A Novel Token-Wise Asymmetric Contrastive Learning For Robust Face Presentation Attack Detection

Libinsha. E[1] and M. K.Dwaraka[2]
[1, 2] Dept of Biomedical Engineering
[1, 2] Udaya School of Engineering, Vellamodi, India – 629204

Abstract- *Face recognition systems have become a fundamental component of modern biometric authentication. However, these systems remain vulnerable to presentation attacks such as printed photographs, replay videos, masks, and AI-generated deepfakes. Existing face anti-spoofing (FAS) methods often exhibit limited generalization when exposed to unseen attack types and cross-domain variations. This paper presents a robust FAS framework that combines token-level feature learning, contrastive representation learning, and angular margin optimization to improve liveness detection performance. The proposed framework learns discriminative feature representations by encouraging compact live-face embeddings and enhanced separation from spoof-face embeddings. Localized facial analysis enables the extraction of fine-grained liveness cues, including texture inconsistencies, illumination artifacts, and reflectance variations. Experimental evaluation demonstrates an accuracy of 90.91%, an ACER of 9.9%, and a ROC-AUC of 0.9903, indicating strong discriminative capability. The proposed system provides an effective and practical solution for enhancing biometric security against presentation attacks.*

Keywords: Face Anti-Spoofing, Presentation Attack Detection, Contrastive Learning, Angular Margin Loss, Biometrics, Deep Learning, Liveness Detection.

I. INTRODUCTION

Face recognition has become one of the most widely adopted biometric technologies because of its convenience, contactless operation, and ease of deployment. Applications include smartphone authentication, banking systems, surveillance platforms, and access-control systems. Despite significant advances in recognition accuracy, face recognition systems remain vulnerable to presentation attacks such as printed photographs, replay videos, masks, and synthetic facial media generated using deep learning techniques [1].

Recent studies have demonstrated that conventional supervised anti-spoofing approaches often struggle when confronted with unseen attack types or domain shifts caused by differences in cameras, illumination conditions, and acquisition environments [1], [2]. Consequently, improving generalization capability has become a primary research focus in modern FAS systems.

Recent advances in prompt learning, domain adaptation, contrastive learning, and domain generalization have significantly improved the ability of anti-spoofing models to learn generalized liveness representations [2]–[7]. These approaches focus on extracting intrinsic facial characteristics associated with liveness rather than relying on attack-specific artifacts.

Motivated by these developments, a face anti-spoofing framework based on token-level feature learning, contrastive representation learning, and angular margin optimization is developed. The framework focuses on capturing localized facial cues and constructing a discriminative feature space that improves robustness against spoofing attacks.

The major contributions are summarized as follows:

1. A token-level feature extraction framework for capturing localized liveness information.
2. Contrastive representation learning for enhanced feature discrimination.
3. Angular margin optimization to improve class separability.
4. A practical image- and video-based anti-spoofing system with quantitative performance evaluation.

II. LITERATURE REVIEW

Face anti-spoofing has evolved from handcrafted texture analysis to advanced deep learning frameworks. A comprehensive survey by Yu et al. reviewed recent developments in deep learning-based anti-spoofing systems and highlighted the importance of domain generalization and robust feature learning [1].

Prompt-learning approaches have recently emerged as an effective solution for improving generalization. Liu et al.

introduced CFPL-FAS, which employs class-free prompt learning to improve robustness against unseen attacks [3]. Guo et al. further enhanced generalization through style-conditional prompt token learning, enabling adaptive feature extraction under varying attack conditions [4].

Domain adaptation methods have also demonstrated significant benefits. Liu et al. proposed domain generalized pretraining combined with source-free domain adaptation to improve cross-domain performance [2]. Similarly, bottom-up domain prompt tuning was introduced to learn domain-invariant facial representations and improve generalization capability [5].

Zheng et al. investigated unified representation learning for missing-modality face anti-spoofing, demonstrating the importance of modality-invariant and modality-specific feature representations [6]. Contrastive learning has increasingly been adopted in anti-spoofing research. Song et al. demonstrated that supervised contrastive learning improves discriminative feature learning in spectral imaging-based anti-spoofing systems [7]. Deb and Jain proposed localized facial analysis for generalized anti-spoofing, showing that local regions provide important liveness cues [8]. Bian et al. introduced multiple explainable and generalizable cues to improve robustness and interpretability in face anti-spoofing systems [9].

Semi-supervised learning approaches have also been investigated to address the limited availability of labeled datasets. Sergievskiy et al. demonstrated that semi-supervised learning can improve generalization performance while reducing dependence on extensive annotations [10].

