SFormer [25]	Windowed Video Transformer	~2.4	~11
AT-HSTNet (Proposed)	Proposed Architecture	CNN + BiLSTM + Action-Transformer	0.45	~30

Temporal processing is hierarchical, such that the BiLSTM layers capture short- and mid-range dependencies, while the Action-Transformer operates on temporally summarized representations rather than raw frame-level features. This substantially reduces dense self-attention over long frame sequences, reducing the memory and computational cost significantly. Further efficiency comes from temporal chunking, gradient accumulation, and mixed-precision training, which together reduce peak memory consumption during training and inference. As a result, AT-HSTNet achieves an average throughput of about 30 frames per second on an NVIDIA GeForce GTX 1650 GPU, making it possible to perform deepfake detection with good performance without specialized hardware acceleration. This means that long-range temporal reasoning can be achieved without compromising runtime efficiency.

Table 3 compares AT-HSTNet to state-of-the-art CNN–Transformer and Transformer-only models for deepfake detection with a focus on computational complexity and runtime efficiency. Values report single-stream inference using uniform input resolutions and sequence lengths. It can be observed that AT-HSTNet exhibits an efficiency advantage relative to Transformer-only spatiotemporal networks. In sharp contrast, AT-HSTNet restricts Transformer operations to a compact sequence of BiLSTM-encoded temporal tokens and thus mitigates attention complexity while maintaining sensitivity to long-range time-based inconsistencies. The proposed AT-HSTNet therefore offers competitive or superior detection accuracy while considerably reducing the computational cost and improving the inference speed, making it well-suited for real-world deepfake detection applications that require efficient deployment and near real-time performance.

4.4. Analysis of Temporal Module Design

Balance between temporal modeling capacity and computational efficiency is the reason for selecting BiLSTM and Action-Transformer modules. To capture the short-range and mid-range temporal dependencies without excessive model complexity, the BiLSTM section is designed. BiLSTM with an increased number of layers and hidden dimensions

was recorded to provide limited performance and increased computational cost, making it less suitable for deepfake recognition.

Instead of using raw frame-level features, the Action-Transformer uses temporally summarized representations generated by the BiLSTM for long-range temporal modelling. This hierarchical structure lessens the quadratic complexity of self-attention and decreases the effective sequence length. Table 2 shows that directly implementing Transformer-based attention (Model B) results in a significant increase in computational cost (~1.80 GFLOPs) with only a moderate improvement in performance. The effectiveness of hierarchical temporal abstraction is demonstrated by the suggested AT-HSTNet, which achieves superior performance (98.7% accuracy) with significantly lower computational cost (0.45 GFLOPs).

To guarantee adequate representational capacity and prevent needless computational overhead, the number of transformer layers and attention heads was chosen. Effective long-range temporal reasoning is made possible by this design without sacrificing real-time performance.

5. Conclusions and Future Directions

This paper presents AT-HSTNet, a hierarchical deepfake video detection framework that integrates EfficientNet for frame-level feature extraction, BiLSTM-based sequence modeling, and an action-aware Transformer for long-range time reasoning. By explicitly separating short- and mid-range time modeling from global dependency analysis, the proposed framework effectively captures fine-grained time inconsistencies that are difficult to detect using frame-based or short-window approaches. Experimental results on the FFIW-10K dataset demonstrate that AT-HSTNet consistently outperforms baseline methods across all evaluation metrics, achieving 98.7% classification accuracy and a 96.9% F1-score. In addition, the proposed model maintains a low computational cost of 0.45 GFLOPs and achieves an inference speed of approximately 30 FPS, enabling effective operation on consumer-grade hardware. These results confirm that hierarchical time modeling combined with action-aware attention provides a reliable and efficient solution.

