

## Article

# Interpretability-Driven Defense Against Physical Adversarial Attacks in Facial Recognition Systems

Maryam Suleiman<sup>1\*</sup>, Peter Ogedebe<sup>2</sup>, Usman Abubakar Idris<sup>3</sup>

1. Department of Computer Science, Baze University, Plot 686, Jabi Airport Road Bypass, Cadastral Zone, Abuja, Federal Capital Territory
2. Department of Computer Science, Baze University, Plot 686, Jabi Airport Road Bypass, Cadastral Zone, Abuja, Federal Capital Territory
3. Department of Computer Science, Baze University, Plot 686, Jabi Airport Road Bypass, Cadastral Zone, Abuja, Federal Capital Territory

\* Correspondence: [maryam8383@bazeuniversity.edu.ng](mailto:maryam8383@bazeuniversity.edu.ng)

**Abstract:** This study aims to develop and evaluate a defense mechanism, the Self-Explaining Gate (SEG), to protect facial recognition (FR) systems against physical adversarial attacks (PAAs) by combining identity recognition, feature attribution, and anatomical landmark detection. SEG uses a three-stream architecture comprising identity feature extraction, real-time Grad-CAM feature attribution, and Haar-cascade-based anatomical landmark detection, integrated through a novel Anatomical Consistency Score (ACS) that measures the overlap between model attention and biological facial landmarks. The approach was evaluated on the CASIA-FASD and SiW benchmark datasets and a custom adversarial dataset, using detection accuracy, false acceptance/rejection rates, and inference latency as evaluation metrics. Our SEG achieved 94% attack detection accuracy, rejecting 96.1% of adversarial images while accepting 90% of genuine images, with inference latency under 160 milliseconds, outperforming adversarial training, feature squeezing, and SSR-FCN baseline defenses. Performance degrades on low-resolution images and under extreme occlusion, and further validation on video-based and larger real-world datasets is required. SEG can be integrated into real-time facial recognition deployments, such as border control, banking, and security systems, to provide interpretable and anatomically grounded rejection of adversarial inputs. By making automated identity verification more transparent and resistant to spoofing, SEG has the potential to increase public trust in facial recognition systems used in security-sensitive contexts. This study is the first to combine intrinsic, feedforward explainability with anatomical landmark validation as a real-time defense mechanism, introducing the Anatomical Consistency Score and a zero-background weighting scheme.

**Keywords:** Physical Adversarial Attacks, Facial Recognition, Explainable AI, Anatomical Landmark Detection, Self-Explaining Gate

**Citation:** Suleiman, M, Ogedebe, P & Idris, U. A. Interpretability-Driven Defense Against Physical Adversarial Attacks in Facial Recognition Systems. International Journal of Human Computing Studies 2026, 8(2), 73-78

Received: 10<sup>th</sup> May 2026Revised: 26<sup>th</sup> Jun 2026Accepted: 08<sup>th</sup> Jul 2026Published: 25<sup>th</sup> Aug 2026

**Copyright:** © 2026 by the authors. Submitted for open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>)

## 1. Introduction

Facial recognition (FR) technology has found its way into many sectors, including border control, finance, and security, among others. Some of the latest models, such as ArcFace, FaceNet, and ResNet, achieve over 99% accuracy on the Labeled Faces in the Wild (LFW) benchmark dataset. Still, this high level of performance does not indicate that these models identify biological faces; rather, they identify pixel patterns learned during training. This lack of grounding in human anatomical features makes them easy targets of Physical Adversarial Attacks (PAAs).

Physical Adversarial Attacks differ from their digital counterparts. While digital attacks modify the entire image at once before it is fed into the model, physical attacks are applied to the subject itself — an adversarial patch, special eyeglasses, fake makeup, or a projected laser — so that the perturbation exists in the real world before the camera captures the image. These attacks have proven effective, with reported success rates exceeding 90% against commercial black-box face recognition systems.

Problem statement: Existing countermeasures each carry weaknesses. Presentation Attack Detection (PAD) fails against unknown attack types, adversarial training requires prior knowledge of attacker patterns and can degrade performance on normal images, and post-hoc attribution methods such as Grad-CAM, SHAP, and LIME produce computationally heavy and diffuse maps that are unsuitable for real-time inference at 30+ frames per second (FPS). Existing studies have also tended to address either adversarial defense or model interpretability in isolation, leaving the joint problem of security, utility, and speed — what we term the “triple-bind” — largely unresolved.