These studies collectively indicate that localized feature extraction, contrastive representation learning, and domain-aware learning strategies provide promising directions for improving anti-spoofing performance.

III. PROPOSED METHODOLOGY

A. System Architecture

The proposed face anti-spoofing framework integrates token-level feature learning, contrastive representation learning, and angular margin optimization to distinguish genuine facial presentations from spoofing attacks. As shown in Fig. 1, the framework consists of seven major stages: input acquisition, face detection and preprocessing, patch/token generation, feature extraction, contrastive representation learning, angular margin optimization, and classification.

The system processes facial images through a sequence of feature learning modules designed to capture both local and global liveness characteristics. The final output is a binary classification indicating whether the input sample belongs to the live-face or spoof-face class.

B. Input Acquisition and Preprocessing

The framework accepts RGB facial images collected from face anti-spoofing datasets and real-world acquisition systems. Since facial images may contain variations in pose, illumination, scale, and background conditions, a preprocessing stage is employed to ensure consistency before feature extraction.

Face detection is performed using the RetinaFace detector, which accurately localizes facial regions within the input image. The detected face is subsequently aligned to standardize facial orientation and reduce geometric variations. Image normalization is then applied to improve robustness against illumination differences and acquisition inconsistencies.

The preprocessing stage produces normalized facial images that provide a consistent input for subsequent token generation and feature learning processes.

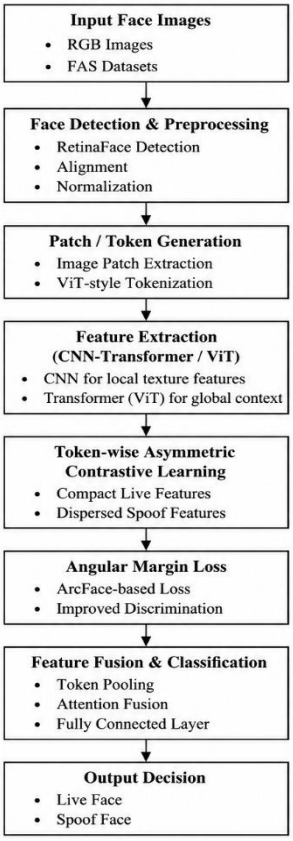

Fig. 1. Proposed face anti-spoofing framework.

C. Patch and Token Generation

Instead of processing the entire facial image as a single entity, the proposed framework partitions each face image into multiple local patches. This strategy enables the model to focus on fine-grained facial characteristics that contain valuable liveness information.

Each image patch is transformed into a token representation following a Vision Transformer (ViT)-style tokenization process. The generated tokens preserve localized facial details such as skin texture, reflection patterns, illumination inconsistencies, and micro-structural artifacts commonly associated with presentation attacks.

Token-level processing provides improved sensitivity to subtle spoofing cues that may not be adequately captured through conventional global feature extraction approaches.

D. Feature Extraction

Feature extraction is performed using a hybrid CNN–Transformer architecture that combines the strengths of convolutional neural networks and transformer-based models.

The convolutional component extracts low-level texture features and local spatial patterns that are essential for identifying spoofing artifacts. Simultaneously, transformer layers model long-range relationships among facial regions, enabling the framework to capture global contextual information.

The resulting feature representation integrates local texture characteristics and global structural information, producing discriminative embeddings suitable for liveness analysis.

E. Contrastive Representation Learning

The extracted token representations are processed through a contrastive representation learning module. The objective of this module is to construct a discriminative feature space in which samples belonging to the same class remain close to one another while samples from different classes remain well separated.

For genuine facial presentations, the learned representations are encouraged to form compact feature clusters due to their consistent biological characteristics. Spoof samples exhibit greater variability because of differences in attack media, printing artifacts, replay devices, and presentation conditions. Consequently, the learned feature space naturally promotes stronger discrimination between live and spoof classes.

F. Angular Margin Optimization

Although contrastive learning improves feature separation, additional optimization is required to maximize inter-class discrimination. Therefore, angular margin loss is incorporated into the training process.

Angular margin optimization increases the angular distance between live and spoof feature embeddings while simultaneously reducing intra-class variation. This process creates a more structured and discriminative embedding space with clearer decision boundaries.

The enhanced feature separability improves classification accuracy and strengthens resistance against challenging presentation attacks that exhibit visual similarities to genuine facial images.

G. Feature Fusion and Classification

The discriminative token-level representations generated by the previous modules are aggregated through token pooling and attention-based feature fusion mechanisms. The fusion process combines information from multiple facial regions to create a compact global representation of the input face.

The fused feature vector is subsequently passed to a fully connected classification layer that estimates the probability of the sample belonging to either the live-face or spoof-face category.

