ISSN: 2582-4686 SJIF 2021-3.261,SJIF 2022-2.889, 2024-6.875 ResearchBib IF: 10.25 / 2025

THE MULTIDISCIPLINARY JOURNAL OF SCIENCE AND TECHNOLOGY

VOLUME-6, ISSUE-3

DETECTION ALGORITHMS AND MODELS FOR SYNTHETICALLY GENERATED VISUAL MEDIA

Khakimbekov Doniyorbek Tursunpulot ugli
PhD student of New Uzbekistan University

Abstract. *The rapid development of artificial intelligence technologies has unprecedentedly expanded the capabilities for the automatic generation and manipulation of visual content. This, in turn, has led to a proliferation of deepfakes and other artificially generated objects that pose a significant threat to the reliability of digital information. This paper analyzes improved algorithms and models that enable the high-accuracy detection of artificially generated visual objects. Within the framework of this study, in contrast to traditional approaches that rely solely on spatial features, a hybrid architecture is proposed that evaluates temporal discontinuities and logical inconsistencies between consecutive video frames. The developed model was tested on prominent open-source datasets, and the results confirmed its superior effectiveness in detecting artificial manipulations compared to baseline methods. The proposed algorithmic solution can be practically implemented to enhance digital security and verify information authenticity across various information systems.*

Key words: *Artificial intelligence, visual object detection, computer vision, deep learning, neural networks, spatiotemporal analysis, digital security.*

INTRODUCTION

In the contemporary digital era, the majority of information exchange occurs through visual media. The advent of generative adversarial networks and diffusion models has granted unprecedented capabilities to synthesize highly realistic artificial images and videos [1]. While these technologies offer vast opportunities in entertainment, education, and digital art, their malicious application has increased exponentially. Artificially generated visual objects, commonly referred to as deepfakes, are increasingly utilized to propagate misinformation, compromise individual reputations, and execute sophisticated social engineering attacks aimed at bypassing biometric access control systems [2-3].

Consequently, verifying the authenticity of visual information in the digital environment has emerged as one of the most critical challenges in computer science and cybersecurity. Currently, numerous algorithms have been developed to detect artificial videos; however, the majority of these solutions operate on isolated frames, meticulously searching for pixel-level anomalies and spatial artifacts. As generative models continue to evolve, the spatial imperfections they produce are diminishing, thereby reducing the efficacy of traditional frame-by-frame detection algorithms. This research addresses the inherent limitations of existing models for detecting artificially generated visual objects by proposing a novel, highly accurate architecture that concurrently analyzes both the spatial and temporal characteristics of visual data. The primary objective of this study is to introduce an algorithmic model capable of discerning microscopic shifts in frame dynamics and systematically analyzing algorithmic errors that remain imperceptible to the human eye, such as unnatural movement trajectories or inconsistencies in biological signals [4].

Method

To achieve the objectives of this study, a novel hybrid neural network architecture was developed to analyze visual information across two distinct dimensions: spatial and temporal. The proposed

ISSN: 2582-4686 SJIF 2021-3.261,SJIF 2022-2.889, 2024-6.875 ResearchBib IF: 10.25 / 2025

methodology operates in a dual-phase pipeline. In the initial phase, a continuous sequence of video frames is subjected to spatial analysis. Each frame undergoes preprocessing, where the primary visual objects, specifically facial regions, are automatically extracted and normalized using advanced tracking algorithms. To extract spatial features, a convolutional neural network based on a residual architecture is employed. This network processes each isolated frame to identify pixel-level anomalies, blending artifacts, and illumination inconsistencies inherent to generative models. In the second phase, the extracted spatial feature vectors are analyzed along the temporal axis [5,6].

Recognizing that generative models struggle to maintain continuous and natural movement across time, the sequence of frame-level features is fed into a vision transformer mechanism. The self-attention modules within the transformer architecture are specifically designed to capture long-range dependencies and temporal inconsistencies. By analyzing the entire sequence rather than isolated images, the model effectively detects unnatural biological signals, such as irregular blink frequencies, asynchronous lip movements, and micro-expressions that deviate from human physiology. The training process of the proposed architecture was conducted using large-scale open-source datasets comprising both pristine videos and highly realistic synthetic content generated by various state-of-the-art manipulation algorithms. Hyperparameters, including the learning rate, sequence length, and batch size, were rigorously optimized to ensure robust convergence and mitigate the risk of overfitting.

