Title: **2nd Adversarial Attacks on Deepfake Detection (AAAD Challenge 2026): Generated Media in Real-World Scenarios**

Organizers**: Luca Guarnera (*), Francesco Guarnera (*), Sebastiano Battiato (*), Giovanni Puglisi (**), Zahid Akhtar (^), Andrea Montibeller (^^), Giulia Boato (^^).**

(*): University of Catania, Italy

(**): University of Cagliari, Italy

(^): State University of New York Polytechnic Institute, USA

(^^): University of Trento, Italy

Brief Description

Deepfake generation technologies are rapidly evolving and are increasingly used to manipulate public opinion, attack individuals, and amplify large scale misinformation. In the next few years, generators will make synthetic faces almost indistinguishable from real ones, especially after being shared on social media platforms that apply compression, resizing, re-encoding and filtering. These widespread transformations expose a gap between lab evaluations and real-world use, underscoring the urgent need for robust deepfake detectors that remain reliable under both adversarial attacks and realistic post-processing. Our AADD Challenge 2025 at ACM MM (https://dl.acm.org/doi/10.1145/3746027.3761983) has shown that state-of-the-art deepfake detectors are still vulnerable to carefully crafted adversarial perturbations under controlled conditions. The AADD Challenge 2026 extends this line of work by explicitly modeling JPEG compression and social-media-like processing, making the evaluation significantly closer to real deployment scenarios. This is directly relevant to the multimedia research community, industry (social media, streaming, content platforms), and society, as it targets the robustness of forensic tools that will be critical in safeguarding information integrity over the coming years. By advancing robust deepfake forensics under adversarial and real-world constraints, AADD 2026 supports the development of dependable tools for protecting information integrity and digital trust.

Research tasks

Generate adversarial perturbations that fool deepfake detectors while maintaining high visual fidelity

The scientific community must develop adversarial perturbation techniques that simultaneously fool deepfake while maintaining strict visual fidelity ensuring cross-generator transferability across synthesis models, while preserving high visual fidelity and perceptual realism and preserving effectiveness post-JPEG compression. This long-term research addresses critical gaps in detector robustness against evolving generation technologies, platform-induced degradation, and coordinated attack scenarios, ultimately enabling reliable forensic systems for social media moderation, law enforcement investigations, and democratic process protection over the next years.

Develop attacks that survive JPEG compression

This task involves developing adversarial attack methodologies that survive realistic JPEG compression pipelines simulating Social Network processing, ensuring perturbations remain effective post-compression while maintaining visual fidelity across dual evaluation scenarios. This addresses the long-term challenge of building deployment-ready deepfake detectors that withstand both adversarial manipulation and platform-induced degradation, enabling robust content moderation systems, reliable forensic evidence preservation, and trustworthy automated filtering at web scale over the next years.

Create attacks that generalize across generation models and unseen detection architectures

Investigate adversarial attack strategies that generalize across multiple deepfake generation frameworks and remain effective against unseen and black-box detection architectures. This addresses universally robust detection systems against evolving generation technologies and detector architectures, ensuring sustained forensic reliability across diverse threat landscapes.

Objective of the Challenge

The 2025 AADD edition attracted 17 international teams from leading universities and industry, generating significant peer-reviewed contributions. The 2026 edition, with enhanced realism and the integrated post-processing scenario, is expected to attract even larger participation, pooling diverse methodological approaches, and accelerating innovation cycles. AADD 2026 aims to accelerate innovation cycles in robust multimedia forensics and adversarial learning.

Current state-of-the-art deepfake detectors achieve impressive accuracy on clean benchmark datasets but reveal critical vulnerabilities when confronted with adversarial perturbations combined with realistic JPEG compression typical of social media platforms, while no standardized benchmark exists that systematically integrates next generation deepfake generators with authentic platform processing pipelines. This Challenge accelerates research by establishing a comprehensive evaluation framework that combines standardized multi-metric assessment (SSIM+LPIPS+evasion accuracy) with realistic threat modeling incorporating actual social media compression. By providing public datasets, blind evaluation protocols, live leaderboards, and integration with ACM Multimedia proceedings, AADD 2026 creates a sustained research infrastructure that enables fair method comparison, fosters healthy competition among international teams, bridges the gap between academic adversarial research and industrial deployment requirements, and driving breakthrough innovations in robust multimedia forensics that can withstand coordinated adversarial attacks in real-world operational environments.