A softmax activation function is employed to produce the final classification scores. The output of the framework is a binary decision indicating whether the input image represents a genuine live face or a spoof presentation attack.

IV. IMPLEMENTATION

A web-based interface was developed to demonstrate real-time anti-spoofing functionality.

A. Image-Based Detection

The image analysis module accepts a facial image and classifies it as either a live or spoof presentation. Prediction confidence and probability distributions are displayed to facilitate result interpretation.

B. Video-Based Detection

The video analysis module extracts sixteen uniformly distributed frames from an uploaded video. Each frame is independently evaluated, and the final decision is obtained through aggregation of frame-level predictions.

C. Evaluation Metrics

Performance is evaluated using standard biometric metrics:

- Attack Presentation Classification Error Rate (APCER)
- Bona Fide Presentation Classification Error Rate (BPCER)
- Average Classification Error Rate (ACER)
- Equal Error Rate (EER)
- Accuracy
- Receiver Operating Characteristic Area Under the Curve (ROC-AUC)

These metrics provide a balanced assessment of both security and usability performance.

V. RESULTS AND DISCUSSION

The developed face anti-spoofing system was evaluated through image-based detection, video-based liveness analysis, and quantitative performance assessment. A web-based interface was implemented to demonstrate the practical deployment of the proposed framework. Figure 2 presents the home page of the developed system, which supports both image and video analysis while providing access to evaluation metrics and performance statistics.

The image analysis module accepts facial images and classifies them as either live or spoof presentations. Figure 3 illustrates an example spoof detection result obtained using the proposed framework. The uploaded facial image was classified as a spoof presentation with a confidence score of 95.46%. The corresponding live and spoof probabilities were 4.54% and 95.46%, respectively. The high confidence associated with the prediction demonstrates the ability of the framework to identify presentation attacks by exploiting localized liveness cues extracted through token-level feature analysis. Texture inconsistencies, illumination variations, and reflectance artifacts contribute to the discrimination between genuine and spoof samples.

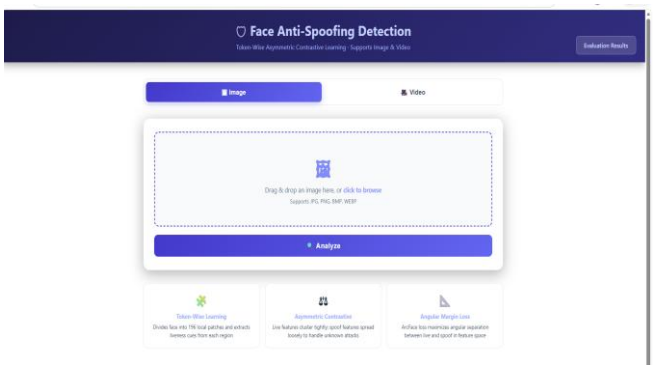

Fig. 2. Home page of the developed face anti-spoofing system showing image and video analysis modules.

Video-based liveness analysis was performed by extracting sixteen uniformly distributed frames from each uploaded video and evaluating them independently. Figure 4 presents a representative live-video prediction result. The uploaded video was classified as a live presentation with an overall confidence score of 99.07%. All sixteen extracted frames were correctly classified as live, while no spoof frames were detected. The frame-wise predictions remained highly consistent throughout the video sequence, indicating that the proposed framework effectively captures temporal evidence and maintains stable performance across multiple observations. Such behavior is particularly important for mitigating replay attacks and dynamic presentation attacks, where decisions based on a single frame may be unreliable.

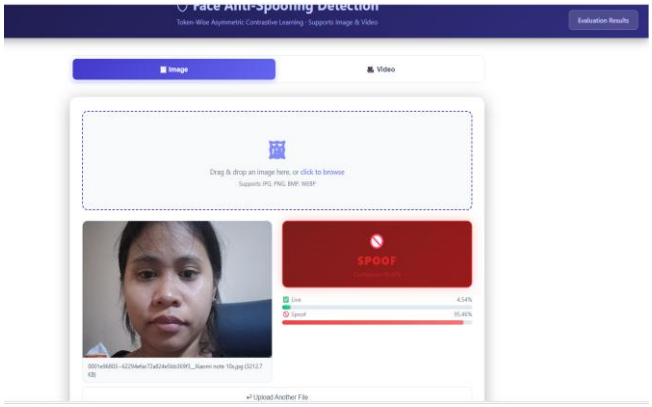

Fig. 3. Image-based face anti-spoofing result showing a spoof sample classified with 95.46% confidence.

