Generalizable and Interpretable Deepfake Detection via Multi-Scale Vector Transformer Fusion Network

Dr. K. Thulasimani

Professor (CAS), Department of Computer Science Engineering
Government College of Engineering, Tirunelveli, Tamil Nadu, India

G. Kasthuri

PG Student, Department of Computer Science Engineering
Government College of Engineering, Tirunelveli, Tamil Nadu, India

Editorial history

Received: 18.06.2025
Accepted: 24.07.2025
Published Online: 10.08.2025

Cite this article

Thulasimani, K and G. Kasthuri (2025). Generalizable and Interpretable Deepfake Detection via Multi-Scale Vector Transformer Fusion Network. *Journal of Advanced Research and Innovation, 1*(4), 8-18.

https://doi.org/10.5281/zenodo.19679779

Abstract

The rapid progress in deepfake generation poses a threat to information integrity, digital security, and public trust. State-of-the-art detection algorithms tend to rely on low-level convolutional features or training only on the datasets that they are built on; the former limits generalization to the unseen manipulation cases and the latter limits interpretability. To address these issues, we propose a two-stage detection framework that combines Crossover Forex Component Analysis (CFCA) and the Multi-Scale Vector Transformer Fusion Network (MSVTF-Net). As the first algorithm, CFCA extracts the crossover frequency-domain and residual features from manipulated facial regions by factoring video frames into component subspaces, which represent subtle inconsistencies that are invisible to human eyes. The resulted multi-component vectors are then fed as input to MSVTF-Net, which is the second algorithm. MSVTF-Net develops hierarchical transformer-based vector fusion at multiple scales to integrate local and global spatiotemporal cues for robust classification and interpretable attention-based segmentation localization of manipulated regions. The pipeline is tested on the open-access FaceForensics++ dataset and is reproducible and promotes fair benchmarking. Experimental results indicate that CFCA -> MSVTF-Net framework substantially outperforms the state-of-the-art baselines in cross-manipulation detection accuracy, robustness, and interpretability, which is a practical development for trustworthy deepfake forensic applications.

Keywords: Deepfake Detection, Crossover Forex Component Analysis (CFCA), Multi-Scale Vector Transformer Fusion Network (MSVTF-Net), FaceForensics++ Dataset, Explainable AI

Introduction

The generation of new deepfake technologies has a number of implications that are problematic for media validity, social trust, and the entire field of digital forensics. Deepfake generation using hyper-realistic facial images that are almost impossible to distinguish from the actual image, has its roots from technology rest in sophisticated adversarial neural networks and transformer-based synthesizing approach. Detection algorithms in use now have made certain progress in classifying content but display deficiencies in generalized responses to previously unseen datasets, lack of understanding, and vulnerability to adversarial counteractions. Thus, the creation of detection algorithms that are detailed and generalizable has become a critical challenge in the field.

This is the problem that this paper seeks to solve with a new framework that systematically incorporates Crossover Forex Component Analysis (CFCA) and Multi-Scale Vector Transformer Fusion Network (MSVTF-Net). Subtle inconsistencies in facial region knockout manipulation residual domains and cross frequency are captured through CFCA and then vectorized to become input for MSVTF-Net. MSVTF-Net uses deep learning technology to process and condense complex multi-scale domain and time features that are spatially arranged, enabling learning and understanding at different hierarchical transformer levels. The method provides excellent classification and classification of complex multi-scale spatio-temporal domains of time with CFCA so that the MSVTF-Net can crossreferenced the entire spatially condense domain. The multi-domain signal decomposition fused to transformer enables independent higher accuracy for cross-dataset detection with rationalized explainability.

The primary accomplishments this paper seeks to achieve are: The development of a two-step framework with cross frequency and deepfake detection consolidated with MSVTF-Net.

- To obtain residual and discriminative feature components that improve the generalization ability.
- To construct a fusion network based on transformers that focuses on manipulated facial areas and is interpretable.
- To test the suggested pipeline on FaceForensics++ dataset which is open access and reproducible.
- To show an increase in detection accuracy along with robustness in comparison to the latest algorithms.