Challenge rules:

Dataset Composition and Accessibility

The AADD 2026 challenge dataset has been meticulously curated to reflect the cutting-edge landscape of synthetic facial image generation, featuring a comprehensive collection of deepfake images produced through contemporary state-of-the-art techniques, with a dedicated training subset reserved exclusively for organizer use in training the challenge detectors and a distinct test subset provided directly to participants for adversarial perturbation. Participants will receive full access

solely to the test set, comprising around 2,000 deepfake images, enabling them to craft and submit attacks tailored to evade detection, while undergoing rigorous, organizer-controlled evaluation to ensure fairness and prevent any premature leakage of assessment insights.

Dataset Access and Documentation

All datasets will be provided privately and securely to registered participants through dedicated channels, ensuring controlled and confidential access that upholds ethical standards and prevents unauthorized use. To foster long-term academic impact and enable future comparisons, the datasets (including the test set and evaluation detectors) along with complete documentation will be made available upon request to the organizers only at the conclusion of the challenge, after the conference, allowing external researchers to replicate and extend the results with proper attribution. This delayed release enables independent replication and extension of results while preserving the integrity of the live competition.

Submission Format and Requirements

Every participating team is required to deliver a structured submission package that includes adversarial versions of the entire test set, meticulously organized within a predefined directory structure that mirrors the original filenames to enable seamless processing and evaluation. Complementing these perturbed images is a concise yet informative 1–2 page methodology abstract, which must articulate the core attack approach, pivotal design decisions, and standout innovations that distinguish the submission. Additionally, participants agree to a mandatory embargo period extending until after the ACM Multimedia conference, prohibiting the publication of any papers referencing the provided challenge dataset during this time to safeguard evaluation integrity and competition fairness.

Objective Evaluation Metrics

Submissions undergo evaluation via three synergistic metrics that collectively balance imperceptibility, perceptual fidelity, and attack efficacy, starting with the Structural Similarity Index (SSIM) to quantify pixel-level resemblance between pristine deepfakes and their adversarial counterparts. Building on this foundation, the Learned Perceptual Image Patch Similarity (LPIPS) metric employs pre-trained deep neural networks to gauge human-like perceptual quality, capturing subtle changes that might evade traditional pixel metrics but could still betray the attack to discerning observers. The cornerstone metric, Detection Evasion Accuracy, calculates the proportion of adversarial images successfully misclassified as authentic across the evaluation classifiers, directly quantifying the potency of the attack in circumventing detection mechanisms.

The ultimate leaderboard ranking emerges from a weighted aggregation of the three core metrics—SSIM for structural fidelity, LPIPS for perceptual alignment, and Evasion Accuracy for outright success—wherein each component is appropriately normalized to a common scale before integration, with weights strategically assigned to prioritize detection evasion as the paramount objective while

imposing firm guardrails on visual distortions through the imperceptibility measures. This balanced formulation ensures that top-performing attacks not only fool detectors with high reliability but also preserve the intrinsic qualities of the original deepfakes, thereby advancing the state-of-the-art in adversarial robustness studies without sacrificing practical realism.

Evaluation

To mirror real-world deployment challenges, all evaluations incorporate post-processing realism through organizer-applied JPEG compression across varying quality factors, rigorously testing the resilience of adversarial perturbations under conditions that simulate common image transmission and storage pipelines. A dedicated robustness score further highlights submissions that sustain elevated evasion rates post-compression, rewarding methods engineered for durability and conferring ranking advantages to those demonstrating superior generalization beyond pristine scenarios.

Ranking and Winner Selection

Teams ascend the leaderboard in descending order of their composite scores, with the uppermost three earners securing invitations to author extended technical papers (comprising six pages plus two for references) for rigorous peer review, potentially culminating in publication within the prestigious ACM Multimedia 2026 proceedings upon acceptance by an expert panel. Ties are adjudicated through a hierarchical cascade of criteria: first, the highest average SSIM; second, the most favorable average LPIPS; third, strongest robustness under post-processing; fourth, earliest submission timestamp; and finally, a qualitative judgment of methodological ingenuity by the organizers.

Information about Challenge

Building on the successful infrastructure established for AADD 2025 the organizing team commits to extending this proven framework within the same multimedia forensics domain for AADD 2026 by establishing and maintaining a dedicated website that provides continuous access to enhanced datasets for participants or applicants. We will collaborate closely with ACM Multimedia 2026 organizers to publicize the challenge through official channels, targeted outreach to forensics/AI security communities, and dedicated conference sessions presenting top results, while maintaining the rigorous review and evaluation process proven effective in 2025 with objective multi-metric assessment across blind classifiers, ensuring fair competition, scientific transparency, and sustained community engagement across challenge editions.

For information contact:

- Luca Guarnera: luca.guarnera@unict.it
- Francesco Guarnera: francesco.guarnera@unict.it