Adversarial Attack Challenge for Secure Face Recognition 2025

João Tremoço[1] Iurii Medvedev[2] Nuno Freitas[1] Andreia Costa[1] Diogo Nunes[2] Niklas Bunzel[3,4,5] Lukas Graner[3,4] Nicholas Göller[3,4] Lorenzo Pellegrini[6] Nicolò Di Domenico[6] Guido Borghi[7] Monson Verghese[8] Shruti Bhilare[8] Avik Hati[9] Miguel Lourenço[1] Nuno Gonçalves[2]

[1]Youverse, Portugal [2]Institute of Systems and Robotics, University of Coimbra [3]Fraunhofer SIT [4]ATHENE [5]TU-Darmstadt [6]University of Bologna, Italy [7]University of Modena and Reggio Emilia, Italy [8]Dhirubhai Ambani University [9]NIT Tiruchirapalli

Abstract

Adversarial attacks pose a significant threat to the reliability of biometric systems, particularly in security-critical applications such as identity verification and access control. Ensuring robustness against such attacks is essential for the safe deployment of face recognition technologies in real-world scenarios. To advance this goal, the 2025 Adversarial Attack Challenge for Secure Face Recognition was organized as part of the International Joint Conference on Biometrics (IJCB) 2025.

The competition focused on two main tracks: Detection, where the objective was to determine whether a given face image is clean or adversarial, and Resilience, which aimed to evaluate recognition systems under adversarial perturbations. Participants were provided with a standardized dataset derived from CelebA and LFW, encompassing both clean samples and adversarial images crafted using ten diverse attack methods targeting evasion and impersonation scenarios. To ensure fairness and reproducibility, all models were trained solely on the data provided, with support from a custom open source adversarial attack package tailored for face recognition.

In addition to benchmarking adversarial robustness, the challenge contributes to the research community by releasing the data set and the extensible attack package, allowing further investigation of secure and reliable face recognition systems.

1. Introduction

Face recognition (FR) systems have become integral to a wide range of real-world applications, including public surveillance, border control, secure facility access, financial authentication, and personal device unlocking. Their convenience, non-invasiveness, and increasing accuracy have contributed to their rapid adoption across both governmental and commercial sectors. As these systems become more prevalent in security-critical environments, ensuring their robustness and trustworthiness is of utmost importance.

Despite their impressive performance under benign conditions, FR systems are notably susceptible to adversarial attacks. Such attacks are carefully crafted perturbations to input images, often imperceptible to the human eye, but capable of causing significant degradation in recognition performance. These perturbations exploit the vulnerabilities of deep learning models, leading to incorrect predictions or misidentifications, as shown in Figure 1. Adversarial attacks on FR systems are typically categorized into two main types:

- **Evasion** attacks aiming to cause a failure in recognizing a legitimate user. For example, an authorized individual's image may be perturbed such that the system fails to match it with their enrolled identity, thereby denying access.
- **Impersonation** attacks that attempt to manipulate an input image to be falsely recognized as a different individual. Such attacks can result in unauthorized users being granted access by mimicking the facial features of a target identity.

The existence of such vulnerabilities poses serious concerns and hinders wider adoption of FR systems, especially in high-stakes scenarios where the consequences of misidentification can be severe, ranging from privacy breaches to national security threats. Furthermore, the transferability of adversarial examples across different models and the potential for physical-world attacks exacerbate the threat landscape.

Most, if not all, FR systems focus on the extraction of a biometric signature of the face that encodes the identity of the person in a resilient and invariant fashion. Adversarial attacks target the **vulnerabilities** of this embedding extrac-

