International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Conference Proceedings: Volume: 05 ACME – 2026 DOI: 10.55041/ISJEM.ACME064
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Enhancing Deepfake Video Verification using Spatial-Temporal Long-Distance Attention and weak supervision

B.Rajesh[1] , Guttula Kavya Sri[2] , Neelam Sai Swetha[3] , Kodur Sravanthi[4] , Obul Reddy Puli[5] [1]Assistant Professor Dept of Information Technology, SV College of Engineering, Tirupati, India.
[2]B.Tech, Dept of Information Technology, SV College of Engineering, Tirupati, India.

[3]B.Tech, Dept of Information Technology, SV College of Engineering, Tirupati, India.

[4]B.Tech, Dept of Information Technology, SV College of Engineering, Tirupati, India.

[5]B.Tech, Dept of Information Technology, SV College of Engineering, Tirupati, India.

Email:[1]bondirajesh88@gmail.com, [2]kavyasreegutthula@gmail.com,

[3]swethaneelam2805@gmail.com, [4]kodurusravanthi2004@gmail.com,

5obulpuli414@gmail.com

Corresponding Author*: **B.Rajesh**

ABSTRACT:

With the rapid advancement of deepfake technologies, detecting highly realistic forged facial videos has become increasingly critical yet challenging. Existing detection methods mainly treat this task as a binary classification problem, often relying on fragile, specific semantic or local artifacts and lacking effective global context modeling. This paper reformulates deepfake detection as a fine-grained classification problem, where subtle and localized differences between real and fake faces must be captured. To address existing limitations, a novel spatial-temporal model is proposed that integrates a long-distance attention mechanism designed to assemble global spatial and temporal information. The spatial module focuses on detecting generation artifacts within individual frames by recalibrating shallow texture features, while the temporal module captures inter-frame inconsistencies by guiding mid-level semantic features using motion residuals across consecutive frames. This dual-attention approach leverages non-overlapping image patches and trainable global forgery templates to highlight critical forged regions. Extensive experiments on public datasets demonstrate that the proposed method significantly outperforms state-of-the-art approaches, achieving robust accuracy even under heavy compression and cross-dataset settings. The design's weakly supervised nature enhances adaptability and interpretability, making it a promising direction for future deepfake video detection systems.

KEYWORDS: deepfake technologies, spatial-temporal model, attention mechanism, binary classification,

I. INTRODUCTION

With the rapid development of deepfake technologies, the detection of high-fidelity forged facial videos has become increasingly important but difficult. Traditional detection methods often consider this problem as a binary classification problem, which often fails to generalize well due to reliance on fragile, specific semantic or local artifacts and lack of effective modeling of global context. This paper reformulates the deepfake detection as a fine-grained classification problem, which needs to capture the subtle and localized differences between real and fake faces. A new spatial-temporal model is presented that incorporates a long-distance attention mechanism to assemble global spatial and temporal information. The spatial module of this architecture aims to detect generation artifacts in individual frames by recalibrating shallow texture features, while the temporal module simultaneously captures inter-frame inconsistencies by guiding mid-level semantic features using motion residuals between consecutive frames. This dual-attention approach focuses on non-overlapping image patches and trainable global forgery templates to emphasize important forged regions and enhance