Key contributions include the following:

- A hierarchical deepfake video detection framework, AT-HSTNet, is introduced to explicitly separate short- and medium-range time modeling from long-range sequence reasoning, enabling robust capture of frame-level visual artifacts and multi-scale time inconsistencies.
- An action-aware Transformer module is introduced to perform long-range time reasoning on BiLSTM-encoded features instead of raw frame features, reducing redundant attention computation and improving training stability compared with conventional CNN–Transformer designs.
- A lightweight spatial feature extraction strategy based on EfficientNet-B0 is introduced to effectively balance detection accuracy and computational efficiency for fine-grained facial artifact analysis.
- A memory-efficient training and optimization framework is developed, incorporating sequence-level MixUp, frame-level Random Erasing, and stabilization techniques to enable efficient training and real-time inference on consumer-grade hardware.

Despite the strong performance achieved by AT-HSTNet, this study has several limitations. In particular, the experimental evaluation is restricted to a single dataset, which limits the assessment of the model's generalization ability across different techniques and data distributions. Additional evaluations on multiple large-scale and diverse datasets, such as FaceForensics++, Celeb-DF, and DFDC, are therefore required to more comprehensively evaluate the robustness of the proposed framework under varying real-world deepfake generation conditions. Second, a detailed component-wise analysis is required

to separately assess the contributions of the BiLSTM and the action-aware Transformer modules. In addition, the applicability of the proposed framework to online or streaming video scenarios where low latency and incremental inference are required has not yet been investigated, and the integration of multimodal information, such as audio and textual indications, can be investigated to provide more comprehensive deepfake detection.

Author Contributions: Conceptualization, S.J.; data curation, S.J., M.C.E.R., and A.E.; formal analysis, S.J. and M.C.E.R.; investigation, S.J. and O.A.-K.; methodology, S.J.; project administration, M.C.E.R., O.A.-K., and M.E.B.; visualization, S.J. and A.B.F.; writing—original draft, S.J., M.C.E.R., and A.E.; writing—review and editing, O.A.-K., A.B.F., and M.E.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not Applicable.

Informed Consent Statement: Not Applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/tfzhou/FFIW, accessed on 14 December 2025. The implementation of the proposed architecture will be made available upon request.

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

Abbreviations

The following abbreviations are used in this manuscript:

3D	3-dimensional
AT-HSTNet	Action-Transformer-based Hierarchical Spatiotemporal Network
AUC	Area Under the Curve
BiLSTM	Bidirectional Long Short-Term Memory
Celeb-DF	Celebrity DeepFake
CNN	Convolutional Neural Network
DFDC	Deepfake Detection Challenge
EMA	Exponential Moving Average
FF	Face Forensics
FFIW	Face Forensics in Wild
FN	False Negative
FP	False Positive
FPS	Frames Per Second
GAN	Generative Adversarial Network
GAP	Global Average Pooling
GB	Giga Bytes
GFLOPs	Giga Floating Point Operations Per Seconds
GPU	Graphical Processing Unit
HCiT	Hybrid Convolutional Neural Network
ISTVT	Interpretable Spatial–Temporal Video Transformer
Leaky RELU	Leaky Rectified Linear Unit
LSTM	Long Short-Term Memory
RNN	Recurrent Neural Network
SForms	Swine-based Transformers
TP	True Positive
UV	Ultraviolet
ViT	Vision Transformer