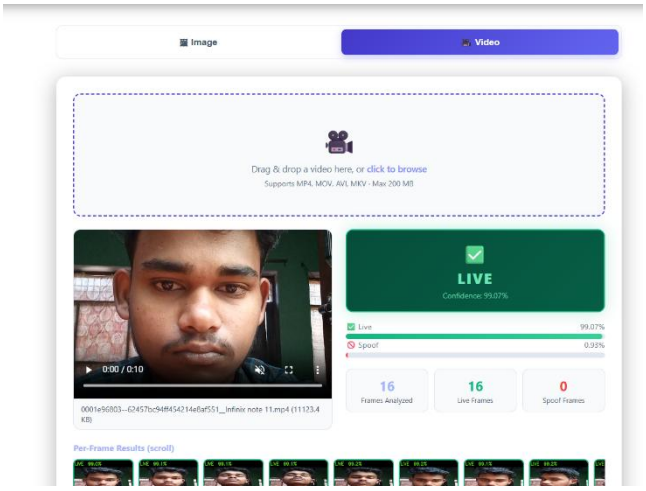

Fig. 4. Video-based liveness detection result showing all extracted frames classified as live.

Comprehensive performance evaluation was conducted using standard face anti-spoofing metrics, including Attack Presentation Classification Error Rate (APCER), Bona Fide Presentation Classification Error Rate (BPCER), Average Classification Error Rate (ACER), Equal Error Rate (EER), overall accuracy, and Receiver Operating Characteristic Area Under the Curve (ROC-AUC). Figure 5 presents the evaluation dashboard containing the performance metrics, confusion matrix, ROC curve, threshold analysis, and test-set distribution.

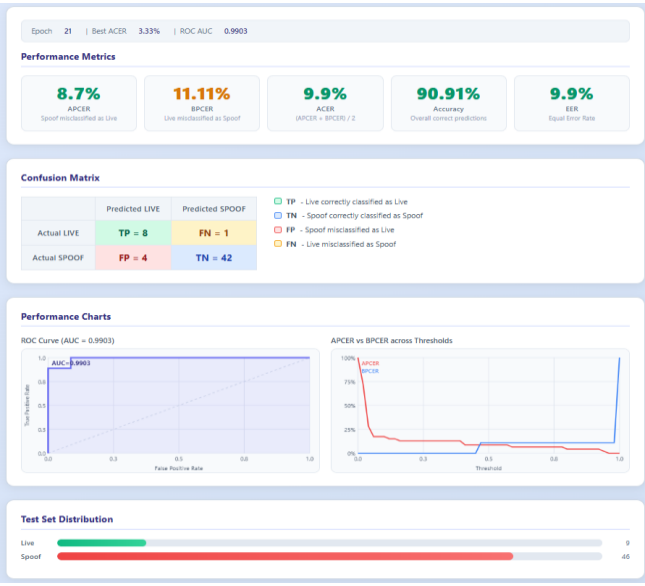

Fig. 5. Performance evaluation results including APCER, BPCER, ACER, ROC-AUC, confusion matrix, and threshold analysis.

The quantitative results are summarized in Table I. The framework achieved an overall classification accuracy of 90.91%, indicating reliable discrimination between live and spoof presentations. The APCER and BPCER values were 8.70% and 11.11%, respectively, resulting in an ACER of 9.90%. The low ACER demonstrates balanced performance between attack detection capability and genuine-user acceptance. Furthermore, the ROC-AUC value of 0.9903 indicates excellent separability between live and spoof classes, confirming the effectiveness of the learned feature representations.

Table I. Quantitative performance evaluation of the proposed face anti-spoofing framework.

Metric	Value
APCER (%)	8.70
BPCER (%)	11.11
ACER (%)	9.90
Accuracy (%)	90.91
EER (%)	9.90
ROC-AUC	0.9903

A detailed analysis of the classification outcomes is presented in Table II. Out of 55 test samples, the framework correctly classified 50 samples. Eight live samples were correctly recognized, while only one live sample was incorrectly classified as spoof. Similarly, forty-two spoof samples were correctly detected, with four spoof samples misclassified as live. The confusion matrix indicates that the framework maintains a strong balance between sensitivity to spoof attacks and preservation of genuine-user accessibility.

Table II. Classification results of the proposed face anti-spoofing framework on the test dataset.

Actual Class	Predicted Live	Predicted Spoof
Live	8	1
Spoof	4	42

The ROC curve shown in Fig. 5 further demonstrates the strong discriminative capability of the proposed framework. The near-perfect ROC-AUC value indicates that the learned feature space effectively separates genuine and attack presentations across varying decision thresholds. In addition, the APCER–BPCER threshold analysis reveals stable performance over a broad operating range, suggesting robustness to threshold selection during deployment.

