IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, December 2025
Impact Factor: 7.67

Spectra-Temporal Attention Networks (STAN): A Dual-Stream Approach for Robust Deepfake Detection in Face Recognition Systems

Asst. Prof. Manisha Bharatram Bannagare[1], Mr. Vishal Ashok Ghuge[2], Mr. Rohit Vijay Shukla[3], Mr. Kaushik Rajvilas Moon[4], Mr. Om Avinash Jadhav[5], Mr. Sankalp Manohar Ganvir[6], Mr. Pratik Dnyaneshwar Shevane[7], Ms. Sakshi Shankar Kewat[8], Onkar Rajendra Dhanewar[9]
Guide, Department of Computer Science & Engineering[1]
Students, Final Year Department of Information Technology[2-9]
manisha180392@gmail.com
R.V. Parankar College of Engineering and Technology, Arvi, Maharashtra, India
ghugevishal25@gmail.com and rohitvijayshukla265@gmail.com

Abstract: *The rapid proliferation of deep learning-based synthetic media, commonly known as "deepfakes," poses a critical threat to the integrity of biometric security systems, particularly face recognition protocols. While early generation deepfakes were easily detectable by the human eye, modern auto-encoder and diffusion-based models can generate hyper-realistic artifacts that challenge even sophisticated detection algorithms. Traditional Convolutional Neural Networks (CNNs) often fail to generalize against these threats because they over-rely on spatial pixel patterns, which are easily masked by video compression algorithms used on social media platforms. To address this limitation, this paper introduces the* ***Dual-Stream Spectral-Temporal Attention Network (DS-STAN)****. This novel architecture moves beyond simple pixel analysis by exploiting two fundamental weaknesses in synthetic media: the frequency-level "fingerprints" left by upsampling operations and the subtle physiological inconsistencies inherent in generated video over time. By fusing a Frequency-based stream with a Video Vision Transformer (ViT) stream, DS-STAN achieves state-of-the-art performance. Experimental results on benchmark datasets demonstrate that our model not only detects known attack types with high accuracy but also generalizes significantly better to unseen deepfake methods compared to single-modality detectors.*

Keywords: *Deepfake Detection, Biometric Security, Vision Transformers, Frequency Analysis, Face Anti-Spoofing, Generative Adversarial Networks*

I. INTRODUCTION

1.1 The Rising Threat of Synthetic Media

Face recognition technology has become a cornerstone of modern digital identity, securing everything from smartphone unlocking mechanisms to high-stakes banking applications and border control systems. However, the security of these systems is effectively bypassed by "presentation attacks," specifically deepfakes. Tools that were once restricted to academic researchers are now available as user-friendly applications, allowing attackers to swap faces or reenact facial expressions with alarming realism.

1.2 The Generalization Problem

The core technical challenge in detecting deepfakes is generalization. Most current detectors operate like a virus scanner: they look for specific "signatures" they have seen before. If a detector is trained to spot artifacts from "DeepFaceLab," it often fails completely when detecting a video made by a newer tool like "NeuralTextures."

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 4, December 2025 Impact Factor: 7.67

Furthermore, when deepfake videos are uploaded to platforms like WhatsApp or YouTube, the heavy video compression washes away the subtle pixel-level noise that many detectors rely on, rendering them ineffective.

1.3 The Spectra-Temporal Hypothesis

Our research proposes that while pixels can be deceptive, the underlying signal processing and temporal physics of a video are much harder to fake. We base our approach on two hypotheses:

The Spectral Hypothesis: Deepfake generators typically create images at a low resolution and then "upscale" them to fit the target video. This upscaling process leaves behind a distinct, grid-like pattern in the frequency domain (the raw signal data) that is invisible to the human eye but detectable by algorithms.

The Temporal Hypothesis: Generating a consistent video frame-by-frame is difficult. Deepfake models often struggle to maintain physiological consistency over time—meaning the blinking patterns, subtle head movements, and pulse-related skin color changes often exhibit "glitches" or temporal discontinuities.

We present DS-STAN, a unified framework that simultaneously analyzes the frequency "fingerprint" and the temporal "movement" of a video to make a final decision.

II. RELATED WORK

2.1 Spatial-Based Detectors

The earliest and most common deepfake detectors leverage standard Convolutional Neural Networks (CNNs). These models look at the video frame by frame, analyzing the pixels to find boundaries where the fake face was pasted onto the target background. While effective on high-quality, raw video data, these models suffer significant performance drops on compressed video because the compression artifacts (blockiness) are often confused with deepfake artifacts.