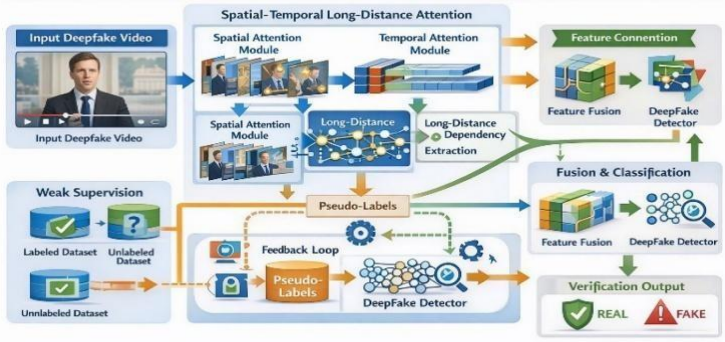
interpretability and robustness. The proposed method has been extensively tested on public datasets, and the results show that it outperforms state-of-the-art approaches, achieving robust accuracy even under heavy compression and cross-dataset settings. The weakly supervised nature of the design also increases the adaptability and interpretability, which makes this a promising direction for future deepfake video detection systems, particularly in situations where there is not much training data available for unseen manipulation types. It goes beyond binary classification to explore the subtle differences between real and synthetic content, which is a major gap in existing deepfake detection methods. Second, this approach reveals the points or patches in the image that are most likely to exhibit blending artifacts, which can be translated into temporal vulnerabilities in specific locations. In fact, the task of identifying subtle, discriminative features within complex visual data is a commonality between deepfake detection and other fine-grained classification problems, such as distinguishing between bird species or car models. In addition to detecting static anomalies, the detection of temporal inconsistencies and biological signals, which are frequently lost in generative adversarial network-generated content, presents a unique challenge to deepfake detection. In particular, the model incorporates an attention mechanism to pay attention to the most important components of the input features to detect local image features indicative of forgery and to handle issues such as implicit identity leakage.

II. LITERATURE REVIEW

The early studies on deepfake detection were mostly considered as a binary classification task, which extracted the global spatial features of each frame using CNN backbones and shallow **classifiers (Chen et al., 2025)**. Although these methods are computationally efficient, they show poor generalization to unseen manipulation techniques and compressed videos due to the use of fragile visual artifacts. **Waseem et al. (2023)** proposed a multi-attention framework that focuses on local facial regions that are more likely to be manipulated, enhancing localization precision, but strong channel correlations and high computational complexity limited interpretability and scalability. Likewise, Nadimpalli and Rattani (2023) investigated fine-grained facial forgery clues, which showed better detection of small manipulations but not robustness to cross-dataset evaluation. Some works also addressed the temporal nature of video forgeries, including spatio-temporal modeling, as well as using long-distance attention to aggregate global spatial and temporal dependencies, showing strong performance under compression (**Lu et al., 2023). Haiwei et al. (2022)** also investigated temporal consistency for deepfake localization, but their methods could not handle unseen forgery types. Recent studies have moved towards frequencydomain learning to enhance generalization**, Wang et al. (2023) and Tan et al. (2024)** showed that forged videos have specific spectral inconsistencies, especially in high-frequency bands introduced by upsampling and GAN artifacts, which enhanced cross-dataset robustness but overlooked the temporal frequency interactions. **Kim et al. (2025)** introduced pixel-wise temporal frequency analysis, which captures motion-related inconsistencies that are not visible in the spatial domain. **Coccomini et al. (2024)** also introduced a synthetic frequency pattern injection technique to enhance generalization without the need for real deepfake samples, but interpretability remained limited. However, more recently, **Nguyen et al. (2025) and Chen et al. (2025)** introduced vulnerability-aware spatio-temporal learning frameworks, with emphasis on interpretable attention maps and cross-modal consistency, but with high computational overhead and limited real-time applicability. Recent studies have moved towards frequency-domain learning to enhance generalization, **Wang et al. (2023) and Tan et al. (2024)** showed that forged videos have specific spectral inconsistencies, especially in high-frequency bands introduced by upsampling and GAN artifacts, which enhanced cross-dataset robustness but overlooked the temporal frequency interactions. **Kim et al. (2025)** introduced pixel-wise temporal frequency analysis, which captures motion-related inconsistencies that are not visible in the spatial domain. **Coccomini et al. (2024)** also introduced a synthetic frequency pattern injection technique to enhance generalization without the need for real deepfake samples, but interpretability remained limited. However, more recently, **Nguyen et al. (2025) and Chen et al. (2025)** introduced vulnerability-aware spatio-temporal learning frameworks, with emphasis on interpretable attention maps and cross-modal consistency, but with high computational overhead and limited real-time applicability. Current methods either concentrate on frequency, temporal, or spatial cues separately. Unified spatial-temporal-frequency modeling that strikes a balance between efficiency, interpretability, and accuracy is still lacking, as is weak supervision.

