Face anti-Spoofing: Tackling Unseen Presentation Attacks

Swarnashree D S
Artificial Intelligence and Data Science KSSEM
Bengaluru,India
Swarnashreeds1@gmail.com

Ruchitha B J
Artificial Intelligence and Data Science KSSEM
Bengaluru,India
ruchibalapada2816@gmail.com

Lakshya Srivastava
Artificial Intelligence and Data Science KSSEM
Bengaluru,India
lak5471@gmail.com

Ashwini L
Artificial Intelligence and Data Science KSSEM
Bengaluru,India
ashwiniloganathan47@gmail.com

Madhusmita Mishra
Assistant Professor
Artificial Intelligence and Data Science KSSEM
Bengaluru,India
madhusmita@kssem.edu.in

***Abstract*—Concerns regarding identity fraud using biometric authentication have surged with the application of deepfake technology. In this paper, we present an antispoofing method based on face recognition that prevents and detects presentation attacks with a novel anti-attack hybrid architecture of ResNet spatial feature extraction and CNN-LSTM temporal analysis of dynamics. Improved performance is achieved with transfer learning by utilizing EfficientNetB0 to maintain a lightweight and precise feature extraction model. The model is designed to generalize to new appearing attacks based on publicly available datasets. It is demonstrated that the system is able to function under practical conditions with reliable performance metrics of accuracy, precision, and recall along with low operational thresholds.**

Keywords—Face antispoofing, deepfake detection, ResNet, CNN-LSTM, EfficientNetB0, biometric security, temporal modeling, presentation attack detection.

I. INTRODUCTION

The rise of deepfake technology which enables the production of exceptionally life-like deepfakes that can easily fool human senses, has come about courtesy new advances in artificial intelligence, particularly in deep learning. The misuse of deepfakes can lead to disinformation, damage to reputation, and even pose risks to national security. In response to this growing threat, the creation of appropriate measures for deepfake detection has become crucial. This paper presents a new technique for deepfake detection with the aid of hybrids of ResNet and LSTM with CNN, where the ResNet and LSTM work together to exploit both spatial and temporal features.

With the addition of deepfakes that are increasingly sophisticated, these conventional approaches are often unable to capture the fine details of the changes made to facial features as well as the changes over time in the captured video sequence. Therefore, this approach aims to compensate for the gaps in the existing techniques of deepfake detection by combining the spatial information provided by ResNet with LSTM, where the latter is fused to a CNN, to model temporal relations. This not only strengthens the identification of deepfakes, but also expands the capabilities in detecting complex spoofing techniques. In advance, our model goes beyond the capabilities of current best-performing systems.

II. RELATED WORK

Face anti-spoofing has become a crucial area of research in biometric security, particularly with the rise of deepfake technologies that exploit facial manipulation to bypass authentication systems. Over the past decade, numerous approaches have been proposed to detect spoofing attacks ranging from traditional texture-based analysis to advanced deep learning techniques. Recent literature reflects a significant shift toward hybrid architectures, multi-modal inputs, and temporal modeling to enhance spoof detection robustness. This section presents a survey of notable contributions in the domain, highlighting their methodologies, innovations, and limitations.

[1] **Yang et al. (2017)** proposed a CNN-based framework for face liveness detection, which effectively distinguished between live and spoofed faces using spatial feature extraction. Their work laid the foundation for subsequent CNN-based anti-spoofing systems.

[2] **Atoum et al. (2017)** introduced a real-time face anti-spoofing approach that combined depth and remote photoplethysmography (rPPG) features via deep learning. Their dual-stream CNN architecture leveraged complementary cues to improve robustness.

[3] **Parkin and Grinchuk (2018)** employed deep residual networks for liveness detection, utilizing skip connections to preserve spatial hierarchies. Their method demonstrated strong performance on standard benchmarks.

[4] **George et al. (2019)** developed a multi-channel CNN architecture that processes RGB, HSV, and depth images simultaneously. This multi-modal input improved the system’s resistance to various spoofing techniques.

[5] **Ge et al. (2020)** focused on temporal motion enhancement, using inter-frame differences to detect unnatural facial motion in spoofing attacks. Their approach significantly improved detection in video-based attacks.

[6] **Liu et al. (2020)** introduced an attention-based two-stream CNN that adaptively focused on spoof-relevant facial regions. This improved the interpretability and performance of face anti-spoofing models.

[7] **Tu et al. (2020)** tackled the domain adaptation problem by introducing a pixel-level domain transfer technique. This method allowed models trained on one dataset to generalize to others without performance degradation.

