Adversarial Attacks on Deepfake Detectors: A Challenge in the Era of AI-Generated Media (AADD-2025)

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

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

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

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

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

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

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

Zahid Akhtar
Department of Electrical and Computer Engineering,
State University of New York Polytechnic Institute
Utica, New York, USA
akhtarz@sunypoly.edu

Abstract

The proliferation of AI-generated media has heightened risks of misinformation, driving the need for robust deepfake detection systems. However, adversarial attacks—subtle perturbations designed to evade detection—remain a critical vulnerability. To address this, we organized the AADD-2025 challenge, inviting participants to develop attacks that fool diverse classifiers (e.g., ResNet, DenseNet, blind models) while preserving visual fidelity. The dataset included 16 subsets of high/low-quality deepfakes generated by GANs and diffusion models (e.g., StableDiffusion, StyleGAN3). Teams were evaluated on structural similarity (SSIM) and attack success rates across classifiers. Thirteen teams proposed innovative solutions leveraging latent-space manipulation, ensemble gradients, surrogate modeling, and frequency-domain perturbations. Challenge's top performers—MR-CAS (1st, score: 2740), Safe AI (2nd, 2709), and RoMa (3rd, 2679)—achieved high SSIM (0.74–0.93) while evading classifiers. MR-CAS's latent diffusion inversion and Safe AI's gradient ensemble framework demonstrated superior transferability, even against Vision Transformers. Key insights revealed latent-space attacks outperform pixel-level methods, ensemble strategies enhance cross-model robustness, and hybrid CNN-transformer attacks are most effective. Despite progress, challenges persist in generalizing attacks across heterogeneous models and maintaining perceptual quality. The AADD-2025 challenge underscores the urgency of developing adaptive defenses and hybrid detection systems to counter evolving adversarial threats in AI-generated media. To facilitate reproducibility and further research, the complete dataset is available for download in the challenge GitHub repository https://github.com/mfs-iplab/aadd-2025.

CCS Concepts

• **Applied computing** → **Computer forensics**.

Keywords

Adversarial Attacks; Deepfake Detection; AI-Generated Media; Latent-Space Manipulation; Transferability; Ensemble Methods; Generative Models; Diffusion Models; Vision Transformers; Digital Forensics; Perceptual Quality

ACM Reference Format:
Sebastiano Battiato, Mirko Casu, Francesco Guarnera, Luca Guarnera, Giovanni Puglisi, Orazio Pontorno, Claudio Vittorio Ragaglia, and Zahid Akhtar. 2025. Adversarial Attacks on Deepfake Detectors: A Challenge in the Era of AI-Generated Media (AADD-2025). In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3746027.3761983

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3761983

1 Introduction

The rapid evolution of generative AI technologies, especially Generative Adversarial Networks (GANs) [8] and diffusion models [11], has greatly enhanced the realism of synthetic media, commonly known as deepfakes [2, 14, 19]. While these models enable applications in entertainment and virtual human creation, they also pose serious risks including misinformation, identity fraud, and erosion of public trust—often manifesting as an 'impostor bias,' where users doubt the authenticity of media content [4, 14]. Consequently, robust deepfake detection has become essential for digital forensics and content moderation [9, 17]. Earlier research also examined manipulations introduced by social media platforms such as Facebook [16]. Despite advances using deep learning architectures such as CNNs and Vision Transformers (ViT), detectors remain highly vulnerable to adversarial attacks—subtle perturbations that mislead models into classifying deepfakes as authentic [10, 18]. Studies show that both white-box and black-box attacks can effectively bypass state-of-the-art classifiers, exposing critical weaknesses [1, 5]. A key challenge lies in achieving strong transferability across models while preserving imperceptibility of perturbations [7, 24]. The Adversarial Attacks on Deepfake Detectors (AADD-2025) challenge was designed to address these issues, requiring participants to craft adversarial examples that evade multiple detectors while maintaining high visual quality. It encouraged innovative methods—latent-space manipulation, ensemble gradients, and surrogate modeling—to advance resilience in detection systems. The competition leveraged a comprehensive dataset of high- and low-quality deepfakes generated by GANs and diffusion models [1], with evaluation based on Structural Similarity Index Measure (SSIM) and attack success rates, promoting balanced optimization between visual fidelity and adversarial effectiveness. The remainder of this paper is organized as follows. Section 2 reviews related work on deepfake detection and adversarial robustness. Section 3 introduces the AADD-2025 challenge, including dataset, protocol, and evaluation metrics. Section 4 summarizes the approaches proposed by participating teams, while Section 5 presents the main results and insights. Finally, Section 6 concludes with final remarks and directions for future research.

