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

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

Andrea Montibeller
Department of Information Engineering and Computer Science, University of Trento
Trento, Italy
andrea.montibeller@unitn.it

Salvatore Alfio Sambataro
Department of Mathematics and Computer Science, University of Catania
Catania, Italy
salvatore.sambataro@phd.unict.it

Dario Samuele Pishvai
Department of Mathematics and Computer Science, University of Catania
Catania, Italy
dario.pishvai@phd.unict.it

Ludovica Beritelli
Department of Mathematics and Computer Science, University of Catania
Catania, Italy
ludovica.beritelli@phd.unict.it

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

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

Giulia Boato
Department of Information Engineering and Computer Science, University of Trento
Trento, Italy
giulia.boato@unitn.it

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

Abstract

The increasing realism of AI-generated media and its widespread dissemination through social media demand deepfake detectors that remain reliable under adversarial attacks and realistic post-processing. We present AADD-2026, a competitive benchmark for evaluating adversarial attacks against deepfake detectors under realistic deployment conditions. The challenge considers synthetic images from five generative families, social-media-inspired processing, JPEG and JPEG-AI compression, and eight heterogeneous detectors operating across spatial, frequency, and foundation-model representations. Ten teams submitted valid solutions, exploring latent-space, frequency-aware, ensemble, manifold-projection, and compression-aware attacks. Results show that explicitly accounting for compression and post-processing substantially improves transferability to unseen detectors. The winning approach achieves the highest attack success through a gradient-free real-face manifold projection strategy. Meanwhile, DINOv2- and SigLIP2-based detectors show substantially greater resilience than conventional architectures, highlighting a clear trade-off between attack transferability and perceptual fidelity. Challenge resources and evaluation details are available at https://iplab.dmi.unict.it/mfs/acm-aadd-challenge-2026/.

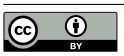
CCS Concepts

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

Keywords

Deepfake, Adversarial attacks, Multimedia Forensics

ACM Reference Format:
Francesco Guarnera, Luca Guarnera, Andrea Montibeller, Salvatore Alfio Sambataro, Dario Samuele Pishvai, Ludovica Beritelli, Giovanni Puglisi, Zahid Akhtar, Giulia Boato, and Sebastiano Battiato. 2026. Adversarial Attacks on Deepfake Detectors: A Challenge in the Era of AI-Generated Media. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3767308.3837702



This work is licensed under a Creative Commons Attribution 4.0 International License.

MM '26, Rio de Janeiro, Brazil

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2213-4/2026/11
https://doi.org/10.1145/3767308.3837702



(a) GAN (b) Diffusion (c) DiT

(d) Rectified Flow (e) Autoregressive

Figure 1: Examples of deepfake images from the AADD-2026 dataset.

1 Introduction

The rapid evolution of generative AI yields hyper-realistic synthetic media [29, 31], increasingly weaponized to amplify misinformation at an unprecedented scale [2, 18]. To counter this, synthetic media detection [3, 35] primarily relies on deep architectures like CNNs and ViTs [8, 11, 14, 28, 30, 36]. While highly accurate on pristine data, these models remain vulnerable to adversarial examples [1, 13, 27]. Unlike traditional pixel-space noise [24], the top submissions of the AADD-2025 challenge demonstrated that advanced latent-space manipulations achieve superior visual fidelity and transferability [4], highlighting cross-model robustness in black-box scenarios as a critical research priority [9].

However, traditional adversarial perturbations lack resilience in real-world environments. Social media platforms routinely apply resizing, filtering, and standard JPEG compression that easily disrupt high-frequency adversarial noise [19, 21, 23], spurring the need for compression-resistant attacks [32]. This vulnerability is further exacerbated by emerging deep learning-driven codecs like JPEG-AI [34], whose non-linear autoencoders alter image statistics in ways that degrade attacks tailored for handcrafted codecs [7].

The AADD-2026 Challenge extends previous efforts by explicitly modeling both AI-driven compression and social media-like processing pipelines. By enforcing these realistic post-processing constraints alongside cross-model evaluation, the challenge fosters the development of robust forensic tools to safeguard information integrity and digital trust [6].