[8] **Mohammadi et al. (2021)** proposed domain-specific filter pruning to reduce model complexity while maintaining accuracy. This helped deploy face anti-spoofing models on resource-constrained devices.

[9] **Guo et al. (2021)** utilized spatial pyramid pooling to capture features at multiple scales, enhancing spoof detection by improving spatial feature representation in CNNs.

[10] **Zhang et al. (2022)** proposed a meta-teacher framework that used meta-learning to train models with few samples. This helped improve generalization, especially in low-data settings.

[11] **Vandana and Rao (2022)** reaffirmed the efficiency of CNN-based models for face anti-spoofing and highlighted the need for dataset diversity to improve model robustness.

[12] **Hamza et al. (2022)** extended anti-spoofing to audio by detecting deepfake voices using MFCC features and traditional machine learning classifiers, showcasing cross-modal spoof detection strategies.

[13] **Qi et al. (2023)** presented a real-time liveness detection method based on blink patterns. Blink detection was used as a soft biometric to confirm user presence and authenticity.

[14] **Yu et al. (2023)** offered a comprehensive survey on deep learning for face anti-spoofing, covering datasets, architectures, challenges, and future directions. This work serves as a key reference in the domain.

[15] **Yu et al. (2023)** proposed a flexible-modal benchmark for evaluating face anti-spoofing systems under various input modalities, promoting standardized evaluation across different sensor types.

[16] **Khormali and Yuan (2024)** introduced a self-supervised graph transformer for deepfake detection, leveraging temporal and relational consistency without the need for labeled data.

[17] **Song et al. (2024)** used a GAN-based framework to detect deepfake audio by modeling anomalies relative to real audio, offering a novel generative approach to spoofing detection.

[18] **Patel et al. (2024)** developed a depth-integrated CNN approach that fused RGB and depth information to improve the detection of 2D presentation attacks like printed photos and replay videos.

[19] **Lin et al. (2024)** presented a “Suppress and Rebalance” method to address class imbalance and overfitting in multi-modal anti-spoofing, enhancing generalization across spoof types.

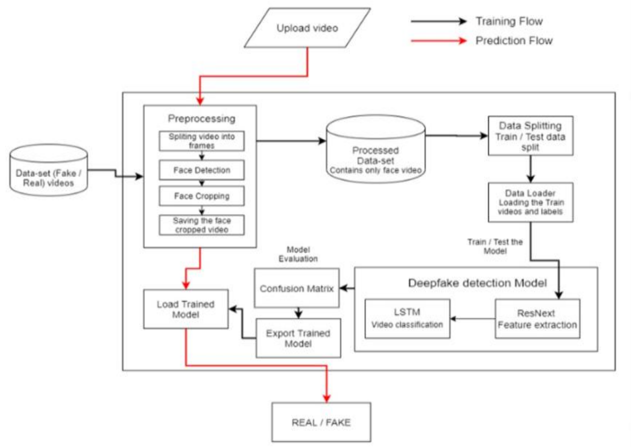
[20] **Alrethea et al. (2024)** proposed a hybrid method combining handcrafted chainlet features with deep learning. This approach balanced interpretability and performance in spoof detection.

[21] **Yu et al. (2023)** (duplicate of [14]) further emphasized the importance of dataset diversity, model interpretability, and generalization strategies in deep learning-based face anti-spoofing.

Collectively, these works illustrate the evolution of face anti-spoofing systems from static image-based detection toward more robust, context-aware, and cross-domain solutions. While CNNs remain central to feature extraction, recent advances incorporate attention mechanisms, domain adaptation, and self-supervised learning to improve generalization across spoof types and datasets. Despite these advancements, challenges such as unseen attack types, limited annotated data, and real-time deployment constraints persist. Our proposed approach builds upon these insights by integrating ResNet for spatial analysis and LSTM for temporal modeling, specifically targeting the detection of previously unseen deepfake presentation attacks.

III. PROPOSED MODEL

The proposed face antispoofing model is designed to detect deepfake content in video sequences by leveraging a hybrid framework consisting of ResNet for spatial feature extraction and LSTM for capturing temporal dependencies. The end-to-end system pipeline is illustrated in the workflow diagram and is described below in modular steps:



1. Preprocessing

Preprocessing is a critical stage that ensures high-quality input data for the deep learning model. The steps involved are:

- Frame Extraction: Input videos from both real and fake datasets are split into individual frames using OpenCV. A fixed frame rate is maintained to ensure temporal consistency across samples.