2 Related Work

The domain of deepfake detection has become a dynamic field of study, with a primary focus on developing classifiers that can identify sophisticated forgeries [1]. However, the adversarial robustness of these detectors is a significant concern, as numerous studies have demonstrated their vulnerability to carefully crafted perturbations [1, 15, 18, 21, 22]. Research has shown that by introducing small, often imperceptible changes to a deepfake, an attacker can cause state-of-the-art detection models, including those based on CNNs like XCeption and ResNet, to misclassify the content as authentic [1, 23]. This vulnerability persists across various attack scenarios, from white-box attacks, where the attacker has full knowledge of the model, to more practical black-box settings where the model's architecture and parameters are unknown [18, 20]. The effectiveness of such attacks is often linked to their transferability, where perturbations created for one model can successfully fool another [5, 18]. Recent surveys and comprehensive evaluations consistently highlight that even top-performing detectors show a significant drop in performance under adversarial conditions, underscoring the urgent need for more resilient detection systems [1, 6].

A critical aspect of generating effective adversarial examples is the trade-off between the attack's success and the preservation of visual quality. The goal for an attacker is to create perturbations that are strong enough to fool a detector but subtle enough to remain invisible to human observers [24]. To this end, recent attack methodologies have moved beyond simple additive noise. For instance, some methods use generative models to create adversarial perturbations that are more structured, like shadows or subtle lighting changes, to better conceal artifacts [7]. Others employ techniques to constrain the magnitude of the perturbations in the perceptual domain, ensuring high-fidelity outputs [24, 25]. The challenge of maintaining this balance is central to modern adversarial attack research and is a key evaluation criterion in competitions like the AADD-2025 challenge. The development of attacks that can evade an ensemble of detectors, including unknown or "blind" models, while maintaining high structural similarity to the original deepfake, represents the current frontier in this arms race, pushing the research community to develop more fundamentally robust detection paradigms.

3 Challenge Description

Participants were tasked with designing adversarial attacks targeting four classifiers: a ResNet50, a DenseNet121, and two previously undisclosed blind models (i.e., models not initially released to participants and used exclusively during evaluation)—a ViT-B-16 and a DenseNet121. Notably, the DenseNet121 blind model differs from the other classifiers by leveraging Discrete Cosine Transform (DCT) features instead of spatial features. These classifiers were trained across diverse generative models, including both GAN-based and diffusion-based architectures.

3.1 Dataset

The released dataset is structured into two main components: fake and real, each further subdivided based on resolution—high-quality (*HQ*) and low-quality (*LQ*). Specifically, the fake component is organized into subsets according to the generative models utilized, which include both diffusion models (DM) and generative adversararial networks (GANs). The *LQ* subsets represent intentionally down-sized or compression-degraded images from high-resolution native generative models. A representative example of images included in the dataset is shown in Figure 1. The fake portion of the challenge dataset originates from the WILD dataset [3]. The real images were sampled from two datasets: FFHQ, originally presented in [13], and CelebA-HQ, as introduced in [12].

3.2 Competition Protocol and Duration

The timeline spanned three months, beginning with a registration phase where teams submitted details (e.g., names, institutions). Upon registration, participants signed a Data Licence Agreement (DLA) to access the training dataset. During the development phase, teams focused on attacking the released classifiers (ResNet, DenseNet) and optimizing perturbations for the blind models. Submissions were limited to three attempts per team, with only the final submission counted for evaluation. The test dataset included unperturbed deepfake images across all 16 subsets, with no ground truth provided.

(a) HQ Original 1 **(b) HQ Original 2** **(c) LQ Original 1** **(d) LQ Original 2**

Figure 1: Examples of deepfake images from the challenge dataset: representative samples from both high-quality (HQ) and low-quality (LQ) generative models.

For evaluation, participants submitted:

- **Attacked Test Set**: Perturbed images adhering to the challenge guidelines.

- **Abstract Paper**: A 1–2 page summary detailing methodology, motivation, and contributions.