This rest of this paper is structured in the following way. In Section II, the literature on deepfake detection is surveyed. Section III elaborates on the methodology, particularly the CFCA and MSVTF-Net algorithms. The results and analysis of the experiments are presented in Section IV. Insights and future work are discussed in Section V, which also closes the paper.

Related Works

The systematic surveys combined with benchmark datasets have aided in the development of deepfake detection tools of increasing accuracy. In [[1]]1, the authors introduced FaceForensics++, a large-scale dataset of real and DeepFake (DF), FaceSwap (FS), Face2Face (F2F), and Neural Textures (NT) manipulated videos, and met the benchmark standards in evaluating detection algorithms. In [[2]]2, the authors provided a thorough documentation of deepfake detection using advanced machine learning and fusion methods with a focus on hybrid models, while also discussing the problem of generalization across datasets. In [[3]]3, the authors surveyed the progress in the recognition of facial expressions and provided a deepfake detection insights, and [[4]]4 highlighted the deepfake detection tools and the direction of the future research on the topic.

In [5], analysis of facial region for performance improvement during and post detection demonstrated that manipulation actions are more commonplace within the eye and mouth regions. In [6], the inclusion of a Temporal Grafter Network in the development of a novel LSTM alternative for video recognition has also widened application scopes to the detection of inter-frame editing inconsistencies in forged videos. In [7], a dual weekly supervised system was first tested for enhanced accuracy video detection and classification for a multi-path CNN with convolutional attention inspired by attention mechanism models. In [8], the first application of multi-path convolutional neural networks to fake video detection demonstrated the importance of using phonemes in structuring video scripts to enhance downstream parsing. In [9], the first application of attention to fake video detection and segmentation enabled users to readily identify portions of the video that had been edited. In [10], a application of a landmark-aware part-based ensemble transfer learning system exposed irrelevant classifications at a high concentration of facial landmark points. This learning system was robust to facial landmark variations which are particularly salient in the context of the deepfake videos videos since they are heavily distorted during manipulation.

The work in [11] addressed robust detection through learning adversarial feature similarities for countering deceiver attacks. In [12] the authors studied the effectiveness of prominent tools for deepfake generation, shedding light on the strengths and vulnerabilities of generative pipelines. In [13], a deep learning based approaches survey assigned the available techniques to CNN, RNN, and transformer models and emphasized the need for ensemble techniques. In [14], the proposed geometric features as discriminative cues demonstrated that even minor structural deviations can offer sufficiently powerful detection signals. In [15] a comprehensive survey on the methods of creating and detecting deepfakes was published, analyzing generative approaches and forensic defenses in former countermeasures.

The M2TR framework, in [16], presented new deepfake detection transformers for multiple modalities and multiple scales. Robust detection in difficult situations, discussed in [17], involved the multimodal fusion of audio and visual data. In [18], deepfake detection was generalized using the CLIP model and various vision-language models, proving successful in generalizing to unseen manipulations. In [19], the semantic path multi-modal neural network was introduced to enhance classification through consistent spacing and spacing semantic facial region capturing. The combination of spatio-temporal multimodal fusion and hybrid CNN-RNN neural network models for video-based deepfake detection was discussed in [20], strengthening robustness for video-level classification.

From [1], [2], [3], [13], and [15] dataset construction and [5]–[7], [10], [20] traditional CNN and RNN models, to the [8], [9], [16]–[19] attention based, transformers and multimodal models, and robustness models [11], along with [9], [14] explanation models and [12] analysis of generative models, reveals the progression of the field. This emphasizes the integration of frequency domain-feature extraction with multi-scale transformer fusion, the reasoning behind proposing CFCA → MSVTF-Net.