References

1. Sharma, V.K.; Garg, R.; Caudron, Q. A systematic literature review on deepfake detection techniques. *Multimed. Tools Appl.* **2025**, *84*, 22187–22229. [CrossRef]
2. Li, M.; Ahmadiadli, Y.; Zhang, X.-P. A Survey on Speech Deepfake Detection. *ACM Comput. Surv.* **2025**, *57*, 165. [CrossRef]
3. Heidari, A.; Navimipour, N.J.; Dag, H.; Unal, M. Deepfake detection using deep learning methods: A systematic and comprehensive review. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2024**, *14*, e1520. [CrossRef]
4. Yan, Z.; Yao, T.; Chen, S.; Zhao, Y.; Fu, X.; Zhu, J.; Luo, D.; Wang, C.; Ding, S.; Wu, Y.; et al. Df40: Toward next-generation deepfake detection. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 29387–29434.
5. Concas, S.; La Cava, S.M.; Casula, R.; Orru, G.; Puglisi, G.; Marcialis, G.L. Quality-based artifact modeling for facial deepfake detection in videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 3845–3854. Available online: https://openaccess.thecvf.com/content/CVPR2024W/DFAD/html/Concas_Quality-based_Artifact_Modeling_for_Facial_Deepfake_Detection_in_Videos_CVPRW_2024_paper.html (accessed on 14 December 2025).
6. Le, B.M.; Kim, J.; Woo, S.S.; Moore, K.; Abuadbba, A.; Tariq, S. SoK: Systematization and Benchmarking of Deepfake Detectors in a Unified Framework. *arXiv* **2025**, arXiv:2401.04364. [CrossRef]
7. Zafar, F.; Khan, T.A.; Akbar, S.; Ubaid, M.T.; Javaid, S.; Kadir, K.A. A Hybrid Deep Learning Framework for Deepfake Detection Using Temporal and Spatial Features. *IEEE Access* **2025**, *13*, 79560–79570. [CrossRef]
8. Al Redhaei, A.; Fraihat, S.; Al-Betar, M.A. A self-supervised BEiT model with a novel hierarchical patchReducer for efficient facial deepfake detection. *Artif. Intell. Rev.* **2025**, *58*, 278. [CrossRef]
9. AlMuhaideb, S.; Alshaya, H.; Almutairi, L.; Alomran, D.; Alhamed, S.T. LightFakeDetect: A Lightweight Model for Deepfake Detection in Videos That Focuses on Facial Regions. *Mathematics* **2025**, *13*, 3088. [CrossRef]
10. Wang, Z.; Cheng, Z.; Xiong, J.; Xu, X.; Li, T.; Veeravalli, B.; Yang, X. A Timely Survey on Vision Transformer for Deepfake Detection. *arXiv* **2024**, arXiv:2405.08463. [CrossRef]
11. Cantero-Arjona, P.; Sánchez-Macián, A. Deepfake Detection and the Impact of Limited Computing Capabilities. *arXiv* **2024**, arXiv:2402.14825. [CrossRef]
12. Ain, Q.U.; Ning, H.; Philipo, A.G.; Daneshmand, M.; Ding, J. Beyond Accuracy: A Deployment-Oriented Benchmark of Deepfake Detection Models. *TechRxiv* **2025**, *18*. [CrossRef] [PubMed]
13. Patel, Y.; Tanwar, S.; Bhattacharya, P.; Gupta, R.; Alsuwian, T.; Davidson, I.E.; Mazibuko, T.F. An improved dense CNN architecture for deepfake image detection. *IEEE Access* **2023**, *11*, 22081–22095. [CrossRef]
14. Tipper, S.; Atlam, H.F.; Lallie, H.S. An investigation into the utilisation of CNN with LSTM for video deepfake detection. *Appl. Sci.* **2024**, *14*, 9754. [CrossRef]
15. Al-Dulaimi, O.A.H.H.; Kurnaz, S. A hybrid CNN-LSTM approach for precision deepfake image detection based on transfer learning. *Electronics* **2024**, *13*, 1662. [CrossRef]
16. Petmezas, G.; Vanian, V.; Konstantoudakis, K.; Almaloglou, E.E.I.; Zarpalas, D. Video deepfake detection using a hybrid CNN-LSTM-Transformer model for identity verification. *Multimed. Tools Appl.* **2025**, *84*, 40617–40636. [CrossRef]
17. Ikram, S.T.; Chambial, S.; Sood, D. A performance enhancement of deepfake video detection through the use of a hybrid CNN Deep learning model. *Int. J. Electr. Comput. Eng. Syst.* **2023**, *14*, 169–178. [CrossRef]
18. Kaddar, B.; Fezza, S.A.; Akhtar, Z.; Hamidouche, W.; Hadid, A.; Serra-Sagristá, J. Deepfake Detection Using Spatiotemporal Transformer. *ACM Trans. Multimed. Comput. Commun. Appl.* **2025**, *20*, 345. [CrossRef]
19. Heo, Y.-J.; Yeo, W.-H.; Kim, B.-G. DeepFake detection algorithm based on improved vision transformer. *Appl. Intell.* **2023**, *53*, 7512–7527. [CrossRef]
20. Khormali, A.; Yuan, J.-S. DFDT: An End-to-End DeepFake Detection Framework Using Vision Transformer. *Appl. Sci.* **2022**, *12*, 2953. [CrossRef]
21. Wang, T.; Cheng, H.; Chow, K.P.; Nie, L. Deep Convolutional Pooling Transformer for Deepfake Detection. *ACM Trans. Multimed. Comput. Commun. Appl.* **2023**, *19*, 174. [CrossRef]
22. Khan, S.A.; Dang-Nguyen, D.-T. Hybrid Transformer Network for Deepfake Detection. In Proceedings of the International Conference on Content-based Multimedia Indexing, Dublin, Ireland, 22–24 October 2022; ACM: Graz, Austria, 2022; pp. 8–14. [CrossRef]
23. 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]
24. Khan, S.A.; Dai, H. Video Transformer for Deepfake Detection with Incremental Learning. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event, China, 20–24 October 2021; ACM: Graz, Austria, 2021; pp. 1821–1828. [CrossRef]
25. Kingra, S.; Aggarwal, N.; Kaur, N. SFormer: An end-to-end spatio-temporal transformer architecture for deepfake detection. *Forensic Sci. Int. Digit. Investig.* **2024**, *51*, 301817. [CrossRef]