Final scores were computed using a weighted combination of Structural Similarity Index (SSIM) and detection accuracy across all four classifiers (including blind models). The formula for the final score is:

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

where:

- C is the set of all classifiers used in the evaluation
- C_f is a specific classifier belonging to the set C
- N is the number of deepfake images in the test dataset
- k is the index identifying the k-th image in the dataset ($k \in \{1, 2, ..., N\}$)
- I_k is the k-th original deepfake image from the test dataset
- I_k^{ADV} is the adversarial image generated from I_k
- $SSIM(I_k, I_k^{ADV})$ is the Structural Similarity Index between the original image I_k and the adversarial image I_k^{ADV} (value between 0 and 1)
- $LABEL_{real}$ is the label of the "real" class (opposite to "deepfake")
- $[c(I_k^{ADV}) = LABEL_{real}]$ is the indicator function that returns 1 if classifier c classifies the adversarial image I_k^{ADV} as "real", 0 otherwise

The formula computes a cumulative score that rewards adversarial attacks which successfully maintain high structural similarity with the original image while fooling the classifiers into predicting the "real" label. The final score is the sum of all SSIM contributions weighted by the success of the attack on each classifier for each image. The top 3 teams were invited to submit extended papers for potential inclusion in the ACM Multimedia 2025 proceedings.

3.3 Evaluation Metrics

The participants' methods were evaluated based on two criteria:

(1) **SSIM Requirement**: Each submission had to include original deepfake images and their corresponding adversarial versions. Only complete image pairs were evaluated. The SSIM measures the structural similarity between two images by comparing their luminance, contrast, and structure. It provides a value between 0 and 1, where 1 indicates perfect similarity and 0 indicates no similarity, and is calculated as:

$$\text{SSIM}(I, K) = \frac{(2\mu_I \mu_K + c_1)(2\sigma_{IK} + c_2)}{(\mu_I^2 + \mu_K^2 + c_1)(\sigma_I^2 + \sigma_K^2 + c_2)} \quad (2)$$

where I and K are the two images being compared, μ_I and μ_K are the mean pixel intensities of images I and K respectively, σ_I^2 and σ_K^2 are the variances of pixel intensities in images I and K respectively, σ_{IK} is the covariance between the pixel intensities of images I and K, and c_1 and c_2 are small positive constants added to avoid division by zero when the denominators are close to zero, ensuring numerical stability. The mean SSIM for each classifier C_f, $\text{SSIM}_{C_f}^{avg}$ was computed as the average structural similarity across all image pairs:

$$\text{SSIM}_{C_f}^{avg} = \frac{1}{N} \sum_{k=1}^{N} SSIM(I_k, I_k^{ADV}) \quad (3)$$

where N is the total number of adversarial images evaluated, I_k is the k-th original image and I_k^{ADV} is the k-th adversarial image. Finally, we defined SSSIM Score (SSIMS) as:

$$\text{SSIMS} = \frac{1}{|C|} \sum_{C_f \in C} \text{SSIM}_{C_f}^{avg} \quad (4)$$

where C is the set of all classifiers.

(2) **Attack Success Rate (ATR) Calculation**: An adversarial image was considered a successful attack if the detection system misclassified it as "real" (i.e., failed to detect it as a deepfake). The attack success rate for each classifier was calculated as:

$$\text{ASR}_{C_f} = \frac{1}{N} \sum_{k=1}^{N} [C_f(I_k^{ADV}) = LABEL_{real}] \quad (5)$$

where N is the total number of adversarial images evaluated, C_f is a specific classifier, I_k is the k-th original image, I_k^{ADV} is the k-th adversarial image, $LABEL_{real}$ is the label for the "real" class, and $[C_f(I_k^{ADV}) = LABEL_{real}]$ is the indicator function that returns 1 if the classifier incorrectly predicts the adversarial image as "real", and 0 otherwise. This metric represents the proportion of adversarial images that successfully evaded detection by fooling the classifier into misclassifying them as authentic content. For an overall evaluation of the methods, we defined the Attack Success Score (ASS) as:

$$ASS = \frac{1}{|C|} \sum_{C_f \in C} \mathrm{ASR}_{C_f} \tag{6}$$

where C is the set of all classifiers.