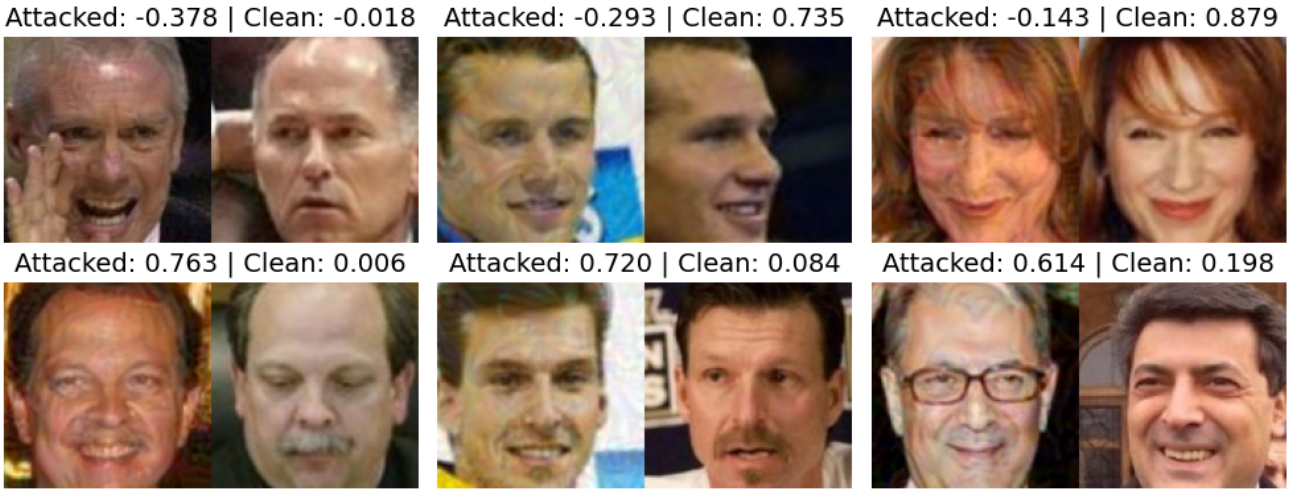

Figure 1. Examples of pair attacks. In each pair, the left image is the attacked one. The top row shows evasion attacks; the bottom row shows impersonation attacks. Similarity scores by an off-the-shelf FR system compare clean and attacked pairs. The clean label refers to the similarity before the attack, while the attack label refers to the similarity post-attack. Best seen in color.

tion stage. By introducing small, carefully computed perturbations to the input image, an attacker can manipulate the resulting embedding [31, 10]. These perturbations are often imperceptible to the human eye but sufficient to cause significant changes in the system's decision. Many adversarial attacks rely on access to the model gradients to optimize the perturbation –these are known as **white box attacks** [11, 2]. However, in real-world applications, such access is often restricted. Consequently, **black-box attacks** have emerged as a potent threat, relying only on input-output queries to the system or using surrogate models trained to mimic the behavior of the target system [26, 4]. The challenge addressed in this paper focuses on black-box scenarios, reflecting the realistic threat model faced by deployed FR systems. This includes attacks that do not require access to the architecture or parameters of the target model, making them harder to defend against and more broadly applicable.

Recent studies show that traditional defenses [1, 34], such as input pre-processing or adversarial training, are often insufficient against unseen and adaptive attacks. This highlights the pressing need for comprehensive evaluation frameworks and standardized benchmarks that can assess and improve the resilience of FR systems under adversarial conditions. In response to these challenges, the 2025 Adversarial Attack Challenge for Secure Face Recognition is presented as part of the International Joint Conference on Biometrics (IJCB) 2025. The goal of this initiative is to foster progress in both detection and mitigation of adversarial threats and to promote reproducible research by providing common datasets, open source tools, and a competitive yet collaborative evaluation mechanisms.

1.1. Purpose

The susceptibility of FR systems to adversarial attacks necessitates the development of robust models capable of withstanding such manipulations. Traditional evaluation metrics, which often focus solely on accuracy under benign conditions, fail to capture a model's resilience to adversarial perturbations. Therefore, there is a pressing need for benchmarks that assess FR systems' robustness against such manipulations.

An alternative defense against such attacks involves detecting adversarial manipulations in input images and rejecting them as invalid inputs. This solution is easier to integrate into existing FR systems as a modular component.

However, the lack of large, publicly available standardized datasets containing adversarial examples and adversarial generation tools hinders the development and comparison of defense mechanisms. Establishing such resources is crucial for advancing research in this domain and for fostering the development of accurate and secure FR systems. Moreover, although methods of evaluating FR systems' performance on adversarial data exist, they are not widespread and often do not address the detection problem comprehensively.

1.2. Contributions

The 2025 Adversarial Attack Challenge makes several significant contributions to the field of adversarial robustness in face recognition:

- **Standardized Evaluation Framework**: The challenge provided a controlled and reproducible envi-

ronment for evaluating adversarial detection and resilience, facilitating fair comparisons between different approaches.

- **Diverse and Realistic Datasets**: Participants were provided with datasets derived from CelebA and LFW, containing a balanced mix of clean images and adversarial examples generated using ten different attack methods. This diversity ensured that the models were tested against a wide range of adversarial scenarios.
- **Attack Diversity**: The provided package and training dataset included ten different attacks. The private test set included three unseen attacks as well. The mix of attack sophistication and variety aimed to aid in generalization capabilities for both attack detection and resilience.
- **Open-Source Adversarial Attack Package**: To support the development and evaluation of robust models, the organizers released an open-source adversarial attack package tailored for face recognition tasks. This resource enabled participants to understand and simulate various attack strategies.

These contributions aim to advance the development of secure and reliable face recognition systems capable of withstanding adversarial threats.

1.3. Paper Structure

The remainder of this paper is organized as follows: Section 2 briefly introduced some benchmarks already available in the field and explains why this challenge and benchmark tools are valuable in the biometrics field. Section 3 details the methodology used for creating and building the challenge datasets. Section 4 presents the results and analysis of the submitted solutions. Finally, Section 5 concludes the paper and discusses future work.

2. Related Work

The growing vulnerability of face recognition (FR) systems to adversarial attacks has stimulated significant research into understanding, evaluating, and mitigating such threats. This section reviews key prior efforts, including benchmarks, adversarial defense strategies, detection mechanisms, and evaluation frameworks.