This study addresses this gap by proposing the Self-Explaining Gate (SEG), a real-time defense that unifies identity recognition, feature attribution, and anatomical landmark detection. The specific contributions are: (1) the Self-Explaining Gate (SEG), a real-time feedforward feature attribution mechanism that bypasses time-consuming post-hoc iterative sampling; (2) the Anatomical Consistency Score (ACS), a new bounded metric that quantifies the alignment between biological facial landmarks and the model's attention map; and (3) a zero-background structural tracking heuristic that eliminates the “mathematical floor trap” of cascade classifiers by assigning no weight to background regions. Together, these contributions reframe face authentication from “black-box matching” into anatomical logic validation, in which verification succeeds only if the model's attention is directed at the correct anatomical regions [1], [2].

### 1.1 Theories Underpinning Study

Realistic physical adversarial attacks have become a major safety concern for deployed facial recognition systems. Wang et al. [1] presented a comprehensive review classifying these attacks by physical medium, including adversarial stickers, accessories such as eyeglasses, infrared lights, and illumination-based methods. Sharif et al. [2] showed how adversarial eyeglasses could cause impersonation attacks on commercial FR systems, and Brown et al. [3] introduced the adversarial patch method, demonstrating that small localized perturbations could cause advanced classifiers to fail. Hu et al. [4] proposed AdvIB, a black-box physical attack method requiring no prior information about the target model, compromising thermal infrared detection through optimal placement of thermally emissive elements and achieving attack success rates of 40.8% at 8.6-metre distances.

Existing defenses fall into several categories. Adversarial training, introduced by Madry et al. [5], trains models on a mixture of adversarial and clean examples; its drawback is that it requires prior knowledge of attack signatures and often degrades performance on benign inputs. Feature squeezing reduces the feature space to suppress adversarial features but can also degrade overall system performance. Deb [6] introduced FaceGuard, which combines adversarial training with detection and purification in a single self-supervised system. For Presentation Attack Detection, methods such as SSR-FCN achieve a genuine acceptance rate of 65% with a false-detection ratio below 3% on unseen spoof attacks, but remain weak against novel classes of threats.

Model interpretability and adversarial robustness have also been shown to be closely connected. Tao et al. [7] proposed an attribute-guided detection method that identifies neurons central to each facial attribute, detecting attacks with 94% accuracy. Jeong et al. [8] developed GNIME, a defense against explanation-aware inversion attacks that reduced transferable classification precision of FR models by up to 84.8%. Borsukiewicz et al. [9] introduced Layer Embedding Activation Mapping (LEAM), showing that face recognition models concentrate attention on a small number of facial regions; nose regions

accounted for 18.9–29.7% of critical regions, and altering just 1% of key pixels reduced cosine similarity by 0.1148.

Taken together, this literature indicates that existing studies have focused on either defending against adversarial attacks or on model interpretability in isolation, and that post-hoc explainability methods remain too computationally heavy to support real-time validity checks. This study builds on this literature by introducing an intrinsic explainability mechanism embedded directly in the feedforward process [10], [11], [12], [13].

## 2. Materials and Methods

### 2.1 Research Design

This is an empirical, experimental study in which the proposed SEG framework was implemented and evaluated quantitatively against benchmark datasets and existing defense methods.

### 2.2 Problem Formulation

A facial recognition system is defined as a function  $f$  mapping an input image  $x$  to an identity prediction  $y$ . A physical adversarial attack seeks a perturbation  $\delta$ , applied in the physical world, such that the prediction on the perturbed image differs from the prediction on the original image, while the perturbed image remains perceptually similar to the original. The defense objective is to learn a gating function that rejects adversarial inputs while accepting benign ones, operating in real time with minimal computational overhead.

### 2.3 SEG Architecture

SEG consists of a three-stream architecture processed concurrently during inference. Stream A, Identity Feature Extraction, uses a standard FR backbone (e.g., ArcFace, FaceNet) to produce identity embeddings. Stream B, Feature Attribution, generates real-time Grad-CAM attention maps showing where the model focuses; unlike traditional post-hoc Grad-CAM, this implementation produces attribution maps during the forward pass, adding less than 50 milliseconds of computational overhead. Stream C, Landmark Detection, applies a zero-background structural tracking protocol in which Haar cascade classifiers localize biological facial landmarks — the primary face bounding box, the left and right ocular regions, and the nasal centre.

A priority weighting scheme assigns a weight of 0.0 to background regions (eliminating the mathematical floor trap), a baseline weight of 1.0 to the face region, and a premium weight of 2.5 to the eye and nose regions. Where the eyes are obscured (e.g., by adversarial glasses), a fallback protocol uses anthropometric proportions to project expected eye positions from known facial geometry.