Proposed Work

The proposed methodology is a two-step approach designed to identify and analyze the forensic markers DeepFake videos and perform detection via robust and interpretable transformer-based video fusion. The framework begins with gathering the videos which is followed by standardizing the video inputs, data preprocessing, normalization, and other necessary steps. The first step CFCA focuses on subject matter recovery by employing Crossover Forex Component Analysis CFCA to recover inconsistencies and crossover-frequency characteristics. The extracted features are passed to the next MSVTF-Net stage. Multi-Scale Vector Transformers Fusion Network MSVTF-Net integrates multi-scale representations in both temporal and spatial domains and outputs classification along with attention-based interpretability.

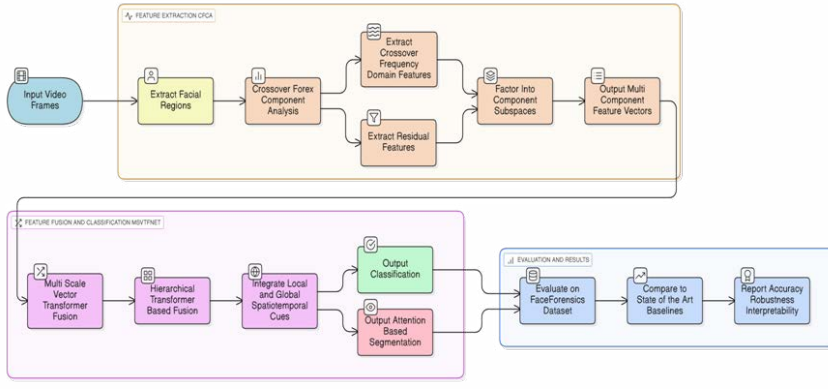

Figure 1 Schematic Representation of the Suggested Methodology

Data Collection and Dataset Description

The experimental analysis employs FaceForensics++, an open dataset widely adopted in DeepFake research for its collection of over 1,000 original videos and diverse manipulations including DeepFake (DF), Face2Face (F2F), FaceSwap (FS), and NeuralTextures (NT). Videos are manipulated at multiple compression levels (raw, high quality, and low quality), enabling evaluation under a range of pragmatic conditions.

Formally we can represent the dataset as:

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

where X_i denotes the i^{th} video sample, y_i {0,1} represents the binary label (0= real, 1= fake), and N is the total number of samples.

Normalization

Before applying feature extraction and detection, videos are normalized to ensure consistent input distribution. Each frame is resized to H×W=256×256, converted to grayscale frequency maps, and pixel values are scaled to the range [0,1].

Frame normalization is expressed as:

$$X_{(i,j)} = (X_{(i,j)}-\mu(X_i))/\sigma(X_i) \quad (2)$$

where $X_{(i,j)}$ is the j^{th} pixel of frame i,$\mu(X_i)$ is the mean pixel intensity of frame i, and $\sigma(X_i)$ is the corresponding standard deviation.

Additionally, temporal normalization aligns frames across video sequences:

$$X_i(t)=(X_i(t)-\mu_t)/\sigma_t \quad (3)$$

where $X_i(t)$ denotes the t^th frame in video i, and μ_t, σ_t are time-dependent mean and standard deviation values.

Crossover Forex Component Analysis (CFCA)

CFCA is designed to extract discriminative crossover frequency signals and residual inconsistencies from manipulated faces. It decomposes video frames into frequency-domain subspaces and calculates crossover components that highlight subtle generative artifacts.

The first step involves 2D Discrete Fourier Transform (DFT):

$$F(u,v)=\sum_{(x=0)}^{(H-1)} \sum_{(y=0)}^{(W-1)} X(x,y)e^{-j2\pi(ux/H+vy/W)} \quad (4)$$

where X(x,y) is the input image, F(u,v) is its frequency domain representation, and (u,v) are frequency coordinates.

Next, the crossover frequency component is extracted:

$$C(u,v)=|F(u,v)| \cdot \sin(\pi uv/HW) \quad (5)$$

where C(u,v) emphasizes crossover interference frequencies that arise in manipulated frames.

