(DFF '25) 1st Deepfake Forensics Workshop: Detection, Attribution, Recognition, and Adversarial Challenges in the Era of AI-Generated Media

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 generative models, particularly Generative Adversarial Networks (GANs) and Diffusion Models, has reshaped multimedia content creation. Alongside creative and commercial opportunities, they have introduced unprecedented risks through the production of highly realistic synthetic content, or deepfakes. These artifacts challenge visual and auditory trust, with major implications for media, security, politics, and law. This workshop provides a forum to examine deepfake technology from forensic, technical, legal, and social perspectives. It will bring together experts to advance robust and explainable detection methods, define benchmarking practices, and address ethical and regulatory frameworks. Topics include detection and attribution, adversarial countermeasures, multimodal analysis, model traceability, legal admissibility of synthetic content, as well as real-world deployment challenges and dataset creation. Further information about the workshop is available at https://iplab.dmi.unict.it/mfs/acm-dff-ws-2025/.

CCS Concepts

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

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3762241

Keywords

Deepfake Detection, Deepfake Attribution, Adversarial Attacks, Generative Models, Digital Forensics

ACM Reference Format:
Sebastiano Battiato, Mirko Casu, Francesco Guarnera, Luca Guarnera, Giovanni Puglisi, Orazio Pontorno, Claudio Vittorio Ragaglia, and Zahid Akhtar. 2025. (DFF '25) 1st Deepfake Forensics Workshop: Detection, Attribution, Recognition, and Adversarial Challenges in the Era of AI-Generated Media. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3746027.3762241

1 Introduction and Motivation

In recent years, Generative AI has rapidly advanced through models such as GANs [8] and Diffusion Models [12, 19, 20], enabling the creation of highly realistic synthetic media, or deepfakes. These include manipulated or fully synthetic images, videos, and audio that convincingly mimic real individuals. While initially developed for creative and accessibility purposes, deepfakes have become dual-use technologies, raising major concerns for cybersecurity, forensics, journalism, and democratic governance [2, 3].

The growing accessibility of user-friendly generative tools, coupled with the increasing quality of the outputs, has significantly lowered the barrier to entry for malicious actors. Deepfakes are now being used to execute a variety of harmful operations, including political disinformation, impersonation fraud, defamation, social engineering attacks, and privacy violations [4, 7]. In one notable

case, cybercriminals used AI-generated speech to mimic the voice of a CEO and trick a bank into transferring $35 million [1]. In another, synthetic audio imitating former US president Biden was disseminated in robocalls to mislead voters ahead of the New Hampshire primaries[2]. The illicit use of synthetic media introduces new forms of digital deception, capable of undermine public trust, destabilizing democratic institutions, and eroding the credibility of legitimate audiovisual evidence.

From a psychological and social perspective, the proliferation of high-quality synthetic content contributes to the emergence of a phenomenon known as *Impostor Bias* [4], where users are increasingly unable or unwilling to trust authentic media. This erosion of epistemic trust affects key sectors such as journalism, legal processes, national security, and interpersonal communication, amplifying the already complex challenge of navigating the modern information ecosystem. The implications are various: in courtrooms, deepfakes can compromise the integrity of evidence; in politics, they can distort democratic debate; in personal contexts, they can be used for extortion.

In contrast to malicious use of deepfakes, the multimedia forensics community has developed numerous detection algorithms that aim to distinguish real content from AI-generated media. Early detection methods focused on identifying spatial and temporal inconsistencies using convolutional neural networks (CNNs) [9, 17, 18, 23], while more recent work has explored interpretable descriptors in the frequency domain, such as statistics from the Discrete Cosine Transform (DCT) [5, 16]. These frequency-based features offer a promising trade-off between accuracy, efficiency, and interpretability, making them particularly appealing in forensic and legal scenarios. However, adversarial research has revealed the fragility of many of these approaches. Attackers can craft subtle perturbations that alter feature statistics while preserving the perceptual quality of the content, thus bypassing detection systems [11]. Such *adversarial examples* expose critical vulnerabilities in systems once considered trustworthy.