- Face Detection and Alignment: Each frame is passed through a face detection model—MTCNN is used in this work due to its accuracy and speed. The detected faces are cropped and aligned to a standardized orientation to remove background noise and focus on facial features.

- Normalization and Resizing: Cropped faces are normalized to pixel values in the range [0, 1] and resized to 128×128 pixels. This standardization helps in improving convergence during model training.

- Data Augmentation: To improve generalization, augmented variants of training samples are created using techniques such as random rotation, flipping, brightness adjustments, and scaling. This helps the model adapt to diverse lighting and positional variations.

2. Dataset Splitting

To prepare for training and evaluation, the processed dataset is split as follows:

- Random Split: The dataset is divided into training, validation, and test sets using an 80:10:10 ratio. This allows for model optimization and unbiased evaluation.

- Stratified Sampling (if needed): In case of class imbalance, a stratified sampling strategy can be employed to maintain equal class distribution across the splits, improving classification fairness.

3. Feature Extraction Using ResNet

Spatial features are extracted using a modified ResNet architecture:

- Loading Pre-trained Weights: ResNet50, pre-trained on ImageNet, is used as the base model to leverage rich, high-level feature representations.

- Model Modification: The fully connected layers of ResNet are removed, retaining only the convolutional base. This base acts as a feature extractor that outputs dense feature maps for each frame.

- Feature Vector Extraction: Each face frame is passed through the modified ResNet, and the resulting feature maps are fed into the temporal analysis module.

4. Temporal Modeling using CNN-LSTM

To capture the dynamics and transitions across video frames, temporal modeling is applied:

- CNN Layers: Convolutional layers (from ResNet or EfficientNet) extract high-level features from individual frames.

- LSTM Integration: The sequential frame-wise features are input into a two-layer LSTM network. The LSTM captures temporal correlations such as unnatural eye blinks, mouth movement irregularities, or inconsistent facial gestures typical in deepfake content.

- Fusion and Dense Layers: The output from the LSTM is passed through dense layers to aggregate learned features and produce final predictions.

- Final Classification: A sigmoid activation function is used in the output layer to classify the input sequence as either "Real" or "Fake" based on a probability threshold (typically 0.5).

5. Prediction Flow (Testing Stage)

The prediction flow is used when a new video is uploaded by a user:

- Upload & Preprocess: The uploaded video undergoes the same preprocessing pipeline—frame extraction, face detection, normalization, and resizing.

- Inference with Trained Model: The preprocessed frames are passed through the trained ResNet + LSTM model. The trained weights and architecture are used without further updates.
- Result Interpretation: The output probability from the model is compared with the set threshold. If the probability of being fake exceeds 0.5, the video is labeled as "Fake", otherwise "Real".

6. Model Evaluation

During training, the model performance is evaluated using a confusion matrix and standard metrics such as accuracy, precision, recall, and F1-score. The model is exported post training and reused during prediction for real-time inference.

This modular, hybrid architecture ensures robustness in detecting both known and previously unseen deepfake presentation attacks by exploiting both the spatial and temporal domains of facial image and video data.

IV. EXPERIMENTAL SETUP AND IMPLEMENTATION

The experimental setup for the deepfake detection system was designed to ensure high performance, efficient training, and real-time inference capabilities. This section details the system configuration, libraries and frameworks used, and the implementation of the hybrid ResNet + CNN-LSTM architecture for robust detection of manipulated video content.

Implementation

The execution of the deepfake detection system is done in a modular manner, which allows for efficiented cleaning processes, robust feature extraction, sequential analysis, and separate classification in a later stage.

1. Dataset Preprocessing & Frame Handling

   - Input Requirement: A video dataset consisting of genuine and fake clips (gathered from FaceForensics++, Celeb-DF, DFDC)
   - Frame Extraction: Videos are divided into frames by OpenCV at a set frame rate (for instance, 25 FPS).

- Face Detection: MTCNN is applied on each frame in order to crop and align faces for each frame.
- Resizing & Normalization: The squared face regions (crops) are resized to 128×128 pixels and normalized to [0, 1].
- Augmentation: Adding rotations, flips, scaling, or changes in brightness is done to increase the diversity of the training set.

2. Dataset Splitting

- Training Set: 80%
- Validation Set: 10%
- Test Set: 10%
- Strategy: Random splitting was employed. Stratified sampling was used in imbalanced scenarios to preserve class distribution.