The experimental findings demonstrate that token-level feature extraction, contrastive representation learning, and angular margin optimization collectively contribute to improved liveness detection performance. Localized feature

analysis enables the identification of subtle spoofing artifacts, while contrastive learning promotes the formation of discriminative feature representations. Angular margin optimization further enhances class separability by increasing inter-class distance and reducing intra-class variation. The resulting framework achieves reliable face anti-spoofing performance and exhibits strong potential for practical biometric authentication applications.

VI. CONCLUSION AND FUTURE WORK

This paper presented a face anti-spoofing framework that integrates token-level feature extraction, contrastive representation learning, and angular margin optimization for reliable liveness detection. The framework was designed to address the limitations of conventional anti-spoofing approaches, particularly their susceptibility to unseen presentation attacks and variations in acquisition conditions. The proposed architecture combines localized facial analysis with discriminative representation learning to improve the separation between genuine and spoof facial presentations.

Experimental evaluation demonstrated the effectiveness of the proposed framework in both image-based and video-based anti-spoofing scenarios. The results achieved an overall accuracy of 90.91%, an ACER of 9.90%, and a ROC-AUC of 0.9903. The confusion matrix analysis further indicated balanced classification performance, with a high number of correctly identified live and spoof samples. These findings confirm that the integration of token-level feature learning, contrastive representation learning, and angular margin optimization provides a robust solution for face anti-spoofing applications.

Future work will focus on improving cross-domain generalization and robustness against emerging attack types, including advanced deepfake and synthetic face generation techniques. The incorporation of multimodal biometric information, such as depth, infrared, and thermal facial data, may further enhance detection capability. In addition, lightweight model optimization and edge-device deployment can facilitate real-time implementation on mobile and embedded platforms. Further evaluation on larger and more diverse benchmark datasets will also contribute to improving the reliability and scalability of the proposed framework for practical biometric authentication systems.

REFERENCES

[1] Z. Yu, Y. Qin, X. Li, C. Zhao, Z. Lei, and G. Zhao, "Deep Learning for Face Anti-Spoofing: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 5609–5631, May 2023, doi: 10.1109/TPAMI.2022.3215850.

[2] Y. Liu, Y. Chen, W. Dai, M. Gou, C.-T. Huang, and H. Xiong, "Source-Free Domain Adaptation With Domain Generalized Pretraining for Face Anti-Spoofing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5430–5448, Aug. 2024, doi: 10.1109/TPAMI.2024.3370721.

[3] A. Liu et al., "CFPL-FAS: Class Free Prompt Learning for Generalizable Face Anti-Spoofing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[4] J. Guo, H. Liu, Y. Luo, X. Hu, H. Zou, Y. Zhang, H. Liu, and B. Zhao, "Style-Conditional Prompt Token Learning for Generalizable Face Anti-Spoofing," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024.

[5] S. Q. Liu, Q. Wang, and P. C. Yuen, "Bottom-Up Domain Prompt Tuning for Generalized Face Anti-Spoofing," in *Computer Vision – ECCV 2024*, A. Leonardis et al., Eds., Lecture Notes in Computer Science, vol. 15128, Springer, Cham, 2025, doi: 10.1007/978-3-031-72897-6_10.

[6] G. Zheng, Y. Liu, W. Dai, C. Li, J. Zou, and H. Xiong, "Towards Unified Representation of Invariant-Specific Features in Missing Modality Face Anti-Spoofing," in *Computer Vision – ECCV 2024*, A. Leonardis et al., Eds., Lecture Notes in Computer Science, vol. 15080, Springer, Cham, 2025, doi: 10.1007/978-3-031-72670-5_6.

[7] C. Song, Y. Hong, J. Lan, H. Zhu, W. Wang, and J. Zhang, "Supervised Contrastive Learning for Snapshot Spectral Imaging Face Anti-Spoofing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 980–985, 2024.

[8] D. Deb and A. K. Jain, "Look Locally Infer Globally: A Generalizable Face Anti-Spoofing Approach," *arXiv preprint*, arXiv:2006.02834, 2020.

[9] Y. Bian, P. Zhang, J. Wang, C. Wang, and S. Pu, "Learning Multiple Explainable and Generalizable Cues for Face Anti-Spoofing," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2310–2314, 2022, doi: 10.1109/ICASSP43922.2022.9747677.

[10] N. Sergievskiy, R. Vlasov, and R. Trusov, "Generalizable Method for Face Anti-Spoofing with Semi-Supervised Learning," *arXiv preprint*, arXiv:2206.06510, 2022.