### 2.4 Anatomical Consistency Score (ACS)

The ACS is a bounded metric between 0 and 1 quantifying the alignment between Stream B (model attention) and Stream C (biological landmarks). It is computed as the weighted sum of the Intersection-over-Union (IoU) between each attention region and its corresponding anatomical landmark region, divided by the sum of the priority weights. Authentication is granted only when the ACS meets or exceeds a calibrated threshold. Previous heuristic masking approaches assigned non-zero weights to background regions, creating a mathematical floor that allowed adversarial images to pass detection thresholds; the zero-background weighting scheme used here eliminates this vulnerability.

Total inference latency is the sum of the recognition latency (under 100 milliseconds on modern GPUs), the attribution latency (under 50 milliseconds), and the landmark-detection latency (under 10 milliseconds using Haar cascades), giving a total latency under 160 milliseconds, which is suitable for real-time deployment at 30+ FPS.

## 2.5 Datasets

Three datasets were used for evaluation. The CASIA Face Anti-Spoofing Dataset (CASIA-FASD) comprises 50 subjects (20 for training, 30 for testing) and 2,408 images evaluated across three attack types — warped photo, cut photo, and video replay — at three resolutions. The Spoof in the Wild (SiW) dataset comprises 165 subjects and 4,478 videos at 1080p resolution across four sessions with varying distance, pose, illumination, and expression, from which 7,586 images were evaluated. A custom adversarial dataset of 164 images (10 genuine, 154 adversarial) featuring adversarial patches, eyeglasses, and face-mask occlusions was also constructed for this study.

## 2.6 Evaluation Metrics and Implementation

Performance was assessed using the False Acceptance Rate (rate of accepting impostors), False Rejection Rate (rate of rejecting genuine users), Attack Detection Rate (rate of correctly identifying adversarial inputs), Inference Latency (end-to-end processing time), and the proposed Anatomical Consistency Score. The FR backbone used was ArcFace with a ResNet-50 encoder; attribution used a modified Grad-CAM with optimized backpropagation; landmark detection used OpenCV Haar cascades (face, eyes, nose); experiments were run on an NVIDIA GTX 1080 Ti GPU using PyTorch and OpenCV.

## 3. Results

SEG achieved 94% detection accuracy across all attack types, comparable to state-of-the-art interpretability-based detection methods. On the custom adversarial test set, SEG correctly rejected 148 of 154 adversarial images (96.1% rejection rate) while accepting 9 of 10 genuine images (90% acceptance rate).

Table 1 compares SEG against three existing defenses. Adversarial Training achieved 55–70% detection accuracy with a 0.5% False Acceptance Rate (FAR), a 12.3% False Rejection Rate (FRR), and 100 ms latency. Feature Squeezing achieved 55% detection accuracy with a 1.2% FAR, 15.7% FRR, and 200 ms latency. SSR-FCN achieved 65% detection accuracy with a 2.0% FAR and under 4 ms latency, though FRR was not reported. SEG achieved 94% detection accuracy with a 0.3% FAR, a 10.1% FRR, and under 160 ms latency — the highest detection accuracy and lowest FAR of the methods compared, with latency remaining within real-time bounds.

**Table 1.** Comparison of SEG with Existing Defense Methods

| Method               | Detection Accuracy | FAR         | FRR          | Latency (ms)   |
|----------------------|--------------------|-------------|--------------|----------------|
| Adversarial Training | 55–70%             | 0.5%        | 12.3%        | 100            |
| Feature Squeezing    | 55%                | 1.2%        | 15.7%        | 200            |
| SSR-FCN              | 65%                | 2.0%        | N/A          | <4             |
| <b>SEG (Ours)</b>    | <b>94%</b>         | <b>0.3%</b> | <b>10.1%</b> | <b>&lt;160</b> |

Analysis of ACS scores showed a clear separation between benign and adversarial inputs: benign samples scored between 0.72 and 0.89, while adversarial samples scored between 0.31 and 0.48. This separation supports the ACS as an effective discriminative metric for adversarial detection. An ablation study further showed that removing either the zero-background weighting or the priority weighting scheme reduced detection accuracy by 12–18%, confirming that both components contribute meaningfully to performance.

#### 4. Discussion

These results indicate that SEG addresses the “triple-bind” of security, utility, and speed identified in the introduction: it provides 94% detection accuracy against diverse physical attacks (security), a 90% acceptance rate on benign samples (utility), and under 160 ms inference latency suitable for 30+ FPS operation (speed). This is consistent with, and extends, the interpretability-based detection literature [7], [9], which established a link between attention localisation and adversarial robustness but did not embed this link in a real-time, feedforward architecture.