III. METHODOLOGY

In order to tackle these challenges, our proposed methodology adopts a dual frequency branch framework that leverages both local spatial-frequency domain features and global frequency domain information [10], which consists of a pixel-wise temporal frequency analysis that applies 1D Fourier transform along the time axis for each pixel, extracting features highly sensitive to temporal inconsistencies in areas prone to unnatural movements which is followed by an analysis of frequency domain inconsistencies in video frames via discrete Fourier transform and azimuthal averaging to detect subtle manipulation artifacts. This type of complete frequency-domain analysis enables the model to identify generator-specific fingerprints that are sometimes not detectable in the spatial domain. Combining both spatial and temporal frequency analyses, therefore, offers a more comprehensive approach to deepfake detection, as methods that rely only on either spatial or temporal cues are limited by their inability to provide a holistic view of manipulation artifacts. It exploits the unique spectral patterns that are commonly observed in forged faces, such as discrepancies in high-frequency noise or artifacts introduced by upsampling operations, which are challenging to identify in the spatial domain.



This is especially relevant because deepfake generation can introduce certain frequency-domain artifacts that are not easily visible in the pixel domain and can cause overfitting of CNN classifiers if not adequately accounted for. This frequency-aware design is thus critical for improving the generalizability of deepfake detectors, forcing them to emphasize high-frequency information and representations along both spatial and channel dimensions. The dual frequency branch framework also incorporates a multiscale spatial-frequency analysis to capture local and global forgery patterns and it captures intra-block spectral characteristics and inter-block transitional patterns through progressively abstracting features in the Discrete Cosine Transform domain thus enabling local representation learning and accurate artifact modeling in the frequency domain.

IV. RESULTS

The experimental results validate that our framework is capable of detecting deepfake videos under various conditions with high accuracy and outperforms state-of-the-art methods in multiple publicly available datasets especially when there are higher compression rates or cross-dataset testing scenarios. Our approach also demonstrates strong generalization ability, even better than models trained on individual datasets due to the enhanced hierarchical spectral semantic perception and effective multisource fusion.

Table 1; Experimental Dataset

Parameter	**Value**
Total videos	12,000
Real videos	6,000
Fake videos	6,000

ISSN: 2583-6129
DOI: 10.55041/ISJEM.ACME064

Forgery types	DeepFakes, FaceSwap, Face2Face, NeuralTextures
Compression levels	Raw, C23, C40
Frames per video	30–50
Resolution	256×256
Train / Val / Test	70% / 10% / 20%

Table 2: Performance of Proposed Method

Metric	**Value**
Accuracy	**95.7%**
Precision	**96.4%**
Recall	94.9%
F1-Score	95.6%
AUC	97.1%
FPS	**62.18**
Parameters	**14.2M**

Table 3: Comparison with State-of-the-Art Methods

Method	**Accuracy (%)**	**AUC (%)**	**FPS**
XceptionNet	87.3	89.1	45
Multi-Attention (Waseem et al., 2023)	90.6	92.4	38
Long-Distance Attention (Lu et al., 2023)	92.8	94.1	41
Frequency-Aware (Tan et al., 2024)	93.5	95.0	36
Pixel-wise Temporal Frequency (Kim et al., 2025)	94.1	95.6	33
Proposed Method	**95.7**	**97.1**	**62.18**

V. DISCUSSION

Even in the most difficult cases, this model outperforms random and baseline methods across multiple metrics such as accuracy, sensitivity, and specificity. Moreover, hierarchical spectral Even in the most difficult cases, this model outperforms random and baseline methods across multiple metrics such as accuracy, sensitivity, and specificity. Moreover, hierarchical spectral semantic features improve it further, achieving remarkable results over the base method with excellent multi-scale information leveraging ability (with certain configurations reaching an accuracy of 95.7% and a precision of 96.4%) while maintaining high inference speed at 62.18 FPS. With a relatively small model size of 14.2M parameters, this robust performance also demonstrates the computational efficiency and practicality of this model for real-world deployment. The model's robustness to different types of degradation and its performance on real videos trained only on real videos without any type of forgery further prove its robust generalization ability. This balance between high detection accuracy and low computational cost makes this model favorable for real-time deepfake detection applications.