4 Participants and Methods

Thirteen teams submitted innovative adversarial approaches. Below, we briefly summarize their key contributions.

DASH: Proposed a region-specialized adversarial attack framework leveraging facial, background, and synthesis-specific perturbations, optimizing via momentum-based gradients and variance-based neighbor sampling to achieve robust transferability.

DeFakePol: Adapted the Fast Gradient Sign Method (FGSM) for targeted multi-model attacks with resampling techniques (downsampling/upsampling), improving transferability across various deepfake detection architectures.

FalseNegative: Implemented a two-stage method combining an enhanced Projected Gradient Descent (PGD) with a U-Net to generate transferable perturbations, integrating constraints based on the Structural Similarity Index Measure (SSIM) to preserve visual fidelity.

GRADIANT: Developed a hybrid attack combining a pixel-level PGD (with Expectation over Transformations) and a feature-level Feature Importance Attack (FIA), using heterogeneous detector ensembles and attention masking to enhance transferability.

MICV: Integrated Nesterov-accelerated Iterative Fast Gradient Sign Method (NI-FGSM) with diverse input augmentations and an ensemble of multiple detection architectures, employing Class-wise Weight Averaging and sample selection based on SSIM.

MILab: Formulated the adversarial task within a constrained perceptual space, using diffusion-based inpainting, attention-guided modifications, and semantic-preserving measures like Learned Perceptual Image Patch Similarity (LPIPS) and SSIM alongside surrogate models for black-box settings.

MR-CAS: Proposed latent-space manipulation via Denoising Diffusion Implicit Models (DDIM) inversion and momentum-based gradient optimization (Momentum Iterative Fast Gradient Sign Method, MI-FGSM), significantly improving visual imperceptibility and transferability of adversarial perturbations.

RoMa: Employed globally distributed adversarial noise optimized through surrogate models, including a Vision Transformer (ViT-B-16) and EfficientNet-B0, refined iteratively using gradient-based methods and the Adam optimizer.

Safe AI: Introduced MIG-COW (Momentum Integrated Gradients with Consensus-Orthogonal Weighting), using Momentum Integrated Gradients and gradient decomposition into consensus and orthogonal components, substantially improving cross-model adversarial transferability.

SecureML: Developed TTDE (Test-Time Distillation Ensemble Attack), distilling knowledge from Convolutional Neural Networks (CNNs) to Vision Transformers, optimizing adversarial examples using combined cross-entropy and SSIM-based losses.

The Adversaries: Proposed MS-GAGA (Metric-Selective Guided Adversarial Generation Attack), employing dual-stream PGD (momentum and saliency-guided) to generate diverse adversarial examples, with metric-based selection ensuring structural fidelity and attack effectiveness.

VYAKRITI 2.0: Utilized ensemble-gradient-based PGD enhanced by SSIM loss and low-frequency perturbations via Fast Fourier Transform (FFT) based filtering, targeting generalization gaps in detection architectures.

WHU_PB: Introduced a lightweight adversarial generator trained via a Rectified Linear Unit (ReLU) based hinge loss and SSIM-based perceptual regularization, optionally employing attention-guided masks for efficient localized perturbations.

5 Competition Results

The competition results reveal several interesting patterns in the performance distribution. The top three teams (MR-CAS, Safe AI, and RoMa) achieved remarkably close scores, with less than 70 points separating the winner from the third-place finisher. This tight competition at the top demonstrates the high quality of solutions and the competitive nature of the challenge. Figure 2 provides a qualitative comparison of adversarial perturbations created by these teams, showcasing their ability to maintain visual fidelity while evading detection systems.

MR-CAS from the University of Chinese Academy of Sciences secured first place with a score of 2740, employing their novel latent diffusion model approach that manipulated images in the latent feature space rather than directly in pixel space. Their DDIM inversion technique proved particularly effective in generating adversarial samples with high visual fidelity and strong transferability.

Safe_AI from UNIST achieved second place with 2709 points, utilizing their Momentum Integrated Gradient with Consensus-Orthogonal Weighting (MIG-COW) framework. Their approach leveraged implementation invariance via Integrated Gradients and sophisticated gradient ensemble techniques to enhance transferability across diverse model architectures.

RoMa from Fraunhofer SIT | ATHENE Center rounded out the top three with 2679 points, implementing a white-box adversarial framework with globally distributed, data-driven noise perturbations optimized through carefully designed surrogate models.