2 Challenge Description

The challenge asks participants to investigate the adversarial vulnerabilities of deepfake detection models by generating perturbed deepfake images. Starting from the fake images of the official test set, participants must apply adversarial perturbations that drive the released and unreleased detectors towards the "real" class, while keeping the perturbation perceptually minimal. Building on the first edition (AADD-2025 [4]), the 2026 challenge moves the evaluation considerably closer to real-world deployment and introduces four main novelties: **1. Broader generator set.** Images span five families (27 generators): GANs [12], DM [15], DiT [26], flow-matching [20], and autoregressive transformers [10], significantly expanding the previous edition; **2. Realistic processing.** Images undergo post-processing (Gaussian noise, geometric transforms, JPEG and JPEG-AI [34] compression) and three social-media-like pipelines to evaluate attack robustness; **3. Eight evaluation detectors.** Teams receive two reference models for development, but final scoring uses eight complementary classifiers (spatial, DCT, foundation backbones, dual-stream), encouraging generalizable attacks; **4. Perceptual scoring.** Visual similarity now combines SSIM [37] and LPIPS [38] for a more faithful perceptual cost estimation.

The two released development models are a spatial `ViT-B/16` and a frequency-domain `DenseNet-DCT`. Final evaluation uses the eight detectors (Section 2.2), hence teams must design highly transferable attacks.

2.1 Dataset

The dataset comprises real and fake images. The fake subset extends the WILD dataset [5] with modern text-to-image generations, totaling ∼27,000 synthetic images across five families (Figure 1): **GANs [12]** (StyleGAN, StyleGAN2, StyleGAN3), **diffusion models [15]** (e.g., DALL-E 3, SDXL, Midjourney, Leonardo AI), **diffusion transformers [26]** (e.g., Adobe Firefly, Qwen-Image, Tencent Hunyuan, NVIDIA Sana), **flow-matching/rectified-flow models [20]** (e.g., FLUX.1, SD 3.5, Stable Cascade), and **autoregressive transformers [10]** (DALL-E 1). Real images (class 0) are sampled from FFHQ [17] and CelebA-HQ [16] to balance the synthetic data.

To simulate realistic deployment, images undergo additive Gaussian noise, JPEG and JPEG-AI [34] compression (QF 50–90), and geometric transformations (mirroring, scaling, and 45°-step rotations), combined with three social-media re-encoding pipelines. These transformations assess robustness to compression and online sharing, which can alter or mask generative spectral traces [19, 21, 23].

The initial dataset contains 24,975 fake and 26,500 real training images, split 80/20 into training and validation sets (Section 2.2). The held-out test set contains 2,025 fake images (∼75 per generator). To ensure task difficulty, only images correctly classified as fake by *all* detectors before attack are retained, yielding 1,600 final challenge images.

For final evaluation, submitted adversarial images are re-processed to assess real-world resilience: 600 remain unaltered, 500 undergo classical filtering (noise, JPEG, JPEG-AI), and 500 are subjected to the social-media pipelines.

2.2 Detectors

The final evaluation uses eight binary detectors (class 0 = Real, class 1 = Fake) spanning complementary domains to assess cross-domain attack transferability. An attack succeeds if the adversarial fake image is predicted as "real".

The eight detectors are grouped into four categories according to their representation domain. **Spatial-domain models (raw pixels):** `ResNet-50` and `DenseNet-121`, using RGB input at 256 × 256 with ImageNet normalization; `ViT-B/16` (released), using RGB input at 224 × 224 (Resize 256 → CenterCrop 224).

Table 1: Overall results and per-classifier Attack Success Rate (ASR). Final Score (FS) combines SSIM and LPIPS ($\alpha = 0.5$). ViT-B/16 and DenseNet-DCT were disclosed; all other classifiers were blind. Bold indicates the top-3 teams and the best ASR per classifier.