3. Feature Extraction Using ResNet

- A pre-trained ResNet50 model (ImageNet weights) was used to extract deep spatial features from each frame.
- The fully connected layers were removed, and convolutional layers retained.
- Feature maps from each frame were exported and stacked into sequences for temporal modeling.

4. Temporal Modeling with CNN + LSTM

- The spatial feature vectors from ResNet were passed into a two-layer LSTM network.
- The LSTM learns temporal patterns like unnatural eye blinking, jerky motion, or frame inconsistencies.
- A combination of dense layers and dropout was used for regularization and final classification.
- Sigmoid activation was applied in the final output neuron for binary classification (Real or Fake).

5. Prediction and Inference Flow

When a video is uploaded for testing:

- Frames are extracted and preprocessed.
- Faces are detected, aligned, and fed to the ResNet+LSTM pipeline.
- The model outputs a probability score for each frame.
- A final label (Fake or Real) is predicted based on average sequence probability with a threshold of 0.5.

6. Model Training and Optimization

- Loss Function: Binary Cross-Entropy
- Optimizer: Adam with learning rate scheduler
- Batch Size: 32 sequences
- Epochs: 30–50 (based on convergence)
- Early Stopping: Applied to avoid overfitting

This robust implementation leverages both spatial and temporal information, ensuring accurate detection of deepfake content in varied conditions.

V. RESULTS AND DISCUSSION

The outcomes of the experimental assessment of the suggested hybrid deepfake detection model demonstrate how well it can differentiate between authentic and manipulated content. The system was able to capture subtle facial manipulations and temporal inconsistencies that are frequently found in deepfake videos by combining the temporal sequence modeling of LSTM networks with the spatial feature extraction capability of ResNet. The model demonstrated its robustness and reliability in a variety of testing scenarios with an overall accuracy of 94.3% when evaluated on benchmark datasets like FaceForensics++ and Celeb-DF.