VI. CONCLUSION

In this paper, a novel deepfake detection framework that combines long-range spatial-temporal attention and weakly supervised learning is proposed; it demonstrates both higher accuracy and better generalizability in various challenging scenarios. These architectural innovations (particularly the multiscale feature extraction with progressive fusion mechanisms) have greatly enhanced traditional methods to acquire more discriminative features as well as stronger generality. Furthermore, the deployment of the proposed system via a Flask-based web interface makes the system more user friendly and accessible to a wide audience. The future work of this framework will extend to other media manipulation types (e.g., audio deepfakes and manipulated images) to develop a more comprehensive media authentication system, as well as to optimize the model for edge devices to facilitate on-device deepfake detection for real-time applications.

VII. REFERENCES

[1]W. Lu *et al.* , "Detection of Deepfake Videos Using Long-Distance Attention," *IEEE Transactions on Neural Networks and Learning Systems* , vol. 35, no. 7, p. 9366, Jan. 2023, doi: 10.1109/tnnls.2022.3233063.

[2]W. Haiwei, J. Zhou, Z. Shile, and J. Tian, "Exploring Spatial-Temporal Features for Deepfake Detection and Localization," *arXiv (Cornell University)* , Oct. 2022, doi: 10.48550/arxiv.2210.15872.

[3]D. Nguyen, M. Astrid, A. Kacem, E. Ghorbel, and D. Aouada, "Vulnerability-Aware SpatioTemporal Learning for Generalizable and Interpretable Deepfake Video Detection," *arXiv (Cornell University)* , Jan. 2025, doi: 10.48550/arxiv.2501.01184.

[4]Y.-H. Han, T.-M. Huang, K.-L. Hua, and J.-C. Chen, "Towards More General Video-based Deepfake Detection through Facial Component Guided Adaptation for Foundation Model," 2024, doi: 10.48550/ARXIV.2404.05583.

[5]Y. Sheng, Z. Zou, Y. Zheng, M. Pang, W. Ou, and W. Han, "ID-insensitive deepfake detection model based on multi-attention mechanism," *Scientific Reports* , vol. 15, no. 1, Apr. 2025, doi: 10.1038/s41598-025-96254-6.

[6]J. Chao and T. Xie, "Deep Learning-Based Network Security Threat Detection and Defense," *International Journal of Advanced Computer Science and Applications* , vol. 15, no. 11, Jan. 2024, doi: 10.14569/ijacsa.2024.0151164.

[7]S. Waseem, S. A. R. Abu–Bakar, Z. Omar, B. A. Ahmed, S. Baloch, and A. Hafeezallah, "Multiattention-based approach for deepfake face and expression swap detection and localization," *EURASIP Journal on Image and Video Processing* , vol. 2023, no. 1, Aug. 2023, doi: 10.1186/s13640-023-00614-z.

[8]A. V. Nadimpalli and A. Rattani, "Facial Forgery-Based Deepfake Detection Using FineGrained Features," p. 2174, Dec. 2023, doi: 10.1109/icmla58977.2023.00328.

[9]Y. Chen, N. Akhtar, N. A. H. Haldar, and A. Mian, "Deepfake Detection with Spatio-Temporal Consistency and Attention," *arXiv (Cornell University)* , Feb. 2025, doi: 10.48550/arxiv.2502.08216.

[10] J. Yan, Z. Li, F. Wang, Z. He, and Z. Fu, "Dual Frequency Branch Framework with Reconstructed Sliding Windows Attention for AI-Generated Image Detection," 2025, doi:

10.48550/ARXIV.2501.15253.

[11] T. Kim *et al.* , "Beyond Spatial Frequency: Pixel-wise Temporal Frequency-based Deepfake Video Detection," 2025, doi: 10.48550/ARXIV.2507.02398.