Numerous studies have demonstrated the effectiveness of adversarial attacks on deep learning models used for face recognition. Goodfellow et al. [11] introduced the Fast Gradient Sign Method (FGSM), laying the foundation for gradient-based perturbations. Carlini and Wagner [2] later proposed more powerful optimization-based attacks. These methods have been adapted to FR systems to create perturbations that are imperceptible to humans but cause identity mismatches.

Although prior efforts have explored adversarial example detection as a defense strategy, challenges remain. Metzen et al.[21] proposed auxiliary networks to detect perturbations, while Pang et al. [25] introduced confidence-based rejection mechanisms. In the context of face recognition, Li et al. [17] analyzed both input-level and feature-level indicators of adversarial manipulation. In the context of FR, both evasion and impersonation attacks have been widely studied. Sharif et al. [31] demonstrated physical attacks using adversarial accessories (e.g., eyeglass frames), while Dong et al. [10] developed decision-based black-box attacks that can fool FR systems with limited information. However, these methods often struggle to generalize across different attack types or transfer effectively between datasets. This highlights the need for a dedicated benchmark that rigorously evaluates detection performance under realistic, black-box, face-specific threat models.

2.1. Benchmarks and Evaluation Frameworks

Several benchmarks have been introduced to evaluate the adversarial robustness of face recognition (FR) systems:

FATE [24] benchmark by NIST conducts large-scale assessments of real-world FR performance, including presentation attack detection, but does not explicitly address adversarial machine learning.

RobFR [38] provides a structured benchmark for evaluating model robustness across a range of digital adversarial attacks and threat models. **FACESEC** [33] enables fine-grained robustness evaluation by considering attacker knowledge, perturbation types (digital and physical), and defense strategies. **TALFW** [42] extends the LFW dataset with transferable adversarial examples to support black-box and cross-model evaluations. **AIoT-Face** [40] targets adversarial robustness in resource-constrained, edge-deployed FR scenarios.

While each of these efforts contributes valuable insights, most focus on either adversarial resilience or attack diversity, without integrating adversarial detection as a standalone evaluation goal. Moreover, many lack openly available datasets or attack-generation pipelines, limiting reproducibility and practical adoption.

To address this gap, the Adversarial Attack Challenge was designed to unify both *detection* and *resilience* evaluation tracks within a standardized black-box framework. It introduces a publicly available dataset, broad attack diversity, and an extensible open-source toolkit—together enabling reproducible, transferable, and generalizable research on adversarial robustness, with a particular emphasis on detection under practical conditions.

2.2. Defense Strategies for FR Systems

Adversarial training remains the most widely studied defense, wherein models are trained on adversarial examples

to improve robustness [20]. In FR, this has been adapted to embedding-space learning [41], though it can be computationally expensive and vulnerable to adaptive attacks.

Other defenses include input preprocessing (e.g., JPEG compression, denoising), robust loss functions [39], and ensemble methods. Yet many of these approaches degrade performance on clean images or fail under unseen attack types, highlighting the importance of thorough benchmarking under diverse and realistic conditions.

3. Methodology

The 2025 Adversarial Attack Challenge was structured into two complementary tracks designed to evaluate different aspects of adversarial robustness in face recognition systems:

- **Detection Track**: Participants developed models to accurately classify face images as clean or adversarial. The adversarial samples included both evasion attacks, aiming to prevent recognition, and impersonation attacks, aiming to falsely match another identity. This track assessed the models' ability to detect adversarial manipulations across varying perturbation levels and attack types.
- **Resilience Track**: In this track, participants trained face recognition models intended to maintain high verification performance even when subjected to adversarial attacks. The focus was on developing models that are robust to adversarial perturbations, ensuring reliable recognition in the presence of both evasion and impersonation attacks.

The adversarial attacks used in this challenge were crafted under a **black-box threat model**, wherein the attacker does not have access to the target face recognition (FR) model's architecture, parameters, or internal gradients. This setting reflects a realistic security scenario, as many deployed systems operate as closed APIs or proprietary software. In such cases, adversaries must rely solely on input-output behavior to design effective attacks.

Although most adversarial attack algorithms are initially designed for white-box settings—where full model access is assumed—numerous studies have demonstrated the phenomenon of **transferability**, whereby adversarial examples generated for one model can also fool other, unseen models [26, 10]. All adversarial examples were generated under this black-box assumption. To achieve this, we used a publicly available, open-source surrogate FR model, specifically the ArcFace model from InsightFace [5], to craft the attacks. The core premise is to test attack transferability—the ability of adversarial examples created on one model to fool other unknown models. This model was not part of the evaluation process and served solely as a stand-in attacker proxy, ensuring that the challenge remains faithful to the black-box assumption while leveraging established attack generation pipelines.

3.1. Attack Generation

To support a robust and diverse evaluation, adversarial examples were created using ten distinct attack algorithms, each selected for its methodological diversity and relevance to FR adversarial research. These attacks were applied to clean face images from CelebA and LFW datasets to produce a balanced corpus of adversarial samples for both training and development phases.