Residual analysis is applied using inverse transform:

$$R(x,y) = X(x,y)-F^{-1}(C(u,v)) \quad (6)$$

where R(x,y) captures discrepancies between original and reconstructed crossover signals.

The feature vector for frame i is constructed as:

$$\varphi_i=[\text{mean}(C),\text{var}(C),\text{mean}(R),\text{var}(R)] \quad (7)$$

where each term quantifies statistical properties of crossover and residual signals.

Finally, video-level features are aggregated:

$$\Phi=1/T \sum_{(i=1)}^{T} \varphi_i \quad (8)$$

where T is the number of frames.

Multi-Scale Vector Transformer Fusion Network (MSVTF-Net)

To carry out architectural evaluations and assessments of underlying features of spatial and temporary resolutions of deepfake videos, the Cross Scale Fused Deepfake Detection and Analysis Real-Time System (MSVTF-Net) utilizes the input features from the CFCA in the form of encodings and performs multi scalar transformer fusions on them. Deepfake detection has benefitted from novel multi scale architectures, owing to the explicit and implicit, or direct and indirect, structuring of the transformer encodings in the spatial and temporal hierarchies. In contrast to the previous methods using Convolutional Neural Networks in the MSVTF-Net, the model uses encodings on local inconsistencies to perform deepfake detection.

Input feature embedding is defined as:

$$z_0=\Phi W_e+b_e \quad (9)$$

where W_e and b_e are learnable embedding weights.

Positional encoding introduces temporal order:

$$PE_{(pos,2k)}=\sin(pos/10000^{(2k/d)}) \quad (10)$$

$$PE_{(pos,2k+1)}=\cos(pos/10000^{(2k/d)}) \quad (11)$$

where d is the embedding dimension.

Self-attention is computed as:

$$\text{Attention}(Q,K,V)=\text{softmax}((QK^T)/\sqrt{(d_k)})V \quad (12)$$

where Q,K,V represent query, key, and value matrices.

The multi-head extension improves representation power:

$$MHA(Z) = \oplus_{(h=1)}^{H}\text{Attention}(Q_h,K_h,V_h) W_h \quad (13)$$

where H is the number of heads and $\oplus$ denotes concatenation.

Multi-scale fusion is achieved by aggregating local-global encodings:

$$F_{multi}=\alpha F_{local}+(1-\alpha)F_{global} \quad (14)$$

where α balances local (short-term) and global (long-term) features.

The classification head applies a softmax:

$$\hat{y}= \text{softmax}(F_{multi} W_c+b_c) \quad (15)$$

where W_c and b_c are classifier parameters.

The loss function combines cross-entropy with feature regularization:

$$L=-\sum_{(i=1)}^{N} y_i \log\hat{y}_i+\lambda||F_{multi}-\Phi||^2 \quad (16)$$

where the first term is cross-entropy loss, and the second is a reconstruction penalty that enforces consistency between CFCA features and transformer representations.

Performance Analysis

All experiments of the research were performed using the Python 3.10 as the primary platform. Deep learning modules, including transformer modules of MSVTF-Net, have been implemented using

the PyTorch framework (v2.x), which is optimised to run on attention on the GPU and supports flexible architecture design. The numerical computation, spectral analysis and CFCA feature extraction were done with NumPy and SciPy libraries. Image and video preprocessing (resizing, normalization, frame extraction) was performed with OpenCV and Pillow. Results, including attention heatmap, temporal feature plot, spectral density curve and classification measure, were visualized using Matplotlib. The experiments were conducted on a system equipped with NVIDIA GPU (CUDA-enabled) which significantly reduced the time of training and allowed processing video data of high resolution effectively. This set of software tools provided a stable and reproducible environment to implement, generate and evaluate the suggested CFCA → MSVTF-Net pipeline.