Result and Discussion

To comprehensively evaluate the performance of the proposed hybrid architecture, a series of quantitative experiments were conducted using standard evaluation metrics, specifically accuracy, precision, recall, and the F1-score. The model was benchmarked against existing state-of-the-art methods on two widely recognized open-source datasets: FaceForensics++ and Celeb-DF. The FaceForensics++ dataset provided a rigorous testing ground with videos subjected to varying degrees of compression, simulating real-world social media environments where digital artifacts are often lost. The experimental results demonstrated that the proposed spatial-temporal model significantly outperformed traditional frame-by-frame analysis techniques across multiple testing scenarios.

Model	Dataset	Video Quality	Accuracy	F1-Score
Standard CNN	FaceForensics++	Uncompressed	89.50%	88.20%
Standard CNN	FaceForensics++	Compressed	76.30%	74.10%
3D-CNN Baseline	Celeb-DF	Mixed	85.10%	84.60%
Proposed Hybrid Model	FaceForensics++	Uncompressed	97.40%	96.80%
Proposed Hybrid Model	FaceForensics++	Compressed	92.10%	91.50%
Proposed Hybrid Model	Celeb-DF	Mixed	94.30%	93.90%

The data presented in the table illustrates a critical vulnerability in conventional convolutional networks: their performance degrades sharply when applied to compressed videos, dropping from 89.5 percent to 76.3 percent. This decline occurs because compression algorithms blur the microscopic pixel-level artifacts that purely spatial models rely upon. In contrast, the proposed hybrid architecture maintained a robust accuracy of 92.1 percent even on highly compressed media. This resilience validates our core hypothesis that temporal discontinuities, such as irregular blink rates or unnatural lip synchronization captured by the vision transformer, remain detectable regardless of

ISSN: 2582-4686 SJIF 2021-3.261,SJIF 2022-2.889, 2024-6.875 ResearchBib IF: 10.25 / 2025

spatial resolution degradation. Furthermore, the high F1-score of 93.9 percent on the challenging Celeb-DF dataset indicates that the model effectively minimizes both false positives and false negatives, making it highly suitable for integration into digital forensic workflows. Despite these promising results, the discussion must also address certain limitations observed during the testing phase. The integration of the vision transformer module significantly increased the computational overhead, requiring substantial processing power for video analysis. Additionally, a slight decrease in detection sensitivity was noted in scenarios involving extreme low-light conditions or when the target object occupied a minimal portion of the frame. Future iterations of the model will need to optimize the computational efficiency of the temporal analysis block to facilitate deployment on resource-constrained devices without sacrificing the high detection accuracy achieved in this study.

Conclusion.

This research paper comprehensively addressed the escalating challenge of identifying artificially generated visual objects by introducing a novel spatial-temporal hybrid architecture. By integrating spatial feature extraction through convolutional neural networks with temporal discontinuity analysis via vision transformers, the proposed model successfully captures both microscopic blending artifacts and unnatural biological signals. The empirical evaluations conducted on standard datasets, including FaceForensics++ and Celeb-DF, confirmed that analyzing frame sequences rather than isolated images is critical for robust detection. Notably, the hybrid approach demonstrated significant resilience against video compression, maintaining high accuracy in scenarios where conventional frame-by-frame models fail. Future research endeavors will focus on optimizing the computational efficiency of the transformer modules to facilitate real-time detection on mobile platforms and edge devices. Additionally, incorporating audio signal analysis to create a multimodal detection framework that evaluates the synchronization of visual movements with acoustic properties presents a promising avenue for further enhancing the reliability of digital forensic tools.

REFERENCES

1. Deressa, D. W., Lambert, P., Van Wallendael, G., Atnafu, S., & Mareen, H. (2024). Improved deepfake video detection using convolutional vision transformer. In 2024 IEEE Gaming, Entertainment, and Media Conference (GEM), 492-497.
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
3. Kaddar, B., Fezza, S. A., & Serra-Sagrista, J. (2024). Deepfake detection using spatiotemporal transformer. ACM Transactions on Multimedia Computing, Communications and Applications, 20(11).
4. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Niessner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF International Conference on Computer Vision.
5. Thakre, A., et al. (2025). CAST: Cross-Attentive Spatio-Temporal feature fusion for Deepfake detection. arXiv preprint arXiv:2506.21711.
6. Zhang, Z., & Laghari, A. A. (2025). Real-Time deepfake detection via gaze and blink patterns: A transformer framework. Computers, Materials & Continua, 85(1), 1457-1493.