This set of attacks was chosen to span a variety of optimization strategies, perturbation norms, and black-box compatibility, thereby encouraging the development of defense methods that generalize beyond a narrow subset of attacks. Table 1 provides a detailed summary of these attacks.

3.1.1 Rationale for Attack Selection

The ten adversarial attacks were selected to cover a broad and complementary spectrum of properties relevant to face recognition robustness evaluation:

Attack Paradigms. The suite includes: *Gradient-based attacks*—IFGSM [20], MIFGSM [8], DI^{2}FGSM [35], and TI-FGSM [9]—which rely on iterative first-order optimization; *Optimization-based methods*, such as Carlini & Wagner (C&W) [2] and LBFGS [32], which frame adversarial generation as a constrained loss minimization problem; *Decision-based techniques* like DeepFool [23], which iteratively push samples across classification boundaries; and *Score-based or heuristic black-box attacks*, including Evolutionary [10] and JSMA [28], which rely on output scores or estimated saliency maps.

Perturbation Characteristics. The attacks span multiple L_p norm constraints—L_0, L_2, and L_∞—resulting in diverse perturbation magnitudes and visual artifacts. This variation ensures models are tested against a range of distortion types, from imperceptible noise to perceptually salient changes.

Transferability Focus. Several attacks (MIFGSM [8], DI^{2}FGSM [35], TI-FGSM [9]) are specifically designed to enhance cross-model transferability, which is critical in black-box scenarios where the attacker has no access to the model internals.

This combination of methodological breadth, norm diversity, and emphasis on transferability was intended to promote the development of defense systems that generalize beyond narrow threat models and encourage participants to build defense models capable of withstanding adaptive and diverse adversarial threats. In doing so, the evaluation environment simulates real-world attack scenarios and promotes

Table 1. Summary of Adversarial Attack Methods Used in the Challenge

Attack (Reference)	Description
C&W (Carlini & Wagner) [3]	Optimization-based attack minimizing perturbation under L_2, L_∞, or L_0 norms; highly effective at evading defenses.
DeepFool [22]	Iteratively finds minimal perturbation to cross a classifier's decision boundary; fast and subtle.
DI^{2}FGSM [36]	Improves transferability of gradient-based attacks by applying random resizing and padding to inputs.
Evolutionary [27]	Gradient-free black-box attack using evolutionary algorithms such as genetic or differential evolution.
IFGSM (BIM) [16]	Iterative version of FGSM with small step size and pixel clipping; effective under white-box settings.
JSMA [29]	Targeted attack that modifies salient pixels based on Jacobian-derived saliency maps.
LBFGS [3]	Early optimization-based attack using L-BFGS to find minimal perturbations that cause misclassification.
MIFGSM [7]	Extends IFGSM by adding a momentum term to gradient updates, improving stability and transferability.
PI-FGSM [15]	Applies adversarial noise in small localized patches to enhance control and reduce detection.
TI-FGSM [6]	Uses convolutional smoothing to make gradient updates robust to translation, increasing attack transferability.

the development of resilient, general-purpose FR systems.

3.1.2 Attacks Used for Evaluation

To rigorously assess the generalization capability of submitted models, the private test set was constructed using a distinct set of unseen attacks, ensuring evaluation against genuinely novel and previously unencountered adversarial threats. In line with the challenge's goal of assessing real-world resilience, the evaluation focused on attacks with high transferability or those that mimic different types of image degradation. The four attacks used are described in Table 2.

The combination of these four methods—spanning non-gradient, occlusion, and advanced gradient-based strategies—provided a comprehensive and challenging testbed for the final evaluation.

3.2. Datasets

The datasets provided to participants and used for evaluation were derived from two well-known publicly available face datasets: CelebA and LFW. A separate private test set was constructed for the final, sequestered evaluation.

- **CelebA (CelebFaces Attributes Dataset)** [19]: A large-scale face attributes dataset comprising over 200,000 celebrity images, each annotated with 40 binary attributes. For this challenge, a curated subset of CelebA was used to generate image pairs for training face recognition models and subsequently creating adversarial examples. Its diversity in terms of identities, pose, illumination, and expression makes it a valuable resource for training generalizable FR systems.
- **LFW (Labeled Faces in the Wild)** [14]: A standard benchmark dataset for unconstrained face verification, containing 13,233 images of 5,749 individuals. It is widely used for evaluating the performance of face verification algorithms. In this challenge, LFW was utilized to create genuine and imposter pairs for both clean and adversarial scenarios, particularly for the Resilience track.

After duplicate identity removal, the training data was supplied to participants included clean images sourced from these datasets, alongside their adversarially perturbed counterparts generated using the ten attack methods detailed in Section 3.1. For the Resilience track, pre-defined genuine and impostors pairs were provided. For the Detection track, images were labeled as either 'clean' or 'adversarial'.