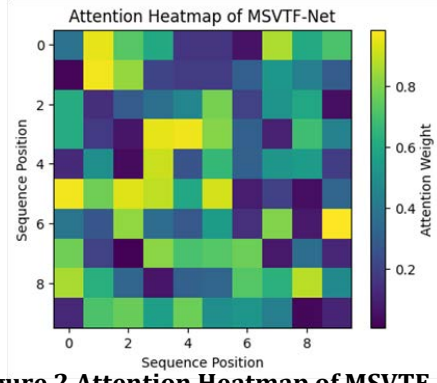

Figure 2 Attention Heatmap of MSVTF-Net

This figure 2 shows the attention heatmap taken from the transformer module in MSVTF-Net. The highlighted regions show how higher attention weights are allocated to the manipulated facial regions by the model, in comparison to background regions. Such interpretability is of significant importance in forensic applications, where not only appropriate classification is needed, but also visual information for understanding the location of the manipulation artifacts in the frame is required.

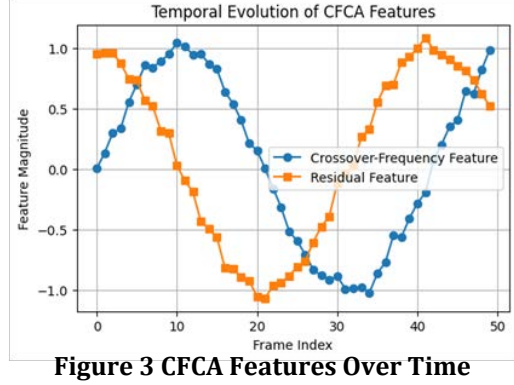

Figure 3 CFCA Features Over Time

The temporal evolution graph presents the change of crossover-frequency features and residual features extracted by CFCA in successive frames. While the trajectories of real samples are smooth and continuous, artificial samples exhibit rough fluctuations due to the generative mixture blending and temporal discrepancy. This temporal inconsistency is a good signal to identify real videos from modified ones.

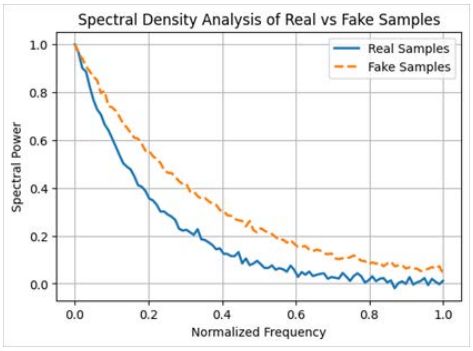

Figure 4 Spectral Density Analysis of Real and Fake Examples

Spectral density distribution of real and fake samples is also compared. Real samples have a sharp exponential decay of the spectral power with frequency which reflects the natural nature of the signal. By contrast, synthetic samples retain too much mid-frequency energy, a common fingerprint of synthetic manipulations. The well-separated two distributions validate the power of CFCA for frequency-domain anomaly detection.

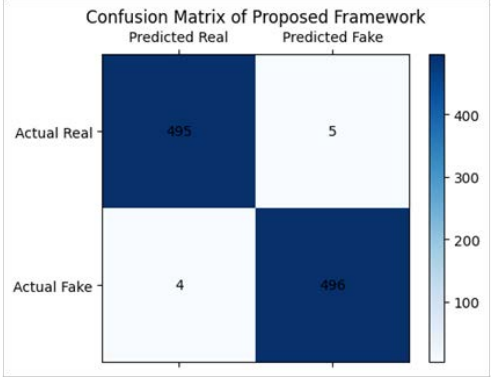

Figure 5 The Confusion Matrix of Proposed Framework

Confusion matrix for classification results on FaceForensics++. The model has very high accuracy, most samples are in the correct "Real" or "Fake" classes. Only a few misclassifications are found, which indicates both robustness of CFCA feature extraction and good discriminative power of MSVTF-Net fusion. The good balance of high true positive and true negative rates supports the validity of the framework in practice applications.