The middle tier of teams (ranks 4-9) showed competitive performance with scores ranging from 2341 to 2631, indicating that multiple viable approaches exist for this challenging problem. These teams employed various sophisticated techniques including hybrid adversarial frameworks, ensemble methods, and advanced loss functions combining classification objectives with perceptual quality measures. A notable performance gap emerged between the top nine teams and the bottom four, suggesting that certain methodological choices and implementation details were critical for achieving

Table 1: Final competition results showing team rankings, SSIM Score (SSIMS), Attack Success Score (ASS), and Final Score (FS).

Rank	Team Name	Organization/Institution	SSIMS	ASS	FS
1	**MR-CAS**	**University of Chinese Academy of Sciences**	**0.742**	**0.672**	**2740**
2	**Safe AI**	**UNIST (Ulsan National Institute of Science and Technology)**	**0.915**	**0.528**	**2709**
3	**RoMa**	**Fraunhofer SIT | ATHENE Center**	**0.934**	**0.509**	**2679**
4	GRADIANT	Gradiant	0.853	0.551	2631
5	DASH	Sungkyunkwan University	0.848	0.543	2618
6	SecureML	University of Cagliari	0.832	0.535	2490
7	MICV	Ant Group	0.738	0.585	2434
8	WHU_PB	Wuhan University	0.834	0.487	2354
9	The Adversaries	Singapore Institute of Technology	0.713	0.590	2341
10	DeFakePol	Samsung Research Poland	0.896	0.332	1665
11	False Negative	The Hong Kong Polytechnic University	0.514	0.555	1602
12	VYAKRITI 2.0	Apex Institute of Technology Chandigarh University	0.298	0.615	1041
13	MILab	University of Science and Technology of China	0.994	0.020	110

Figure 2: Adversarial perturbations generated by top-performing teams on high-quality (HQ) and low-quality (LQ) deepfake samples. Original images (top row) are compared with adversarial examples from MR-CAS, Safe AI, and RoMa teams. Values show SSIM scores and binary predictions for ResNet-50, DenseNet-121, ViT-B-16, and DenseNet-121-DCT models. ✗ indicates successful attack (misclassification), ✓ indicates failed attack (correct classification).

high performance in this competition. Teams that struggled typically faced challenges in balancing attack effectiveness with visual quality preservation, or in achieving robust transferability across diverse detector architectures. The analysis of top-performing solutions reveals several critical methodological patterns that distinguished successful approaches from less effective ones. The winning MR-CAS team's approach demonstrated the significant effectiveness of operating in latent feature spaces rather than directly in pixel space, providing superior transferability and visual quality compared to traditional pixel-based perturbation methods. This latent space manipulation approach fundamentally changed how adversarial examples could be generated while maintaining imperceptibility. Multiple top teams successfully employed ensemble methods, either for generating attacks or for improving transferability across different model architectures. These ensemble approaches proved particularly valuable in creating adversarial examples that could fool diverse detector types, from traditional CNNs to modern Vision Transformers. Advanced optimization techniques including momentum-based optimization, diverse input transformations, and sophisticated gradient aggregation methods proved essential for achieving high performance. Teams that incorporated these techniques showed notably better results in both attack success rates and visual quality preservation. Furthermore, teams that explicitly designed their approaches to handle both CNN and Vision Transformer architectures achieved better overall performance, recognizing the diverse landscape of modern deepfake detection systems. As shown in Figure 3, the top three teams demonstrated markedly different performance patterns between white-box and black-box attacks, with white-box attacks achieving near-perfect success rates while black-box transferability remained a significant challenge. The complete ranking of all participating teams is presented in Table 1, showing the final scores achieved by each team.