Private Test Set. A carefully constructed private test set focused on good quality images, similar to VISA applications, was used for the definitive evaluation. This set contained 11,000 clean image pairs and 100,000 negative image pairs. From this set, 44,000 evasion attacks were performed using 4 different methods mentioned earlier and 220,000 impersonation attacks using two methods were performed. The exact distribution and parameters of the attacks in the private test set were not disclosed to participants to ensure a blind and fair evaluation of model generalization.

4. Challenge Results and Analysis

This section outlines the evaluation protocols and presents the results of the 2025 Adversarial Attack Challenge. We analyze the performance of the submitted solutions, incorporating methodological insights to better understand the factors contributing to their effectiveness. A total of ten requests were made to access the data and the top five teams were selected to publish their results in this evaluation.

Table 2. Summary of Adversarial Attack Methods Used in the Private Evaluation Set

Attack (Reference)	Description
Most Significant Bit (MSB)	Manipulates the most significant bits of pixel values, introducing visible, structured artifacts. Included to test robustness against non-gradient, direct data corruption.
Grid Occlusion	Overlays a grid of black lines, simulating structured occlusion. Evaluates resilience to patterned missing facial information.
SI-FGSM [18]	Improves gradient-based attack transferability by computing gradients over multiple scaled-down input images, enhancing robustness to scale variations.
PI-FGSM [15]	Applies adversarial noise in localized patches, testing the model's ability to handle spatially constrained and adaptable perturbations.

4.1. Evaluation Metrics

To ensure a fair and comprehensive evaluation, standardized metrics were adopted to align with international protocols for presentation attack detection (e.g., ISO/IEC 30107-3) and face recognition system performance. Separate metric suites were defined for each challenge track, as summarized in Table 3

Table 3. Evaluation Metrics for the Detection and Resilience Tracks.

Track	Metric	Description
Detection	APCER ↓	Attack Presentation Classification Error Rate
	BPCER ↓	Bona Fide Presentation Classification Error Rate
	F1-Score ↑	Harmonic mean of precision and recall
	AUC-ROC ↑	Area Under the ROC Curve
Resilience	EER (%) ↓	Equal Error Rate under adversarial attack
	ASR (%) ↓	Attack Success Rate of adversarial examples
	Robustness Score ↑	$(1 - \text{EER}) \times (1 - \text{ASR})$

4.2. Performance in the Detection Track

The Detection track required participants to classify whether input images were clean or adversarial. Final rankings were determined using a composite scoring scheme that aggregated performance across four key metrics: APCER, BPCER, F1-Score, and AUC-ROC. As shown in Table 4, each team received a rank (1 to 5) for each individual metric, and the Combined Score was computed as the sum of these ranks. Lower combined scores indicate superior overall performance.

Table 4. Final ranked performance of submissions in the Detection Track. Each cell shows the metric value and the corresponding rank (in parentheses). The table is sorted by the Combined Score, where lower is better.

Team ID	APCER ↓	BPCER ↓	F1-Score ↑	AUC-ROC ↑	Combined Score ↓
BioLab-0	0.0019 (2)	0.0347 (3)	**0.9670 (1)**	**0.9995 (1)**	**7**
BioLab-1	**0.0016 (1)**	0.1698 (5)	0.8597 (4)	0.9992 (2)	12
Team-Roma	0.0873 (4)	**0.0043 (2)**	0.9502 (2)	0.9888 (4)	12
Polish Samurai	0.0049 (3)	0.1521 (4)	0.8708 (3)	0.9967 (3)	13
SaeidUCC	1.0000 (5)	**0.0000 (1)**	0.0000 (5)	0.4661 (5)	16

Most submitted solutions demonstrated strong generalization capabilities in detecting adversarial inputs across both unseen attack types and image domains, although some performance trade-offs were observed.

Team **BioLab-0** achieved the best overall balance, securing the highest F1-Score (0.9670) and AUC-ROC (0.9995).

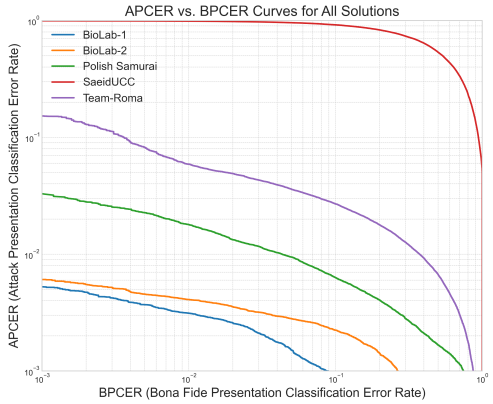

Figure 2. APCER-BPCER plot for the tested solutions, illustrating the trade-off between missing attacks and falsely flagging clean images.

Their winning solution, *PerturbationNet*, was a custom-designed lightweight CNN incorporating small kernel sizes and dilated convolutions to capture the fine-grained artifacts characteristic of adversarial perturbations. Uniquely, they formulated the problem as a multi-class classification task, training the model to recognize each of the ten attack types individually. The final binary decision (clean vs. adversarial) was derived by summing the scores assigned to the attack classes. This multi-class-to-binary strategy, combined with a carefully tuned and relatively mild data augmentation pipeline, likely contributed to the model's superior generalization and top-ranking F1 performance.