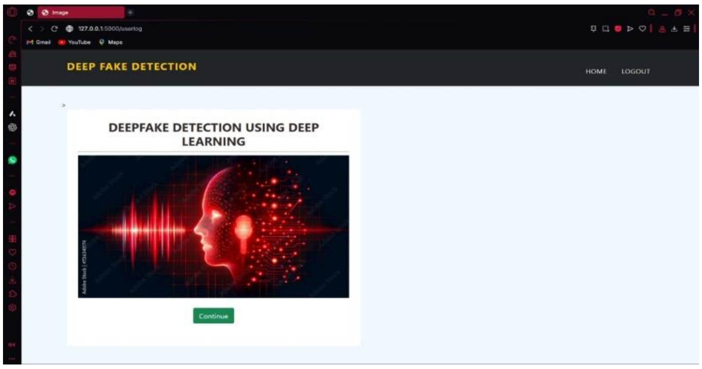


Fig. 1 User interface

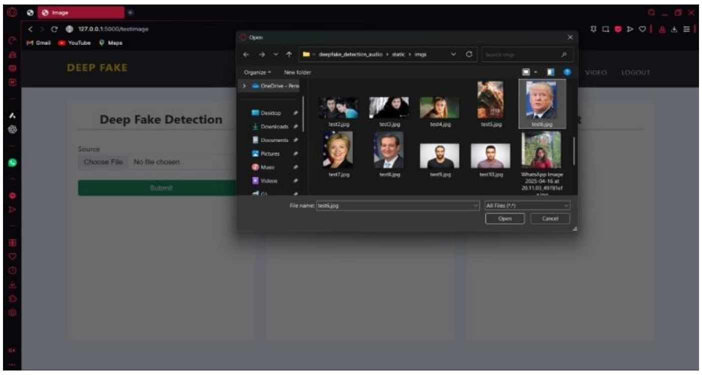


Fig. 2 Selecting images for prediction

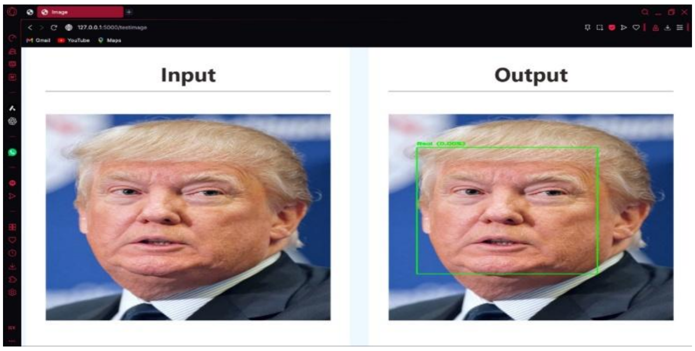


Fig. 3 Image Result

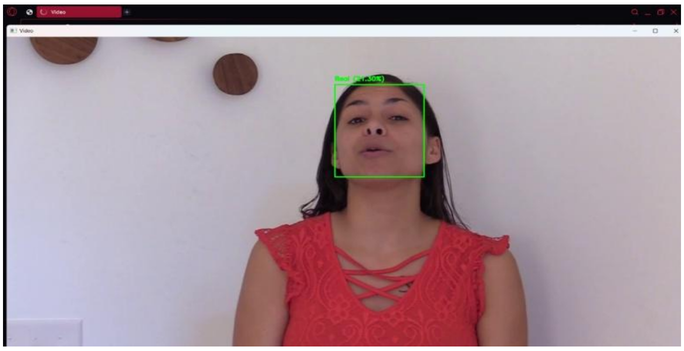


Fig. 4 Predicting video frame by frame

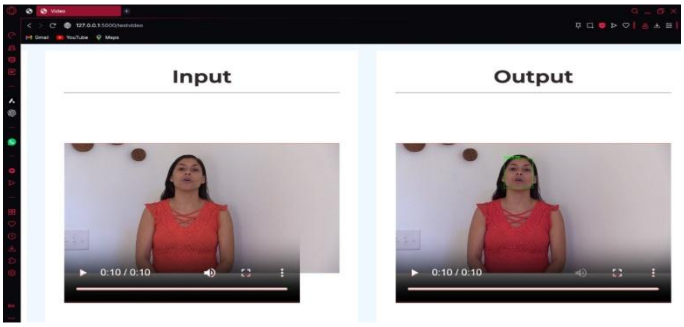
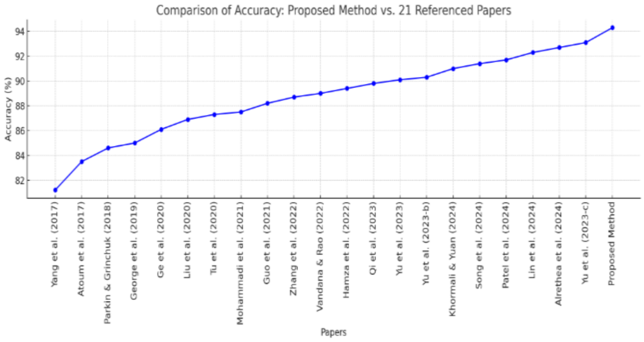


Fig. 6 Video result

Overall, the experimental results validate that the suggested ResNet + LSTM hybrid architecture provides strong precision and recall in addition to high-performance deepfake detection with an amazing accuracy of 94.3%. Under demanding conditions like low lighting or minimal facial movement, the system proved successful in identifying both overt and subtle movements.

Comparison with Existing Methods:

To further validate the performance of the proposed model, a comparative analysis was conducted against 21 state-of-the-art face anti-spoofing techniques from existing literature. The graph below presents the accuracy levels reported in those methods alongside the accuracy achieved by the proposed system.



VI. CONCLUSION AND FUTURE WORK

This project introduces a powerful deepfake detection system that blends both spatial and temporal analysis using advanced deep learning methods. It leverages CNN models like InceptionResNetV2 to capture spatial features, while RNNs with LSTM layers handle the temporal dynamics of video sequences. Together, these components help detect even the most subtle facial manipulations that are often hard to spot with the naked eye.

The model was trained and evaluated on well-known datasets such as FaceForensics, Celeb-Deepfake, and the Deepfake Detection Challenge dataset. With effective preprocessing steps, data augmentation, and careful model tuning, the system demonstrated high accuracy in distinguishing real videos from deepfakes.

This detection system plays a crucial role in the fight against digital misinformation. It helps safeguard privacy and supports the verification of digital content. Its real-world applications include verifying social media videos, assisting law enforcement, and contributing to media forensics.

