1st AADD Challenge: Adversarial Attacks on Deepfake Detectors: A Challenge in the Era of AI-Generated Media – ACM Multimedia 2025

News and info

- 2025/06/11 Evaluation script released (Download)
- 2025/05/15 Extended deadline at May 22
- 2025/04/10 We started to release the data to the participants.
- 2025/03/15 Registration opened!
- 2025/03/04 Deadlines updated!
- 2025/03/04 Information for Authors are available here
- 2025/03/03 WEBSITE online!

Brief Description

The goal of this challenge is to investigate adversarial vulnerabilities of deepfake detection models by generating adversarial perturbed deepfake images that evade state-of-the-art classifiers while maintaining high visual similarity to the original deepfake content. Given the increasing reliance on deepfake detectors in forensic analysis and content moderation, ensuring their robustness against adversarial attacks is of paramount importance. Over the next 3-5 years, this challenge will contribute to the development of more resilient deepfake detection methodologies, mitigating risks associated with adversarial manipulations.

The aim of this challenge is to expose and address vulnerabilities in current deepfake detection systems by designing adversarial attacks that alter deepfake images—rendering them unrecognizable as synthetic content to 4 proposed classifiers—while preserving high visual similarity to the original images. Participants will be provided with one dataset, divided into sixteen subsets: four GAN-based models and four diffusion-based models in high quality resolution, and four GAN-based models and four diffusion-based models in low quality resolution. Participants must focus on the entire dataset. In addition to developing effective adversarial perturbations, participants are required to submit an abstract that outlines their methodology and attack the adversarial images.

Registration Process

To register for the challenge, please send an email to the main contact, Mirko Casu (challenge.dfl@gmail.com), providing the following information:

- Name of Team
- List of teammates (Name, Surname, Nationality, Organization)
- Team email
- Team referent (a teammate)
- Organization/Institution

Participation in the AADD-2025 challenge offers an opportunity to contribute to cutting-edge research in Adversarial Attacks on Deepfake Detectors while showcasing your expertise and innovative solutions.

Chairs

Luca Guarnera, Research Fellow, luca.guarnera@unict.it (Department of Mathematics and Computer Science, University of Catania, Italy)

Francesco Guarnera, Research Fellow, francesco.guarnera@unict.it (Department of Mathematics and Computer Science, University of Catania, Italy)

Co-Chairs

Sebastiano Battiato, Full Professor, sebastiano.battiato@unict.it (Department of Mathematics and Computer Science, University of Catania, Italy)

Giovanni Puglisi, Associate Professor, puglisi@unica.it (Department of Mathematics and Computer Science, University of Cagliari, Italy)

Zahid Akhtar, Associate Professor, akhtarz@sunypoly.edu (State University of New York Polytechnic Institute, USA)

Technical Commitee

Mirko Casu, PhD Student, , mirko.casu@phd.unict.it (Department of Mathematics and Computer Science, University of Catania, Italy)

Orazio Pontorno, PhD Student, orazio.pontorno@phd.unict.it (Department of Mathematics and Computer Science, University of Catania, Italy)

Claudio Vittorio Ragaglia, PhD Student, claudio.ragaglia@phd.unict.it (Department of Mathematics and Computer Science, University of Catania, Italy)

Main Contact

Name : Mirko Casu

Email : challenge.dfl@gmail.com

Address : Dipartimento di Matematica e Informatica Cittadella Universitaria - Viale A. Doria 6 - Italy

Important dates

Website online [March 03]

Registration [March 15 - May 22] - Extended

To register for the challenge, please send an email to the main contact, Mirko Casu (challenge.dfl@gmail.com), providing the following information:

- Name of Team
- List of teammates (Name, Surname, Nationality, Organization)
- Team email
- Team referent (a teammate)
- Organization/Institution

Participation in the AADD-2025 challenge offers an opportunity to contribute to cutting-edge research in Adversarial Attacks on Deepfake Detectors while showcasing your expertise and innovative solutions.