At the extremes of the trade-off, team **BioLab-1** produced the most sensitive detector with the lowest APCER (0.16%). This submission is a variant of the same PerturbationNet architecture, but with a different number of training epochs and different operation threshold to account for 32-bit inference instead of 16-bit of the first submission.

Conversely, team **Team-Roma** submitted a highly specific model with the lowest BPCER (0.43%). Their solution used a ResNet18 [12] backbone but augmented the input with multiple pre-processed versions of the image, including different JPEG compression levels, grayscaling, and morphological operations, named

Resnet18MoreThanRGB. This multi-feature input likely helped the model build a more robust representation of what constitutes a 'clean' image, thereby minimizing false alarms and ensuring a better user experience.

A deeper, per-attack analysis of the AUC scores provides crucial insight into the specific strengths and weaknesses of the detection models. The most striking result is the universal difficulty models faced with the **Grid Occlusion** attack. While most gradient-based and bit-manipulation attacks were almost perfectly detected (AUC approaching 1.0) by the top models, the structured occlusion proved to be a significant challenge. This single attack type is largely responsible for the performance differences observed in Table 4.

Team-Roma's results offer a compelling case study. Their model's overall AUC was 0.9888, but its AUC against Grid Occlusion was only 0.7730. This specific vulnerability explains why their APCER was higher than other top models; their system was less capable of identifying this particular unseen attack. Conversely, the BioLab and Polish Samurai models demonstrated fundamentally stronger discriminative power against Grid Occlusion ($AUC > 0.93$), even if their chosen decision thresholds led to higher BPCER values. This analysis reinforces two key insights: first, that even high-performing detectors can exhibit architectural blind spots when faced with adversarial patterns outside their training distribution; and second, that structured occlusion remains a potent and underappreciated adversarial strategy—particularly when omitted from training data.

4.3. Performance in the Resilience Track

In the Resilience track, the goal was to develop FR models that maintained high verification accuracy under attack. The final ranking was determined by the Robustness Score, which combines EER and ASR.

As shown in Table 6, team **Team-Roma** was the clear winner, achieving the highest Robustness Score (0.5123). Their solution employed a FaceNet model [30] with an InceptionResNet backbone, pretrained on the large-scale CASIA-WebFace dataset. Crucially, they used Gaussian blurring as a preprocessing defense to mitigate adversarial noise. This classic strategy proved highly effective, combining a powerful feature extractor with a simple yet potent defense that disrupts gradient-based perturbations.

The submissions from **Gradient Ascent** took a more complex, two-stage approach. They first used a transformer-based image restoration model [37] to denoise the input image. The restored image was then passed to an IR-SE50 recognition model (a ResNet50 with Squeeze-and-Excitation blocks) [13] for embedding. While sophisticated, this pipeline was less effective than Team-Roma's simpler method, suggesting that the denoising process may not have fully removed the adversarial structures targeted in the evaluation.

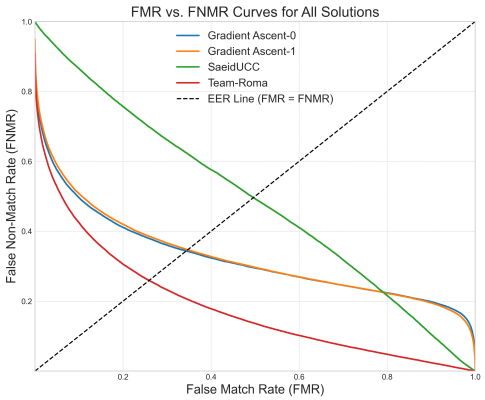

Figure 3. FMR-FNMR plot for the tested solutions under adversarial conditions. The ideal system operates closer to the origin.

Interestingly, the submission from **SaeidUCC**, which utilized a Siamese network with PGD-based adversarial training [20], struggled in comparison. While adversarial training is a theoretically strong defense, this result highlights that its success is highly dependent on implementation details and its alignment with the specific attacks encountered during evaluation.

4.3.1 Per-Attack Resilience Analysis

To understand the factors driving overall performance, we conducted a detailed analysis of each model's resilience against the specific attack types. The results, presented in Table 5, reveal critical insights into the current state of adversarial defenses for face recognition.

The breakdown reveals several key findings. First, the foundational strength of **Team-Roma's** solution is evident in its exceptionally low Bona Fide EER of just 4.11%. This superior performance on clean data provided a significant advantage, as the model started from a much more accurate baseline before any attacks were introduced.

Second, the results expose a universal vulnerability to the non-gradient-based **Grid Occlusion** and **Most Significant Bit (MSB)** attacks. This indicates that current defense strategies are largely ineffective against these structured, non-gradient perturbations.