		Overall Competition Metrics				Per-Classifier Attack Success Rate (ASR)							
Rank	Team	SSIM	LPIPS	ASS	FS	ViT-B/16	Dense Net-DCT	Res Net-50	Dense Net-121	DINOv2	SigLIP2	FMSI	FreqViT
1	**FakeBusters@UNISA**	**0.726**	**0.426**	**0.672**	**5561.78**	0.959	0.891	**0.986**	**0.978**	0.049	**0.074**	**0.953**	**0.490**
2	**CASALab**	**0.805**	**0.319**	**0.498**	**4725.74**	0.915	**0.951**	0.357	0.469	**0.063**	0.061	0.748	0.419
3	**Mizhi Lab**	**0.842**	**0.327**	**0.295**	**2845.52**	**0.979**	0.921	0.037	0.051	0.036	0.018	0.211	0.107
4	BalzanaBits	0.900	0.213	0.252	2725.45	0.913	0.706	0.071	0.076	0.039	0.043	0.099	0.070
5	Fraunhofer SIT	0.826	0.139	0.248	2694.42	0.873	0.863	0.028	0.048	0.011	0.005	0.039	0.118
6	GhostChamber	0.805	0.386	0.281	2543.99	0.946	0.645	0.054	0.082	0.038	0.026	0.229	0.229
7	Harmony Forge	0.852	0.284	0.236	2356.23	0.839	0.615	0.083	0.074	0.049	0.026	0.137	0.061
8	Safe AI	0.931	0.178	0.202	2286.86	0.843	0.674	0.009	0.016	0.001	0.002	0.014	0.059
9	PRLAB	0.888	0.195	0.210	2282.89	0.902	0.518	0.014	0.030	0.013	0.003	0.091	0.108
10	AdverSiena	0.927	0.145	0.143	1652.26	0.844	0.153	0.032	0.031	0.013	0.011	0.034	0.027

Frequency-domain models (DCT spectrum): `DenseNet-DCT` (released), a DenseNet-121 adapted to a single-channel log-scaled DCT map ($1 \times 128 \times 128$); `freqvit`, a lightweight ViT using a 3-channel log-DCT ($3 \times 224 \times 224$).

Foundation-model backbones: `DINOv2` [25] (freezed), with an MLP head and native 518×518 RGB input; `SigLIP2` [33] (freezed), with an MLP head and 384×384 RGB input.

Dual-stream fusion: `fmsi`, a dual-branch network [22] fusing a 4-channel ViT-B/16 (RGB + frequency) with a high-frequency DCT convolutional branch.

2.3 Competition Protocol and Duration

The competition spans approximately three months. After registration and the Data Licence Agreement (DLA), participants receive the test set, two reference detectors, and evaluation configurations. Teams may submit up to three times, with only the final submission determining the ranking. Each submission includes the attacked test set and a 1–2 page Technical Abstract. Final scoring follows Section 2.4, and top-ranked teams are invited to submit extended papers to the conference proceedings.

2.4 Evaluation Metrics

Each submission is evaluated on the perceptual quality of the adversarial images and the attack's success against the detector pool. Perceptual quality combines SSIM [37] and LPIPS [38] (AlexNet backbone). For each original deepfake I_k and its adversarial counterpart I_k^{ADV}, the similarity is:

$$\text{sim}(I_k, I_k^{ADV}) = \alpha \cdot \text{SSIM}(I_k, I_k^{ADV}) + (1-\alpha)\left(1 - \text{LPIPS}(I_k, I_k^{ADV})\right) \tag{1}$$

where $\alpha = 0.5$ gives equal weight to both metrics, ensuring higher values correspond to more imperceptible perturbations.

Final score. The Final Score (FS) rewards imperceptible attacks that successfully fool the detectors. It aggregates the similarity across all N test pairs, weighted by the attack's success on each detector C_f in the pool C:

$$FS = \sum_{k=1}^{N} \sum_{C_f \in C} w_{C_f} \cdot \text{sim}(I_k, I_k^{ADV}) \cdot \left[\, C_f(I_k^{ADV}) = LABEL_{real} \,\right] \tag{2}$$

Here, $w_{C_f} = 1.0$ is the detector weight, and the indicator function equals 1 if C_f misclassifies I_k^{ADV} as "real". Thus, only complete and successful pairs contribute to FS.

Reported metrics. Team performance (Table 1) is summarized by four aggregate quantities: mean SSIM, mean LPIPS, Final Score (FS), and Attack Success Score (ASS). ASS is the average of the per-detector Attack Success Rate (ASR) across all $|C|$ detectors:

$$\text{ASS} = \frac{1}{|C|} \sum_{C_f \in C} \text{ASR}_{C_f}, \tag{3}$$

where $\text{ASR}_{C_f} = \frac{1}{N}\sum_{k=1}^{N}\left[\, C_f(I_k^{ADV}) = LABEL_{real} \,\right]$. Table 1 provides a finer breakdown reporting the per-detector ASR. This reveals which representation domains best resist specific attacks, evaluating performance across the two released reference models and the six blind ones.