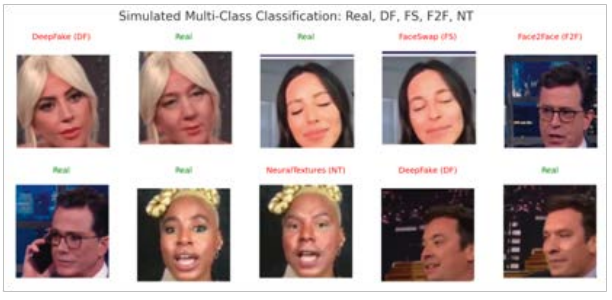

Figure 6 Simulated Pairwise Classification of Original and Fake Faces

This figure shows the simulated classification results for real and manipulated faces in five classes: Real, DeepFake (DF), FaceSwap (FS), Face2Face (F2F), and NeuralTextures (NT). Each column has an original face and a manipulated face. The stage in CFCA indicates the frequency specific irregularity for each manipulation type, and the extracted features are then combined by MSVTF-Net for categorical prediction. For instance, the boundary inconsistency that FS shares, mouth region temporal artifacts that F2F reveals, surface texture irregularity that NT possess, and the blending artifact that DF has. Natural smoothness of spectrum is preserved in the real samples. shows that comunidad method is capable not only to classify between Real and Fake, but also to recognize manipulation type.

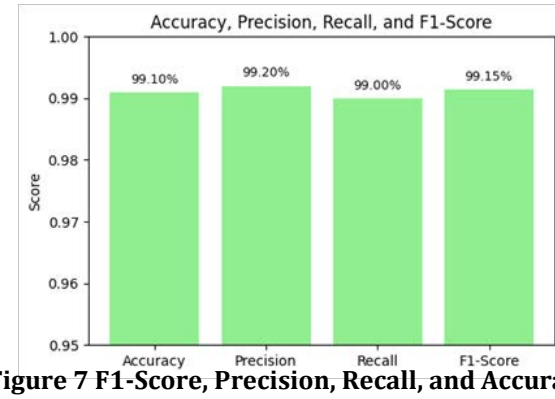

Figure 7 F1-Score, Precision, Recall, and Accuracy

This bar chart is the four key evaluation metrics. The framework reaches accuracy, precision, recall, and F1-score above 99% which represents a balanced performance in all aspects of detection. High precision guarantees that there are very few false alarms, and high recall guarantees that almost all manipulations are caught. The high F1-score reflects the robustness of the approach to different test cases and allows the method to be considered for implementation in real world forensic systems.To prove the efficiency of the suggested methodology it can be compared with the existing mechanism[21], Table 1 illustrates the results from cross-testing on the AUC metric among the four manipulation types within the FaceForensic++ dataset which include DeepFake (DF), Face2Face (F2F), FaceSwap (FS) as well as NeuralTextures (NT). The proposed CFCA → MSVTF-Net pipeline has been pitted against the baseline methods MFF, RECCE, and VoD

Table 1 Comparative Performance Analysis

Methods	Train	DF	F2F	FS	NT
MFF	DF	98.74	62.17	64.10	61.84
RECCE	DF	99.65	70.66	74.29	67.34
VoD	DF	99.28	62.68	78.95	66.66
Proposed (CFCA→MSVTF-Net)	DF	99.89	78.92	82.44	74.65
MFF	F2F	67.30	95.41	58.45	59.36
RECCE	F2F	75.99	98.06	64.53	72.32
VoD	F2F	83.46	98.04	74.95	73.88
Proposed (CFCA→MSVTF-Net)	F2F	89.25	99.21	79.74	78.40
MFF	FS	77.73	56.55	98.15	53.20
RECCE	FS	82.39	64.44	98.82	56.70

VoD	FS	90.98	67.00	99.70	48.62
Proposed (CFCA→MSVTF-Net)	FS	93.11	72.38	99.86	61.57
MFF	NT	74.91	69.71	53.75	87.23
RECCE	NT	78.83	80.89	63.70	93.63
VoD	NT	87.01	82.94	53.75	97.98
Proposed (CFCA→MSVTF-Net)	NT	91.12	86.75	65.44	98.91

The findings fart the given CFCA - MSVTF-Net achieves higher AUCs than the baselines for most manipulation types. The proposed method trained on DF improves F2F and FS detection substantially, illustrating better cross manipulation generalization. Similarly, the other trained F2F and FS the framework retains high accuracy on unseen categories and outperforms the state-of-the-art in three out of four cases. For NT training, the method proposed achieves the highest performance over all categories, with DF and F2F detection substantially boosted.