Compared with adversarial training and feature squeezing [5], which trade off accuracy against benign-sample performance, SEG's anatomically grounded attention validation achieves higher detection accuracy without a corresponding rise in false rejections. Compared with SSR-FCN, which offers lower latency but substantially lower detection accuracy, SEG offers a more favourable balance for security-critical deployments where missed detections carry a higher cost than a modest increase in processing time.

Nevertheless, SEG has limitations. Landmark detection degrades on low-resolution images (below 320×240 pixels), and heavy occlusions such as full-face masks trigger fallback protocols that may reduce accuracy. Although SEG is attack-agnostic in general, fundamentally new attack scenarios may require re-calibration of the priority weighting scheme or the ACS threshold. These limitations echo the broader literature's concern that anatomical or landmark-dependent defenses may struggle under degraded imaging conditions [14], [15].

#### 5. Conclusion

This study introduced the Self-Explaining Gate (SEG), an interpretability-based defense against physical adversarial attacks on facial recognition systems. Unlike black-box pattern-matching techniques, SEG validates anatomical logic by cross-checking model attention against biological landmarks. This approach achieved a 94% detection rate across diverse physical attack types while maintaining real-time performance under 160 milliseconds. The Anatomical Consistency Score provides a bounded mathematical measure of the match between model attention and biological landmarks, and the background-free structural tracking technique removes the mathematical floor trap that affected previous approaches. Future work should extend SEG to video-based facial recognition through temporal consistency modelling, incorporate differentiable attention methods for adaptive landmark weighting, and evaluate SEG on edge devices for outdoor security deployments. Overall, SEG offers a viable, interpretable path toward protecting real-world facial recognition systems from physical attacks.

#### REFERENCES

- [1] M. Wang, J. Zhou, T. Li, G. Meng, and K. Chen, “A survey on physical adversarial attacks against face recognition systems,” *Neurocomputing*, vol. 669, Art. no. 132485, 2026, doi: 10.1016/j.neucom.2026.132485.
- [2] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, “Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition,” in *Proc. ACM Conf. Computer and Communications Security (CCS)*, 2016, pp. 1528–1540.
- [3] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, “Adversarial patch,” *arXiv preprint arXiv:1712.09665*, 2017. [Online]. Available:
- [4] C. Hu, W. Shi, T. Jiang, W. Yao, L. Tian, X. Chen, J. Zhou, and W. Li, “Adversarial infrared blocks: A multi-view black-box attack to thermal infrared detectors in physical world,” *Neural Networks*, vol. 175, Art. no. 106310, 2024, doi: 10.1016/j.neunet.2024.106310.
- [5] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” *arXiv preprint arXiv:1706.06083*, 2017. [Online]. Available:
- [6] D. Deb, “Towards robust and secure face recognition: Defense against physical and digital attacks,” Ph.D. dissertation, Michigan State University, East Lansing, MI, USA, n.d.
- [7] G. Tao, S. Ma, Y. Liu, and X. Zhang, “Attacks meet interpretability: Attribute-steered detection of adversarial samples,” in *Advances in Neural Information Processing Systems*, vol. 31, 2018.

- [8] H. Jeong, S. Lee, S. J. Hwang, and S. Son, "Learning to generate inversion-resistant model explanations," in *Advances in Neural Information Processing Systems*, vol. 35, 2022.
- [9] P. J. Borsukiewicz, J. Klein, J. Samhi, and T. F. Bissyandé, "Explainable AI for analyzing person-specific patterns in facial recognition tasks," *arXiv preprint arXiv:2509.17457*, 2025. [Online]. Available:
- [10] R. R. Selvaraju, A. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017, pp. 618–626.
- [11] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4690–4699.
- [12] D. Zucker, "Deeply explainable artificial neural network," *arXiv preprint arXiv:2505.06731*, 2025.
- [13] A. T. Balci, M. M. Ebeid, P. V. Benos, D. Kostka, and M. Chikina, "An intrinsically interpretable neural network architecture for sequence-to-function learning," *Bioinformatics*, vol. 39, no. S1, pp. i413–i422, 2023, doi: 10.1093/bioinformatics/btad271.
- [14] Y. Chong, Y. Wang, and J. Li, "Automated anatomical landmark detection on 3D facial images using U-Net-based deep learning algorithm," *Quantitative Imaging in Medicine and Surgery*, vol. 14, no. 3, pp. 2466–2474, 2024.
- [15] Z. Zhang, J. Yan, S. Liu, Z. Lei, and S. Z. Li, "CASIA-FASD: A dataset for face anti-spoofing," in *Proc. IEEE Int. Conf. Biometrics: Theory, Applications and Systems (BTAS)*, 2012, pp. 26–31.