3 Participants and Methods

Ten of the 21 registered teams submitted valid adversarial solutions. We briefly summarize their core methodologies below.

AdverSiena: Used Jacobian-induced Mahalanobis Attack (JMA) and Frequency-Consistent Regularization (FCR) on a diverse surrogate ensemble (CNNs, ViTs, CLIP, grayscale DCT). JMA optimizes perturbations via Non-Negative Least Squares; FCR aggregates Jacobians from low-pass views to prevent high-frequency overfitting.

BalzanaBits: Applied JMA and FCR, integrating a LoRA-finetuned SigLIP2 into the surrogate ensemble to target vision-language models. A dynamic stopping criterion halts optimization if per-channel SSIM drops below 0.85 or MSE exceeds predefined bounds, ensuring perceptual quality.

CASALab: Introduced LOKI, a dual-path generator (spatial U-Net, differentiable DCT branch) yielding single-pass adversarial perturbations without iterative optimization. A three-stage curriculum with aggressive EOT (JPEG, WebP, blur, social-media filters) explicitly targets compression robustness.

FakeBusters@UNISA: Proposed a gradient-free, detector-agnostic attack via real-face manifold projection. A dual-branch autoencoder (RGB and DCT with self-attention), trained on FFHQ real faces, adds its reconstruction residual to deepfakes at inference, requiring no surrogate classifier.

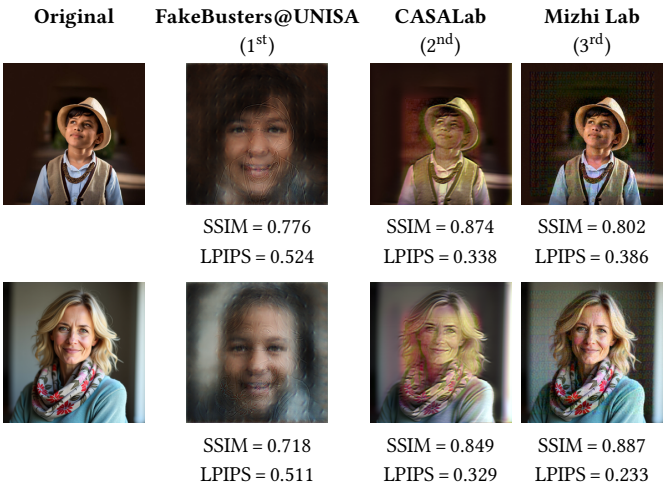


Figure 2: Qualitative comparison of adversarial images from the top-three teams. FakeBusters (rank 1) shows visible identity drift, CASALab (rank 2) introduces spectral and color distortions, while Mizhi Lab (rank 3) generally preserves the highest perceptual fidelity.

Fraunhofer SIT: Combined MI-FGSM with Spectrum Simulation Attack (differentiable DCT dropping/rescaling frequencies) across a 7-surrogate dynamically upweighted ensemble. EoT randomly applies 15 social-media filters during optimization, backed by margin-based early stopping.

GhostChamber: Developed ORACLE, a 5-phase pipeline: DCT-domain optimization, bicubic resolution lifting, border-masked ViT attack with mid-frequency enhancement, anti-forensic post-processing (fingerprint suppression, camera noise), and adaptive rollback verification.

Harmony Forge: Used TI-FGSM against ViT-B/16 and NI-FGSM against DenseNet-DCT, fusing updates via a gradient-consensus rule. Sobel edge-guided budget extension and DCT feature-map saliency amplify frequency perturbations.

Mizhi Lab: Proposed MeD-Fuse, merging two gradient-isolated branches (pixel-space MIG-COW and DCT-aligned) at inference. A ProfileMixed DCT-EOT module simulates compression (JPEG, resize, blur), regularized by LPIPS-VGG.

PRLAB: Applied MI-FGSM (1:4 spatial-to-frequency emphasis) with a JPEG-Aware Straight-Through Estimator (JSTE) for forward-pass compression. Multi-restart selection and LPIPS polishing joint-optimize attack success and perceptual metrics.

Safe AI: Introduced LCD-Attack (Luminance-Chrominance Decomposition), targeting DenseNet-DCT via luminance and ViT-B/16 via orthogonal chrominance perturbations. This mathematically guarantees mutual non-interference within a shared ℓ_∞ budget.