Third, all models demonstrated significantly better resilience against the gradient-based FGSM variants. Team-Roma's simple Gaussian blur defense was remarkably effective against the **SI-FGSM** impersonation attack, achieving a near-perfect Robustness Score of 89.69% by keeping both ASR and EER exceptionally low. This suggests that while classic defenses can handle certain transferable gradient attacks, they are not all-encompassing.

Table 5. Detailed Resilience Track performance, broken down by attack type. Each cell shows Attack Success Rate / Equal Error Rate / Robustness Score, all in percent (%). The best-performing cell in each row, determined by the highest Robustness Score (RS), is highlighted in **bold**.

Attack Scenario	Team-Roma ASR ↓ — EER ↓ — RS ↑ (%)	Gradient Ascent-0 ASR ↓ — EER ↓ — RS ↑ (%)	Gradient Ascent-1 ASR ↓ — EER ↓ — RS ↑ (%)	SaeidUCC ASR ↓ — EER ↓ — RS ↑ (%)
Bona Fide	N/A / **4.11** / N/A	N/A / 12.90 / N/A	N/A / 13.11 / N/A	N/A / 39.24 / N/A
Evasion Attacks				
Grid Occlusion	71.09 / 24.17 / 21.92	73.15 / 30.09 / 18.77	72.95 / 30.35 / 18.84	**53.83 / 42.82 / 26.40**
Most Significant Bit	81.75 / 22.89 / 14.07	99.95 / 50.65 / 0.02	99.95 / 50.63 / 0.03	**94.98 / 55.49 / 2.23**
PI-FGSM	40.63 / 11.68 / 52.44	**27.57 / 16.65 / 60.37**	30.52 / 17.51 / 57.32	41.15 / 39.74 / 35.46
SI-FGSM	**14.00 / 6.46 / 80.45**	20.81 / 15.03 / 67.29	21.90 / 15.43 / 66.05	41.30 / 39.78 / 35.35
Impersonation Attacks				
PI-FGSM	**19.84 / 7.09 / 74.47**	13.87 / 13.04 / 74.89	14.07 / 13.36 / 74.45	46.82 / 41.09 / 31.33
SI-FGSM	**6.10 / 4.48 / 89.69**	10.86 / 12.45 / 78.04	11.18 / 12.65 / 77.58	47.08 / 41.15 / 31.15

Table 6. Performance of submissions in the Resilience Track. Best performance for each metric is in **bold**.

Team ID	EER (%) ↓	ASR (%) ↓	Robustness Score ↑
Team-Roma	**25.93**	**30.84**	**0.5123**
Gradient Ascent-0	34.35	35.44	0.4239
Gradient Ascent-1	34.71	35.78	0.4193
SaeidUCC	49.62	52.50	0.2393

Table 7. Top-performers in the Detection and Resilience tracks.

Rank	Team Name	Participants	Achievement
Detection Track			
1st & 2nd	BioLab	Lorenzo Pellegrini, Nicolò Di Domenico, Guido Borghi	Combined Scores: 7 & 12
3rd	Team-Roma	Niklas Bunzel, Lukas Graner, Nicholas Göller	Combined Score: 12
Resilience Track			
1st	Team-Roma	Niklas Bunzel, Lukas Graner, Nicholas Göller	Robustness Score: 0.5123
2nd & 3rd	Gradient Ascent	Monson Verghese, Shruti Bhilare, Avik Hati	Robustness Scores: 0.4239 & 0.4193

Finally, the results consistently indicate that **evasion attacks were more damaging than impersonation attacks**. Across all teams, Robustness Scores were consistently higher for impersonation scenarios, confirming that it is generally easier for an attacker to disrupt a legitimate match than to impersonate a specific identity. This asymmetry has important implications for the design of future defense strategies and the prioritization of threat models.

The top-performing teams in both the Detection and Resilience tracks are highlighted for their noteworthy achievements. In the Detection track, BioLab (Pellegrini *et al.*) and Team-Roma (Bunzel *et al.*) delivered strong results in identifying adversarial threats, demonstrating effective detection of both known and novel perturbations. In the Resilience track, Team-Roma led the field, with Gradient Ascent (Verghese *et al.*) also presenting promising approaches to robustness—despite the inherent challenges posed by non-differentiable attacks.

5. Conclusion

The 2025 Adversarial Attack Challenge for Secure Face Recognition successfully established a unified benchmark for evaluating system robustness, providing a public dataset and open-source tools to foster reproducible research. The results from both the Detection and Resilience tracks revealed a significant gap between performance on known threats and generalization to novel attacks. In detection, while models effectively identified familiar perturbations, they struggled with unseen, structured attacks like Grid Occlusion. Similarly, the Resilience track showed that all systems were highly vulnerable to non-gradient perturbations, even as some classic defenses proved effective against specific gradient-based attacks, highlighting that current strategies are not yet universally robust.

The challenge's primary contributions—the public benchmark dataset and open-source attack suite—are intended to accelerate research in this critical area. Our findings point to two clear priorities for future work: expanding adversarial training to address diverse, non-differentiable attacks, and developing hybrid systems that integrate robust detection with resilient recognition. These advancements are critical for building secure and trustworthy face recognition systems ready for real-world deployment.