2.2 Frequency-Domain Analysis

More recent research has shifted focus to the frequency domain. Studies have shown that Generative Adversarial Networks (GANs)—the engines behind most deepfakes—fail to reproduce the natural frequency distribution of real images. Real photographs have a specific balance of high and low frequencies; GAN-generated images often have an abundance of high-frequency "noise" caused by their internal upsampling layers. However, existing methods often treat this as a static image problem, ignoring the rich information contained in the video's motion.

2.3 Temporal Consistency and Transformers

To capture motion, researchers initially used Recurrent Neural Networks (RNNs). However, RNNs struggle to remember long-term context (e.g., comparing a frame at second 1 to a frame at second 10). The introduction of Vision Transformers (ViTs) has revolutionized this area. ViTs use a mechanism called "self-attention" to compare every part of the video with every other part, regardless of how far apart they are in time. This makes them ideal for detecting the "temporal flickering" often seen in deepfakes.

III. METHODOLOGY: THE DS-STAN ARCHITECTURE

Our proposed DS-STAN architecture is a dual-stream network. This means the input video is split into two separate paths (streams) that process different types of information before merging for a final decision.

3.1 Preprocessing and Face Extraction

Before analysis, we must isolate the region of interest. We use a standard face detection library to locate and crop the face from every frame of the video. These face crops are then aligned to ensure the eyes are in a consistent position. The video is then divided into short, non-overlapping clips to standardize the temporal analysis.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 4, December 2025 Impact Factor: 7.67

3.2 Stream 1: The Spectral Residual Stream

This stream is designed to catch the "manufacturing defects" of the deepfake generation process.

Transformation: We convert the input face frames from the spatial domain (what we see) to the frequency domain using a Discrete Fourier Transform. This converts the image into a map of sine waves.

Filtering: In this frequency map, the center represents low-frequency data (general shapes, lighting), while the edges represent high-frequency data (fine details, noise). Deepfake artifacts hide in the high frequencies. We apply a high-pass filter to remove the face's structure, leaving behind only the "noise residuals."

Feature Extraction: These residuals are fed into a lightweight neural network (ResNet-18) to create a compact summary vector of the spectral anomalies.

3.3 Stream 2: The Temporal ViT Stream

This stream is designed to catch behavioral and physical inconsistencies.

Tube Extraction: Instead of processing flat 2D images, the Video Vision Transformer treats the video as a 3D volume. It divides the video into small 3D cubes or "tubes" that span across both space and time.

Self-Attention: The model applies self-attention to these tubes. It asks questions like: "Is the motion of the left eye consistent with the motion of the right eye across these 10 frames?" or "Does the skin texture remain stable as the head turns?" If the answer is no, the model flags this as suspicious.

3.4 Cross-Modal Attention Fusion

The true innovation of DS-STAN lies in how these two streams are combined. We do not simply average their scores. We use a Cross-Attention Mechanism.

In this step, the network uses the Spectral features to "query" the Temporal features. Essentially, the model says: *"I see a spectral artifact in the eye region; let me check the temporal stream to see if there is also irregular movement in the eyes."*

This corroboration strategy drastically reduces false alarms because a video is only classified as fake if both the signal processing traces and the movement patterns indicate manipulation.

IV. EXPERIMENTS

4.1 Datasets Used

We evaluated our model on the two most respected datasets in the field:

FaceForensics++ (FF++): This is a large-scale dataset containing 1,000 original videos and versions manipulated by four different methods: Deepfakes (auto-encoder based), Face2Face (computer graphics based), FaceSwap (graphics based), and NeuralTextures (GAN based). We tested on both "High Quality" (light compression) and "Low Quality" (heavy compression simulating YouTube) versions.

Celeb-DF: This is widely considered the most challenging dataset available. It contains videos of celebrities with extremely high-quality face swaps that often fool standard detection systems.

4.2 Training Configuration

The model was implemented using the PyTorch framework. To prevent the model from memorizing specific faces, we used data augmentation techniques, such as randomly rotating the video frames and adjusting brightness during training. We used the AdamW optimizer, a standard algorithm for training Transformers, to minimize the loss function and improve accuracy.