4 Competition Results

The results (Table 1) reveal a stratified landscape: the top two teams significantly outperformed the rest by explicitly targeting compression robustness and cross-model transferability, rather than overfitting to the released detectors.

FakeBusters@UNISA (1st, FS 5561.78): Achieved the broadest attack coverage via manifold projection, but incurred the highest perceptual cost (SSIM 0.726, LPIPS 0.426), as the projection visibly alters image appearance.

CASALab (2nd, FS 4725.74): Their compression-aware curriculum yielded the highest DenseNet-DCT ASR (95.1%) and strong generalization to several blind classifiers.

Mizhi Lab (3rd, FS 2845.52): Recorded the highest ViT-B/16 ASR (97.9%) using a gradient-isolated dual-branch strategy specifically targeting the disclosed models.

The mid-tier teams (**BalzanaBits, Fraunhofer SIT, GhostChamber, Harmony Forge**; ranks 4–7, FS 2356–2725) succeeded primarily on disclosed classifiers with limited transferability. Notably, BalzanaBits and Fraunhofer SIT achieved top perceptual metrics (SSIM 0.900 and LPIPS 0.139, respectively), while GhostChamber's anti-forensic post-processing gained traction on frequency-based blind models (22.9% ASR on FMSI/FreqViT).

The remaining teams (**Safe AI, PRLAB, AdverSiena**; ranks 8–10, FS 1652–2286) demonstrated sophisticated attacks on disclosed classifiers but near-zero transferability to blind ones. Safe AI reached the highest overall visual quality (SSIM 0.931), whereas AdverSiena's low DenseNet-DCT performance (15.3% ASR) severely constrained its score.

An analysis across all submissions highlights four key patterns: 1) **Foundation model resilience**: DINOv2 and SigLIP2 proved impervious to standard CNN/ViT-targeted perturbations (all teams < 10% ASR). 2) **Incidental frequency susceptibility**: FMSI and FreqViT blind models were highly vulnerable to methods altering frequency statistics (FakeBusters reached 95.3% and 49.0% ASR, respectively; CASALab 74.8% and 41.9%), despite not being explicitly targeted. 3) **Compression robustness drives transferability**: Only teams explicitly handling severe transformations (FakeBusters, CASALab) generalized broadly. 4) **Perceptual vs. effectiveness trade-off**: The final ranking nearly inverts the SSIM ranking. Teams with the highest visual fidelity (AdverSiena, Safe AI) placed last, while the winner recorded the lowest SSIM. This underscores the difficulty of evading eight diverse detectors after severe compression without degrading image quality (shown in Figure 2).

5 Conclusions

The AADD-2026 challenge highlighted the importance of modeling realistic post-processing for robust deepfake detection. Top-performing methods achieved strong cross-model transferability through gradient-free manifold projection and compression-aware training. A clear trade-off emerged between perceptual fidelity and attack effectiveness, while foundation models proved more resilient than frequency-based detectors. Developing defenses that withstand severe compression while preserving image quality remains a key challenge for future research.

Acknowledgments

This work was supported by the DEFORM project, funded by the European Union's Horizon Europe Research and Innovation Programme under Grant Agreement No. 101308502, and by the Xunta de Galicia and the European Regional Development Fund under Project ED431C 2025/41.

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), 1225.

[2] Sifat Muhammad Abdullah, Aravind Cheruvu, Shravya Kanchi, Taejoong Chung, Peng Gao, Murtuza Jadliwala, and Bimal Viswanath. 2024. An Analysis of Recent Advances in Deepfake Image Detection in an Evolving Threat Landscape. In *2024 IEEE Symposium on Security and Privacy (SP)*. 91–109. doi:10.1109/SP54263.2024.00194

[3] Irene Amerini, Mauro Barni, Sebastiano Battiato, Paolo Bestagini, Giulia Boato, Vittoria Bruni, Roberto Caldelli, Francesco De Natale, Rocco De Nicola, Luca Guarnera, et al. 2025. Deepfake media forensics: Status and future challenges. *Journal of Imaging* 11, 3 (2025), 73.

[4] 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*. 13714–13719.