6 Conclusion

The AADD-2025 challenge highlighted the vulnerability of deepfake detectors to adversarial attacks, while advancing strategies for more robust forensic systems. Top teams leveraged latent-space manipulation, ensemble gradients, and surrogate modeling to evade diverse classifiers with high visual fidelity. Key insights showed

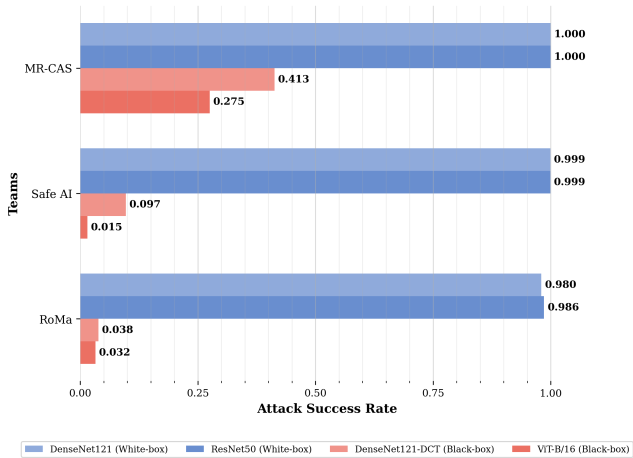

Figure 3: ASRs for top three teams across different classifier architectures. White-box attacks achieve significantly higher success rates than black-box attacks, highlighting transferability challenges in adversarial deepfake generation.

that latent-space attacks outperform pixel-level methods, ensembles improve cross-model robustness, and optimization can balance imperceptibility with attack success. Nonetheless, generalization across heterogeneous models and preservation of structural coherence remain open challenges, underscoring the need for adaptive defenses, hybrid detectors, and standardized benchmarks. Future directions may involve neurosymbolic integration and foundation models trained with adversarial examples for universal and real-time deepfake defense.

Acknowledgments

Orazio Pontorno is a PhD candidate enrolled in the National PhD in Artificial Intelligence, XXXIX cycle, organized by Università Campus Bio-Medico di Roma. The work of Orazio Pontorno and Francesco Guarnera has been supported by MUR in the framework of PNRR PE0000013, under project “Future Artificial Intelligence Research – FAIR".
Luca Guarnera and Sebastiano Battiato: this work was partially supported by project SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU, and by project FOSTERER, funded by MUR within the PRIN 2022 program under contract 202289RHHP.

References

[1] Maryam Abbasi, Paulo Váz, José Silva, and Pedro Martins. 2025. Comprehensive Evaluation of Deepfake Detection Models: Accuracy, Generalization, and Resilience to Adversarial Attacks. *Applied Sciences* 15, 3 (2025). https://doi.org/10.3390/app15031225

[2] Zahid Akhtar, Thanvi Lahari Pendyala, and Virinchi Sai Athmakuri. 2024. Video and audio deepfake datasets and open issues in deepfake technology: being ahead of the curve. *Forensic Sciences* 4, 3 (2024), 289–377. https://doi.org/10.3390/forensicsci4030021

[3] Pietro Bongini, Sara Mandelli, Andrea Montibeller, Mirko Casu, Orazio Pontorno, Claudio Vittorio Ragaglia, Luca Zanchetta, Mattia Aquilina, Taiba Majid Wani, Luca Guarnera, Benedetta Tondi, Giulia Boato, Paolo Bestagini, Irene Amerini, Francesco De Natale, Sebastiano Battiato, and Mauro Barni. 2025. WILD: a new in-the-Wild Image Linkage Dataset for synthetic image attribution. arXiv:2504.19595 [cs.MM] https://arxiv.org/abs/2504.19595

[4] Mirko Casu, Luca Guarnera, Pasquale Caponnetto, and Sebastiano Battiato. 2024. GenAI mirage: The impostor bias and the deepfake detection challenge in the era of artificial illusions. *Forensic Science International: Digital Investigation* 50 (Sept. 2024), 301795. https://doi.org/10.1016/j.fsidi.2024.301795

[5] Muhammad Umar Farooq, Awais Khan, Kutub Uddin, and Khalid Mahmood Malik. 2025. Transferable Adversarial Attacks on Audio Deepfake Detection . In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*. IEEE Computer Society, Los Alamitos, CA, USA, 1555–1564. https://doi.org/10.1109/WACVW65960.2025.00178

[6] Nigel Francis. 2025. Deepfake Detection and Defense: An Analysis of Techniques and Robustness. (2025).

[7] Chiara Galdi, Michele Panariello, Massimiliano Todisco, and Nicholas Evans. 2024. 2D-Malafide: Adversarial Attacks Against Face Deepfake Detection Systems. https://arxiv.org/abs/2408.14143

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (2020), 139–144.