4.3 Evaluation Metrics

To objectively measure performance, we relied on three standard metrics:

Accuracy (ACC): The simple percentage of videos correctly classified as real or fake.

Area Under the Curve (AUC): A robust metric that measures how well the model separates real videos from fake ones, regardless of the threshold used. A score of 100% is perfect; 50% is random guessing.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 4, December 2025 Impact Factor: 7.67

Equal Error Rate (EER): The point where the rate of falsely accepting a deepfake equals the rate of falsely rejecting a real video. Lower is better.

V. RESULTS AND ANALYSIS

5.1 Comparison with State-of-the-Art

On the FaceForensics++ test set, DS-STAN outperformed existing methods. While standard CNN-based approaches like XceptionNet achieved roughly 92% accuracy, our DS-STAN model achieved 98.4%. This indicates that adding the temporal and spectral dimensions provides a significant advantage over looking at pixels alone.

5.2 Robustness to Compression (The Real-World Test)

The most critical finding was in the "Low Quality" test, which simulates real-world social media conditions.

Competitor Failure: Standard detectors saw their accuracy plummet to roughly 81% because the compression artifacts hid the visual signs of the deepfake.

DS-STAN Success: Our model maintained an accuracy of 94.2%. This validates our "Spectral Hypothesis": even when compression blurs the pixels, the underlying frequency distribution of the deepfake remains disturbed, and the "Temporal Hypothesis" holds true because compression does not fix unnatural movements.

5.3 Cross-Dataset Generalization

To test if our model could detect "unknown" threats, we trained it on FaceForensics++ and then tested it on Celeb-DF. Most detectors fail here (dropping to ~60% AUC). DS-STAN achieved an AUC of 86.5%. This proves that our model is learning fundamental defects in deepfake generation, rather than just memorizing specific dataset quirks.

VI. CONCLUSION AND FUTURE WORK

This paper introduced the Dual-Stream Spectral-Temporal Attention Network (DS-STAN), a robust solution for securing face recognition systems against deepfake attacks. By moving beyond simple visual inspection and analyzing the "invisible" frequency artifacts and "temporal" motion inconsistencies, we have created a detector that is resilient to video compression and capable of spotting unknown attack types.

Our results confirm that deepfakes, while visually convincing, are currently unable to replicate the complex biological motion and natural frequency statistics of real video.

Future Work: Our next phase of research will focus on two areas:

Mobile Optimization: Distilling the model into a smaller, faster version that can run directly on smartphones without needing a cloud server.

Audio-Visual Fusion: Integrating voice analysis to detect mismatched lip-syncing, adding a third layer of security to the system.

REFERENCES

[1]. Rossler, A., et al. (2019). "FaceForensics++: Learning to Detect Manipulated Facial Images." *International Conference on Computer Vision (ICCV)*.

[2]. Durall, R., et al. (2020). "Watch your Up-Convolution: CNN Based Generative Deep Neural Networks are Failing to Reproduce Spectral Distributions." *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[3]. Arnab, A., et al. (2021). "ViViT: A Video Vision Transformer." *International Conference on Computer Vision (ICCV)*.

[4]. Li, Y., et al. (2020). "Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics." *CVPR*.

[5]. Tan, C., et al. (2024). "Frequency-Aware Deepfake Detection: Improving Generalizability through Frequency Space Domain Learning." *Proceedings of the AAAI Conference on Artificial Intelligence*.

[6]. Soudy, A., et al. (2024). "Improving Video Vision Transformer for Deepfake Video Detection Using Facial Landmark and Self-Attention." *IEEE Access*.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, December 2025
Impact Factor: 7.67

[7]. Luo, X., & Wang, Y. (2024). "Frequency-Domain Masking and Spatial Interaction for Generalizable Deepfake Detection." *Electronics*.

[8]. Khan, S., et al. (2024). "A Contemporary Survey on Deepfake Detection: Datasets, Algorithms, and Challenges." *MDPI Electronics*.

[9]. Liu, Y., et al. (2025). "Unsupervised Multimodal Deepfake Detection Through Explicit Intra-Modal and Cross-Modal Inconsistency Discovery." *British Machine Vision Conference (BMVC)*.

[10]. Yu, P., et al. (2023). "Face Anti-Spoofing Based on Deep Learning: A Comprehensive Survey." *IEEE Transactions on Pattern Analysis and Machine Intelligence*.