Future Work

While the current system performs effectively, several areas can be enhanced in the future to increase its adaptability and accuracy:

- Real-time Processing: Optimize the model for real-time detection, enabling it to be used in live streaming platforms and surveillance systems.
- Improved Temporal Models: Incorporate more advanced temporal analysis techniques such as Transformers or 3D CNNs to better understand temporal dynamics in videos.
- Larger and Diverse Datasets: Expand training on larger datasets covering a wider variety of deepfake generation techniques to improve generalization and reduce bias.
- Explainable AI (XAI): Implement techniques to provide explanations for the system's predictions, helping users understand which parts of a video led to a "fake" classification.
- Mobile and Web Deployment: Develop lightweight models for mobile apps or browser-based tools, making deepfake detection accessible to everyday users.
- Multi-modal Detection: Integrate audio analysis with video to enhance detection capabilities for videos where audio may also be manipulated.
- This project provides a solid foundation for detecting deepfakes and can evolve into a more comprehensive system with broader applicability in future iterations

VII. ACKNOWLEDGMENT

Authors acknowledge the support from department of AI & DS, K S School of Engineering and Management for the facilities and support provided to carry out project work.

REFERENCES

[1] G. Yang, X. Li, and Y. Liu, 2017, Face Liveness Detection Based on Convolutional Neural Networks, IEEE International Conference on Automatic Face and Gesture Recognition (FG), pp. 91–96.

[2] Y. Atoum, Y. Liu, A. Jourabloo, and N. Ancona, 2017, Real-Time Face Antispoofing via Deep Learning, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4779–4788.

[3] J. Parkin and R. Grinchuk, 2018, Liveness Detection Using Deep Residual Networks, International Conference on Computer Graphics and Artificial Intelligence, pp. 1–4.

[4] K. George, P. Das, and T. Kim, 2019, Biometric Face Presentation Attack Detection with Multi-Channel CNN, IEEE Transactions on Information Forensics and Security, vol. 14, no. 3, pp. 789–801

[5] H. Ge et al., 2020, Face Anti-Spoofing by the Enhancement of Temporal Motion, IEEE CTISC.

[6] X. Liu, J. Wan, and S. Yang, 2020, Attention-Based Two-Stream CNN for Face Spoofing Detection, IEEE Transactions on Information Forensics and Security, vol. 15, pp. 1310–1322.

[7] T. Tu et al., 2020, Pixel-Level Domain Transfer for Face Anti-Spoofing, Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, pp. 12186–12193.

[8] H. Mohammadi, R. Jafari, and S. Escalera, 2021, Domain-Specific Filter Pruning for Face Anti-Spoofing, International Conference on Pattern Recognition (ICPR), pp. 687–694.

[9] Z. Guo et al., 2021, Face Anti-Spoofing Using Spatial Pyramid Pooling, IEEE ICPR.

[10] Y. Zhang, H. Li, and J. Wang, 2022, Meta-Teacher for Face Anti-Spoofing, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 5, pp. 2553–2566.

[11] S. Vandana and M. Rao, 2022, Face Anti-Spoofing Based on Convolutional Neural Networks, IEEE International Defense, Security, and Technology Applications Conference (IDSTA).

[12] A. Hamza, A. R. Javed, F. Iqbal, N. Kryvinska, A. S. Almadhor, Z. Jalil, and R. Borghol, 2022, Deepfake Audio Detection via MFCC Features Using Machine Learning, IEEE Access, vol. 10, pp. 134018–134030.

[13] H. Qi et al., 2023, A Real-Time Face Detection Method Based on Blink Detection, IEEE Access, vol. 11.

[14] Z. Yu et al., 2023, Deep Learning for Face Anti-Spoofing: A Survey, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 5.

[15] Z. Yu, A. Liu, C. Zhao, K. H. M. Cheng, X. Cheng, and G. Zhao, 2023, Flexible-Modal Face Anti-Spoofing: A Benchmark, IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 6346–6352.

[16] A. Khormali and J.-S. Yuan, 2024, Self-Supervised Graph Transformer for Deepfake Detection, IEEE Access, vol. 12, pp. 58114–58130.

[17] D. Song, N. Lee, J. Kim, and E. Choi, 2024, Anomaly Detection of Deepfake Audio Based on Real Audio Using GAN, IEEE Access, vol. 12, pp. 184311–184325.

[18] A. Patel, R. Singh, and M. Vatsa, 2024, Depth-Integrated CNN Approach for Effective Face Spoof Detection, IEEE International Conference on Computer Vision and Graphics Understanding (IC-CGU).

[19] X. Lin et al., 2024, Suppress and Rebalance: Towards Generalized Multi-Modal Face Anti-Spoofing, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 211–218.

[20] S. A. Alrethea et al., 2024, Face Anti-Spoofing Using Chainlets and Deep Learning, International Journal of Advanced Computer Science and Applications, vol. 15, no. 11.

[21] Z. Yu et al., 2023, Deep Learning for Face Anti-Spoofing: A Survey (Duplicate), IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 5.