[9] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. 2024. Mastering Deepfake Detection: A Cutting-Edge Approach to Distinguish GAN and Diffusion-Model Images. *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 11 (2024), 1–24.

[10] Luca Guarnera, Francesco Guarnera, Alessandro Ortis, Sebastiano Battiato, and Giovanni Puglisi. 2024. Evasion Attack on Deepfake Detection via DCT Trace Manipulation. In *International Conference on Pattern Recognition*. Springer, 157–169.

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.

[12] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Hk99zCeAb

[13] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4401–4410.

[14] Abdullah Ayub Khan, Asif Ali Laghari, Syed Azeem Inam, Sajid Ullah, Muhammad Shahzad, and Darakhshan Syed. 2025. A survey on multimedia-enabled deepfake detection: state-of-the-art tools and techniques, emerging trends, current challenges & limitations, and future directions. *Discover Computing* 28, 1 (April 2025), 48. https://doi.org/10.1007/s10791-025-09550-0

[15] Sarwar Khan, Jun-Cheng Chen, Wen-Hung Liao, and Chu-Song Chen. 2024. *Adversarially Robust Deepfake Detection via Adversarial Feature Similarity Learning*. Springer Nature Switzerland, 503–516. https://doi.org/10.1007/978-3-031-53311-2_37

[16] Marco Moltisanti, Antonino Paratore, Sebastiano Battiato, and Luigi Saravo. 2015. Image manipulation on facebook for forensics evidence. In *International Conference on Image Analysis and Processing*. Springer, 506–517.

[17] Zehra Moğulkoç and Beyzanur Yuce. 2024. A Comparative Analysis of State-of-the-Art Algorithms for Robust Deep Fake Detection. (03 2024).

[18] Paarth Neekhara, Brian Dolhansky, Joanna Bitton, and Cristian Canton Ferrer. 2021. Adversarial Threats to DeepFake Detection: A Practical Perspective. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 923–932. https://doi.org/10.1109/CVPRW53098.2021.00103

[19] Gan Pei, Jiangning Zhang, Menghan Hu, Zhenyu Zhang, Chengjie Wang, Yunsheng Wu, Guangtao Zhai, Jian Yang, Chunhua Shen, and Dacheng Tao. 2024. Deepfake Generation and Detection: A Benchmark and Survey. https://arxiv.org/abs/2403.17881

[20] Ben Pinhasov, Raz Lapid, Rony Ohayon, Moshe Sipper, and Yehudit Aperstein. 2024. XAI-Based Detection of Adversarial Attacks on Deepfake Detectors. https://arxiv.org/abs/2403.02955

[21] Orazio Pontorno, Luca Guarnera, and Sebastiano Battiato. 2024. DeepfeatureX Net: Deep Features Extractors Based Network for Discriminating Synthetic from Real Images. In *International Conference on Pattern Recognition*. Springer, 177–193.

[22] Orazio Pontorno, Luca Guarnera, and Sebastiano Battiato. 2025. DeepFeatureX-SN: Generalization of deepfake detection via contrastive learning. *Multimedia Tools and Applications* (2025), 1–20.

[23] Ngan Hoang Vo, Khoa D. Phan, Anh-Duy Tran, and Duc-Tien Dang-Nguyen. 2022. Adversarial Attacks on Deepfake Detectors: A Practical Analysis. In *MultiMedia Modeling*, Björn Þór Jónsson, Cathal Gurrin, Minh-Triet Tran, Duc-Tien Dang-Nguyen, Anita Min-Chun Hu, Binh Huynh Thi Thanh, and Benoit Huet (Eds.). Springer International Publishing, Cham, 318–330.

[24] Run Wang, Ziheng Huang, Zhikai Chen, Li Liu, Jing Chen, and Lina Wang. 2022. Anti-Forgery: Towards a Stealthy and Robust DeepFake Disruption Attack via Adversarial Perceptual-aware Perturbations. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, Lud De Raedt (Ed.). International Joint Conferences on Artificial Intelligence Organization, 761–767. https://doi.org/10.24963/ijcai.2022/107 Main Track.

[25] Yang Yang, Norisma Binti Idris, Chang Liu, Hui Wu, and Dingguo Yu. 2024. A destructive active defense algorithm for deepfake face images. *PeerJ Computer Science* 10 (Oct. 2024), e2356. https://doi.org/10.7717/peerj-cs.2356