26. Javed, M.; Zhang, Z.; Dahri, F.H.; Laghari, A.A.; Krajčík, M.; Almadhor, A. Real-Time Deepfake Detection via Gaze and Blink Patterns: A Transformer Framework. *Comput. Mater. Contin.* **2025**, *85*, 1457–1493. [CrossRef]
27. Zhou, T. tfzhou/FFIW. 26 September 2025. Available online: https://github.com/tfzhou/FFIW (accessed on 14 December 2025).
28. Huang, J.; Yang, P.; Xiong, B.; Lv, Y.; Wang, Q.; Wan, B.; Zhang, Z.-Q. Mixup-based data augmentation for enhancing few-shot SSVEP detection performance. *J. Neural Eng.* **2025**, *22*, 046038. [CrossRef]
29. Ma, G.; Wang, Z.; Yuan, Z.; Wang, X.; Yuan, B.; Tao, D. A comprehensive survey of data augmentation in visual reinforcement learning. *Int. J. Comput. Vis.* **2025**, *133*, 7368–7405. [CrossRef]
30. Kumar, A.; Yadav, S.P.; Kumar, A. An improved feature extraction algorithm for robust Swin Transformer model in high-dimensional medical image analysis. *Comput. Biol. Med.* **2025**, *188*, 109822. [CrossRef]
31. Chen, X.; Liu, C.; Xia, H.; Chi, Z. Burn-through point prediction and control based on multi-cycle dynamic spatio-temporal feature extraction. *Control Eng. Pract.* **2025**, *154*, 106165. [CrossRef]
32. Zhang, Y.; Liu, K.; Zhang, J.; Huang, L. Self-attention mechanism network integrating spatio-temporal feature extraction for remaining useful life prediction. *J. Electr. Eng. Technol.* **2025**, *20*, 1127–1142. [CrossRef]
33. Li, L.; Xu, M.; Chen, S.; Mu, B. An adaptive feature fusion framework of CNN and GNN for histopathology images classification. *Comput. Electr. Eng.* **2025**, *123*, 110186. [CrossRef]
34. Xiao, J.; Sang, S.; Zhi, T.; Liu, J.; Yan, Q.; Luo, L.; Yuan, B. Coap: Memory-efficient training with correlation-aware gradient projection. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 10–17 June 2025; pp. 30116–30126.
35. Ubaid, M.T.; Javaid, S. Precision Agriculture: Computer Vision-Enabled Sugarcane Plant Counting in the Tillering Phase. *J. Imaging* **2024**, *10*, 102. [CrossRef] [PubMed]