6. Acknowledgments

Youverse wants to acknowledge the Portuguese Recovery and Resilience Plan (RPP) who partially funded this project under the program 'Agendas para a Inovação Empresarial', reference no. 62 - 'Center for Responsible AI'. Additionally, this project was also partially financed by the Portuguese Fundação para a Ciência e Tecnologia under the contract UID/00048 - Instituto de Sistemas e Robótica - Coimbra (ISR-UC)

References

[1] A. Athalye, N. Carlini, and D. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pages 274–283. PMLR, 2018.

[2] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *IEEE SP*, pages 39–57, 2017.

[3] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. *IEEE Symposium on Security and Privacy (SP)*, 2017.

[4] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *ACM Workshop on Artificial Intelligence and Security*, 2017.

[5] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019.

[6] Y. Dong, Q. Fu, X. Yang, T. Pang, H. Su, J. Zhu, and J. Bao. Improving transferability of adversarial examples with translation-invariant attacks. In *CVPR*, 2019.

[7] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li. Boosting adversarial attacks with momentum. In *CVPR*, 2018.

[8] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9185–9193, 2018.

[9] Y. Dong, T. Pang, H. Su, and J. Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4312–4321, 2019.

[10] Y. Dong, H. Su, J. Zhu, and J. Bao. Efficient decision-based black-box adversarial attacks on face recognition. In *CVPR*, pages 7714–7722, 2019.

[11] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015.

[13] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[14] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on Faces in 'Real-Life' Images (ICCV)*, 2008.

[15] L. Huang, S. Gong, W. Liu, and J. Yang. Pi-fgsm: Patch-wise iterative fast gradient sign method for targeted adversarial attacks. In *ACM Conference on Multimedia (MM)*, 2020.

[16] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. In *Workshop on Artificial Intelligence and Security (AISec), ICLR*, 2017.

[17] R. Li, J. Deng, L. Xie, and S. Zafeiriou. Exploring adversarial example detection in deep face recognition models. *Pattern Recognition*, 113, 2021.

[18] J. Lin, C. Song, K. He, L. Liu, and J. E. Hopcroft. Nesterov accelerated gradient and scale-invariance for adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 593–602, 2019.

[19] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[20] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

[21] J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff. Detecting adversarial perturbations with pattern recognition. In *International Conference on Learning Representations (ICLR)*, 2017.

[22] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *CVPR*, 2016.

[23] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2574–2582, 2016.

[24] NIST. Face recognition vendor test (frvt). `https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt`, 2023. Accessed: 2025-06-17.

[25] T. Pang, Y. Du, J. Zhu, and C. Chen. Towards robust detection of adversarial examples. In *NeurIPS*, 2018.

[26] N. Papernot, P. McDaniel, and I. Goodfellow. Practical black-box attacks against machine learning. In *ACM Asia CCS*, 2017.

[27] N. Papernot, P. McDaniel, and I. Goodfellow. Practical black-box attacks against machine learning. *Proceedings of the ACM on Asia Conference on Computer and Communications Security (AsiaCCS)*, 2017.

[28] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami. The limitations of deep learning in adversarial settings. In *IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 372–387. IEEE, 2016.

[29] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. The limitations of deep learning in adversarial settings. In *IEEE European Symposium on Security and Privacy*, 2016.

[30] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015.

[31] 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 *ACM CCS*, pages 1528–1540, 2016.

[32] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

[33] L. Tong, Z. Chen, J. Ni, W. Cheng, D. Song, H. Chen, and Y. Vorobeychik. FACESEC: A fine-grained robustness evaluation framework for face recognition systems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13254–13263, 2021.

[34] F. Tramèr, N. Carlini, W. Brendel, and A. Madry. Adaptive attacks to compare the robustness of neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 16336–16347, 2020.

[35] C. Xie, Z. Wang, Z. Zhang, Y. Ren, and A. L. Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2730–2739, 2019.

[36] C. Xie, Z. Zhang, Y. Wang, Z. Wei, and S. Lin. Improving transferability of adversarial examples with input diversity. In *CVPR*, 2019.

[37] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022.

[38] K. Zhang, W. Liu, J. Deng, and S. Zafeiriou. Robfr: Benchmarking robustness of face recognition models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[39] K. Zhang, W. Liu, and S. Zafeiriou. Adversarial defense through network profiling based path regularization. In *ECCV*, 2020.

[40] R. Zhang, S. Li, and S. Han. Aiot-face: Benchmarking adversarial robustness of face recognition for edge ai. In *NeurIPS*, 2022.

[41] S. Zheng, J. Deng, L. Xie, and S. Zafeiriou. Boosting the adversarial robustness of face recognition via noise-to-semantic adversarial training. In *CVPR*, 2022.

[42] Y. Zhong and W. Deng. Towards transferable adversarial attack against deep face recognition. *IEEE Transactions on Information Forensics and Security*, 2020.