[5] Pietro Bongini, Sara Mandelli, Andrea Montibeller, Mirko Casu, Orazio Pontorno, Claudio Vittorio Ragaglia, Luca Zanchetta, Mattia Aquilina, Taiba Majid Wani, Luca Guarnera, et al. 2025. WILD: a new in-the-Wild Image Linkage Dataset for synthetic image attribution. In *2025 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[6] 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 (2024), 301795.

[7] Tong Chen and Zhan Ma. 2023. Toward Robust Neural Image Compression: Adversarial Attack and Model Finetuning. *IEEE Transactions on Circuits and Systems for Video Technology* 33, 12 (Dec. 2023), 7842–7856. doi:10.1109/tcsvt.2023.3276442

[8] Davide Alessandro Coccomini, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. 2022. Combining efficientnet and vision transformers for video deepfake detection. In *International conference on image analysis and processing*. Springer, 219–229.

[9] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. 2018. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 9185–9193.

[10] Patrick Esser, Robin Rombach, and Björn Ommer. 2021. Taming Transformers for High-Resolution Image Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12873–12883.

[11] Claudio Giusti, Luca Guarnera, and Sebastiano Battiato. 2026. Proto-LeakNet: Towards signal-leak aware attribution in synthetic human face imagery. *Computer Vision and Image Understanding* (2026), 104848.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 27.

[13] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. arXiv:1412.6572 [stat.ML] https://arxiv.org/abs/1412.6572

[14] 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.

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

[16] 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 (ICLR)*.

[17] 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 (CVPR)*. 4401–4410.

[18] 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 (2025), 48.

[19] Manyi Li, Renshuai Tao, Yufan Liu, Chuangchuang Tan, Haotong Qin, Bing Li, Yunchao Wei, and Yao Zhao. 2025. Pay Less Attention to Deceptive Artifacts: Robust Detection of Compressed Deepfakes on Online Social Networks. arXiv:2506.20548 [cs.CV] https://arxiv.org/abs/2506.20548

[20] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2023. Flow Matching for Generative Modeling. In *International Conference on Learning Representations (ICLR)*.

[21] Jun Liu, Jiantao Zhou, Haiwei Wu, Weiwei Sun, and Jinyu Tian. 2023. Generating robust adversarial examples against online social networks (osns). *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 4 (2023), 1–26.

[22] Xin Luo and Yi Wang. 2025. Frequency-Domain Masking and Spatial Interaction for Generalizable Deepfake Detection. *Electronics* 14, 7 (2025), 1302. doi:10.3390/electronics14071302

[23] 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.

[24] Paarth Neekhara, Brian Dolhansky, Joanna Bitton, and Cristian Canton Ferrer. 2021. Adversarial threats to deepfake detection: A practical perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 923–932.

[25] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2024. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research (TMLR)* (2024).

[26] William Peebles and Saining Xie. 2023. Scalable Diffusion Models with Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 4195–4205.

[27] G Pei, J Zhang, M Hu, Z Zhang, C Wang, Y Wu, G Zhai, J Yang, C Shen, and D Tao. 2024. Deepfake Generation and Detection: A Benchmark and Survey. ArXiv, 1-28.

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

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10684–10695.

[30] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[31] Abu Taib Mohammed Shahjahan, Mohammad Mannan, Abdessamad Ben Hamza, and Amr Youssef. 2026. On Improving Robustness of Deepfake Image Detectors. arXiv:2606.02797 [cs.CR] https://arxiv.org/abs/2606.02797

[32] Richard Shin, Dawn Song, et al. 2017. Jpeg-resistant adversarial images. In *NIPS 2017 workshop on machine learning and computer security*, Vol. 1.

[33] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. 2025. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. *arXiv preprint arXiv:2502.14786* (2025).

[34] Daria Tsereh, Mark Mirgaleev, Ivan Molodetskikh, Roman Kazantsev, and Dmitriy Vatolin. 2024. JPEG AI Image Compression Visual Artifacts: Detection Methods and Dataset. arXiv:2411.06810 [cs.AI] https://arxiv.org/abs/2411.06810

[35] Luisa Verdoliva. 2020. Media forensics and deepfakes: an overview. *IEEE journal of selected topics in signal processing* 14, 5 (2020), 910–932.

[36] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. 2020. CNN-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8695–8704.

[37] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612. doi:10.1109/TIP.2003.819861

[38] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 586–595.