List of registered teams (until now)

	Team Name	Organization/Institution	Country
1	KR-NSI	ETRI	South Korea
2	WHU_PB	Wuhan University	China
3	DASH	Sunghyunkwan University	South Korea
4	MICV	Ant Group	China
5	MBZUAI	Mohamed bin Zayed University of Artificial Intelligence	Abu Dhabi
6	MuhuLab	National University of Defense Technology	China
7	GRADIANT	Gradiant	Spain
8	RoMa	Fraunhofer SIT	ATHENE Center	Germany
9	VYAKRITI 2.0	Apex Institute of technology Chandigarh University	India
10	Safe AI	UNIST (Ulsan National Institute of Science and Technology)	South Korea
11	MiLab	University of Science and Technology of China	China
12	FalseNegative	The Hong Kong Polytechnic University	China
13	DeFakePol	Samsung Research Poland	Poland
14	MI-CAS	University of Chinese Academy of Sciences	China
15	Robust	IIT Jammu	India
16	The Adversaries	Singapore Institute of Technology	Singapore
17	SecureML	University of Cagliari	Italy

Test set and classificator release [April 10]

Test Set Description:

High Quality (HQ) generators: Adobe Firefly, DeepAI, Flux 1.1 Pro, HotPotAI, NvidiaSanaPAG, StableDiffusion 3.5, StyleGAN 2, StyleGAN 3, Tencent Hunyuan
Low Quality (LQ) generators: DeepAI, Flux.1, Freepik, HotPotAI, NvidiaSanaPAG, Stable Diffusion Attend and Excite, StyleGAN, StyleGAN 3, Tencent Hunyuan

Although the challenge involves a total of four classifiers, only two of them (ResNet and DenseNet) are released with the dataset. The remaining two are used as blind models by the organizers and are not made available to participants. So, these blind classifiers are used only in the evaluation phase.

Notes:
LQ images are created by resizing the images followed by variable Quality Factor (QF) compression. This combination is designed to simulate social media compression. Note that the number of images generated by GAN and Diffusion Models is numerically balanced.

Submission results [June 15]

By this date, all participants are required to submit the following:

-Attacked Test Set: A version of the provided test set that reflects the participant's attack strategy, following the challenge guidelines.
-Abstract Paper: A short abstract paper (1-2 pages) that briefly describes the methodology, motivation, and key contributions of the proposed approach.
Submissions must be made sent an email to the challenge referent Mirko Casu (challenge.dfl@gmail.com).

Leaderboard Publication [June 22]

By this date, the official challenge leaderboard will be released based on the evaluation of the submitted attacked test sets. The top 3 teams will be notified via email.

Please note: All accepted papers will be considered for publication in the ACM Multimedia 2025 proceedings. For more information, please visit the official website.

	Team Name	Organization/Institution	FINAL SCORE
1	MI-CAS	University of Chinese Academy of Sciences	2740
2	Safe AI	UNIST (Ulsan National Institute of Science and Technology)	2709
3	RoMa	Fraunhofer SIT	ATHENE Center	2679
4	GRADIANT	Gradiant	2631
5	DASH	Sunghyunkwan University	2618
6	SecureML	University of Cagliari	2490
7	MICV	Ant Group	2434
8	WHU_PB	Wuhan University	2354
9	The Adversaries	Singapore Institute of Technology	2341
10	DeFakePol	Samsung Research Poland	1665
11	False Negative	The Hong Kong Polytechnic University	1602
12	VYAKRITI 2.0	Apex Institute of technology Chandigarh University	1041
13	MiLab	University of Science and Technology of China	110

Final Paper Submission (Top 3 Teams Only) [June 30]

The top 3 teams will be invited to submit a full-length paper describing their method in detail. This paper will undergo a review process managed by the challenge organizers.

Announcement regarding full paper submission [July 24]

The organizers will notify the top 3 teams about the outcome of the review process. Based on the reviews, zero, one, more, or all of the submitted papers may be accepted for inclusion in the ACM Multimedia 2025 proceedings.

Camera ready - Grand Challenge Solutions (Top 3 Teams Only) [August 01]

Evaluation criteria

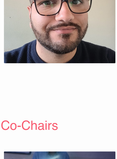
- **SSIM Requirement & Submission**
  Each submission must include original deepfake images and their perturbed versions. Only complete image pairs will be evaluated.
- **Accuracy Calculation**
  An attacked image is a positive case if a detection system misclassifies it as "real." Accuracy is the proportion of these cases within the dataset.
- **Final Score Composition**
  The score is a weighted average of SSIM and detection accuracy (across four classifiers). Higher SSIM indicates greater similarity. Weight distribution will be disclosed upon acceptance. The final score is calculates as follow:

$$\sum_{C_f \in C} \sum_{k=1}^{N} SSIM(I_k, I_k^{ADV}) \cdot [C_f(I_k^{ADV}) = LABEL_{real}]$$

where:

- C_f is the set of all classifiers
- N is the number of deepfake images in the test dataset
- I_k is the k-th image from the deepfake test dataset
- I_k^{ADV} is the adversarial image generated from I_k
- $LABEL_{real}$ is the label of class real
- [] is the indicator function which equals to 1 when predicate is true, otherwise equals to 0

Chairs



Co-Chairs

Dataset Committee