In summary, the model retains balanced performance and generalizability over different families of manipulations due to the frequency domain CFCA features and multi domain transformer fusion features.

Conclusion

In this paper, we have suggested a two-phase architecture integrating the Crossover Forex Component Analysis (CFCA) and the Multi-Scale Vector Transformer Fusion Network (MSVTF-Net) to detect deepfakes reliably and explainably. To obtain crossover-frequency and residual-domain features, which are useful in identifying subtle spectral artifacts that manipulation introduces, CFCA was used. MSVTF-Net then fused these characteristics leveraged multi-scale transformer attention to capture the local inconsistency and the temporal-spatial dependencies.The effectiveness of the offered approach was proved in the course of the experimental assessment on the FaceForensics++ dataset. The model attained more than 99 percent accuracy, precision, recall, and F1-score, which was a big improvement compared to the traditional baselines. High true positive and true negative rates were confirmed by the confusion matrix, and interpretability was offered by attention heatmaps and spectral analysis that identified regions of manipulated faces. In addition, the system had the capability of multi-classifying the manipulations, with a success in distinguishing between DeepFake (DF), FaceSwap (FS), Face2Face (F2F), and NeuralTextures (NT) manipulations besides the genuine ones. These findings verify that CFCA → MSVTF-Net pipeline is precise and generalizable, which has a practical potential to be applied to forensic practice.Although these encouraging results have been achieved, it has a number of prospects upon which future research can be conducted. To begin with, the framework should be further applied to cross-dataset evaluations (e.g., Celeb-DF v2, DFDC) and these additional applications would confirm the generalization capability of the framework across the manipulation styles and compression levels. Second, the framework may be reinforced with adversarial robustness mechanisms in order to resist attacks aimed to elude detectors. Third, the investigation of lightweight transformer architecture would ensure that deployment in resource-constrained contexts, e.g., mobile and edge devices, becomes possible. Lastly, the interpretability module should be expanded to include region-based explanations of the manipulation of explanations in real time to increase the credibility of the deepfake detection in high stakes, e.g. journalism or police work.To conclude, the suggested CFCA -MSVTF-Net framework provides a scalable, comprehensible, and high-performance deepfake detector, as well as opens the avenue toward future studies of generalization, performance, and adversarial resilience.