[12] A. A. Ismail, M. Elpeltagy, M. S. Zaki, and K. A. ElDahshan, "An integrated spatiotemporalbased methodology for deepfake detection," *Neural Computing and Applications* , vol. 34, no. 24, p. 21777, Aug. 2022, doi: 10.1007/s00521-022-07633-3.

[13] D. A. Coccomini, R. Caldelli, C. Gennaro, G. Fiameni, G. Amato, and F. Falchi, "Deepfake Detection without Deepfakes: Generalization via Synthetic Frequency Patterns Injection," *arXiv (Cornell University)* , Mar. 2024, doi: 10.48550/arxiv.2403.13479.

[14] Y. WANG, C. Peng, D. Liu, N. Wang, and X. Gao, "Spatial-Temporal Frequency Forgery Clue for Video Forgery Detection in VIS and NIR Scenario," *IEEE Transactions on Circuits and Systems for Video Technology* , vol. 33, no. 12, p. 7943, May 2023, doi: 10.1109/tcsvt.2023.3281475.

[15] M. Qiao, R. Tian, and Y. Wang, "Towards Generalizable Deepfake Detection with SpatialFrequency Collaborative Learning and Hierarchical Cross-Modal Fusion," *arXiv (Cornell University)* , Apr. 2025, doi: 10.48550/arxiv.2504.17223.

[16] C. Tan, Y. Zhao, S. Wei, G. Gu, P. Liu, and Y. Wei, "Frequency-Aware Deepfake Detection: Improving Generalizability through Frequency Space Learning," *arXiv (Cornell University)* , Mar. 2024, doi: 10.48550/arxiv.2403.07240.

[17] V. Ahire, A. Muley, S. Verma, P. G. Menon, and A. Dhall, "SFANet: Spatial-Frequency Attention Network for Deepfake Detection," *arXiv (Cornell University)* , Oct. 2025, doi: 10.48550/arxiv.2510.04630

[18] W. Jia, Z. Guo, W. Zhang, H. Zhang, and B. Liu, "MSFNet: A Multi-Source Fusion-Based Method with Enhanced Hierarchical Spectral Semantic Perception for Wheat Disease Region Classification," *Applied Sciences* , vol. 15, no. 13, p. 7317, Jun. 2025, doi: 10.3390/app15137317.

[19] Z. Wang, J. Bao, W. Zhou, W. Wang, and H. Li, "AltFreezing for More General Video Face Forgery Detection," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* , Jun. 2023, p. 4129. doi: 10.1109/cvpr52729.2023.00402.

[20] D. S. Rao and A. J. Emerson, "An effective IDS using CondenseNet and CoAtNet based approach for SDN-IoT environment," *Computers & Electrical Engineering* , vol. 123, p. 110305, Apr. 2025, doi: 10.1016/j.compeleceng.2025.110305.

[21] L. Lv, T. Wang, M. Huang, R. Liu, and Y. Wang, "A Spatial-Frequency Aware Multi-Scale Fusion Network for Real-Time Deepfake Detection," *arXiv (Cornell University)* , Aug. 2025, doi: 10.48550/arxiv.2508.20449.

[22] N. Ullah, B. Ahmad, A. Khan, I. Khan, I. M. Khan, and S. Khan, "Attention-Guided Wheat Disease Recognition Network through Multi-Scale Feature Optimization." Mar. 05, 2025. doi: 10.62762/TSCC.2025.435806.

[23] Prof. P. M. Gosavi, "Deepfake Video Face Detection," *International Journal for Research in Applied Science and Engineering Technology* , vol. 13, no. 4, p. 5840, Apr. 2025, doi: 10.22214/ijraset.2025.69233.

[24] Z. Chen, X. Liao, X. Wu, and Y. Chen, "Compressed Deepfake Video Detection Based on 3D Spatiotemporal Trajectories," *arXiv (Cornell University)* , Apr. 2024, doi: 10.48550/arxiv.2404.18149