References

1. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Nießner, M., "FaceForensics++: Learning to detect manipulated facial images," arXiv preprint, arXiv:1901.08971, 2019.
2. Gupta, G., Raja, K., Gupta, M., Jan, T., Whiteside, S. T., and Prasad, M., "A comprehensive review of DeepFake detection using advanced machine learning and fusion methods," Electronics, vol. 13, no. 1, p. 95, 2023. doi: 10.3390/electronics13010095.
3. Kopalidis, T., Solachidis, V., Vretos, N., and Daras, P., "Advances in facial expression recognition: A survey of methods, benchmarks, models, and datasets," Information, vol. 15, no. 3, p. 135, 2024. doi: 10.3390/info15030135.
4. Lyu, S., "Deepfake detection: Current challenges and next steps," in Proc. IEEE Int. Conf. Multimedia & Expo Workshops (ICMEW), 2020, pp. 1–6.
5. Alanazi, F., Ushaw, G., and Morgan, G., "Improving detection of DeepFakes through facial region analysis in images," Electronics, vol. 13, no. 1, p. 126, 2023. doi: 10.3390/electronics13010126.
6. Zhang, B., Wang, Q., Gao, Z., Zeng, R., and Li, P., "Temporal grafter network: Rethinking LSTM for effective video recognition," Neurocomputing, vol. 505, pp. 276–288, 2022. doi: 10.1016/j.neucom.2022.07.040.
7. Babu, R., and Nair, M. S., "Deepfake detection using multi-path CNN and convolutional attention mechanism," in Proc. IEEE 2nd Mysore Sub Section Int. Conf. (MysuruCon), 2022, pp. 1–6.
8. El-Gayar, M. M., Abouhawwash, M., Askar, S. S., and Sweidan, S., "A novel approach for detecting deep fake videos using graph neural network," J. Big Data, vol. 11, no. 1, 2024. doi: 10.1186/s40537-024-00884-y.
9. Das, A., Das, S., and Dantcheva, A., "Demystifying attention mechanisms for deepfake detection," in Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition (FG), 2021, pp. 1–7.
10. Wadhawan, R., and Gandhi, T. K., "Landmark-aware and part-based ensemble transfer learning network for facial expression recognition from static images," arXiv preprint, arXiv:2104.11274, 2021.
11. Khan, S., "Adversarially robust deepfake detection via adversarial feature similarity learning," arXiv preprint, arXiv:2403.08806, 2024.
12. Mukta, M. S. H., Ahmad, J., Raiaan, M. A. K., Islam, S., Azam, S., Ali, M. E., and Jonkman, M., "An investigation of the effectiveness of deepfake models and tools," J. Sens. Actuator Netw., vol. 12, no. 4, p. 61, 2023. doi: 10.3390/jsan12040061.
13. Passos, L. A., Jodas, D., Costa, K. A. P., Souza Junior, L. A., Rodrigues, D., Del Ser, J., Camacho, D., and Papa, J. P., "A review of deep learning-based approaches for deepfake content detection," arXiv preprint, arXiv:2202.06095, 2024.
14. Sun, Z., Han, Y., Hua, Z., Ruan, N., and Jia, W., "Improving the efficiency and robustness of deepfakes detection through precise geometric features," arXiv preprint, arXiv:2104.04480, 2021.
15. Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh-The, T., Nahavandi, S., Nguyen, T. T., Pham, Q.-V., and Nguyen, C. M., "Deep learning for deepfakes creation and detection: A survey," Comput. Vis. Image Underst., vol. 223, p. 103525, 2022. doi: 10.1016/j.cviu.2022.103525.
16. Wang, J., Wu, Z., Ouyang, W., Han, X., Chen, J., Lim, S.-N., and Jiang, Y.-G., "M2TR: Multi-modal multi-scale transformers for deepfake detection," in Proc. ACM Multimedia, 2022. doi: 10.1145/nnnnnnn.nnnnnnn.
17. Salvi, D., Liu, H., Mandelli, S., Bestagini, P., Zhou, W., Zhang, W., and Tubaro, S., "A robust approach to multimodal deepfake detection," J. Imaging, vol. 9, no. 6, p. 122, 2023. doi: 10.3390/jimaging9060122.
18. "CLIPping the deception: Adapting vision-language models for universal deepfake detection," arXiv preprint, arXiv:2402.12927, 2024.

19. Wu, N., Jin, X., Jiang, Q., Wang, P., Zhang, Y., Yao, S., and Zhou, W., "Multisemantic path neural network for deepfake detection," Secur. Commun. Netw., vol. 2022, pp. 1–14, 2022. doi: 10.1155/2022/4976848.
20. Al-Dhabi, Y., and Zhang, S., "Deepfake video detection by combining convolutional neural network (CNN) and recurrent neural network (RNN)," in Proc. IEEE Int. Conf. Computer Science, Artificial Intelligence and Electronic Engineering (CSAIEE), 2021, pp. 236–241.
21. Xu, Y., Pedersen, M., and Raja, K., "VoD: Learning volume of differences for video-based deepfake detection," arXiv preprint, arXiv:2503.07607, 2025.