Furthermore, while in the past research has focused mainly on binary detection (true vs. false), today the task of *attribution* is becoming increasingly important, i.e., determining the generative model, technique, or pipeline responsible for a given deepfake. Deepfake attribution is critical in forensic investigations: it allows analysts to infer responsibility, trace the origin of synthetic content, and identify specific models [10]. Promising directions include lightweight spatio-temporal networks for model attribution in face-swap scenarios, as FAME [1]. However, it presents significant challenges. Generative models evolve rapidly, often with open source code being optimized and redistributed, producing subtle and overlapping artifacts. In the audio domain, this complexity is further amplified. Voice cloning technologies can synthesize speech that mimics pitch, rhythm, and idiosyncrasies with near-human precision, making the impersonation extremely convincing and difficult to detect [6, 13, 15, 21].

This increasingly complex landscape, in which identification, attribution, and legal accountability are constantly being surpassed by generative innovation, requires interdisciplinary action. Although technical solutions continue to evolve, they must be integrated with robust forensic frameworks, resilient assessment protocols, and legal mechanisms capable of responding to real-world threats. Furthermore, the ability to interpret, explain and trust these tools becomes critical not only for investigators and computer forensic experts, but also for journalists, legal authorities, and policy makers.

Recent studies have highlighted that attribution can be addressed through architectural fingerprinting approaches such as DNA-Det [24], or by leveraging contrastive open-world strategies like CPL [22], both of which aim to recognize subtle model specific artifacts.

The DFF-2025 workshop aims to encourage interdisciplinary debate by bringing together researchers from the fields of computer vision, multimedia forensics, machine learning, adversarial robustness, and digital law. Researchers are invited to explore innovative approaches, challenge existing assumptions, and contribute to the development of reliable artificial intelligence systems for detecting, attributing, and mitigating synthetic media.

The main contributions of the DFF-2025 Workshop are:

- A multi-perspective forum linking technical, forensic, and regulatory insights on the deepfake phenomenon.
- Focus on cross-modal detection and attribution methods, covering both visual and audio deepfakes.
- Sessions dedicated to adversarial attacks and defense strategies, with a focus on robustness and generalization.
- Emphasis on explainable and interpretable AI techniques for synthetic media forensics.
- Creation and benchmarking of datasets, protocols, and tools for reproducible evaluation.

2 Workshop Objectives

This workshop aims to address the challenges posed by synthetic media by promoting collaboration between different disciplines. Our goal is to bring together researchers and professionals from the fields of computer vision, multimedia forensics, cybersecurity, law, and ethics to discuss new strategies for detecting, attributing, and mitigating deepfakes. These issues align with earlier forensic investigations on the impact of social network processing on digital images [14], now further amplified by the rise of AI-generated synthetic content. We aim to bridge the gap between cutting-edge technological innovation and social resilience, highlighting both the opportunities and risks inherent in generative AI.
Topics of interest include (but are not limited to):

- Deepfake Detection on Images, Video, and Audio
- Multimodal Deepfake Detection
- Deepfake Model Recognition and Attribution
- Forensic Ballistics for Deepfake Analysis
- Adversarial Forensics and Counter-Forensic Techniques
- Generative Models for Deepfake Creation
- Multimodal Datasets for Deepfake Detection and Generation
- Explainability and Interpretability in AI Forensics
- Passive Deepfake Authentication Methods
- Active Deepfake Authentication Methods
- Deepfakes Detection Method on Realistic Scenarios
- Legal and Ethical Implications of Deepfakes: Detection, Regulation, and Accountability

[1] https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/, Last accessed: 02 August 2025
[2] https://www.reuters.com/world/us/fake-biden-robo-call-tells-new-hampshire-voters-stay-home-2024-01-22/, Last accessed: 02 August 2025

Consistent with the objectives of *ACM Multimedia 2025*, this workshop addresses the urgent global need for standards, tools, and policies capable of managing the risks and responsibilities introduced by generative AI. It advocates for collaborative research that combines technical rigor with ethical foresight to support the credibility and resilience of digital media worldwide.

3 Conclusion

The advancement of generative models has made synthetic media increasingly realistic and accessible, intensifying the threat posed by deepfakes across visual and audio domains. Although detection methods have improved, many still lack robustness and generalization, particularly when faced with new or malicious content. Beyond detection, attribution plays a critical role in forensics, enabling the identification of specific generative models or pipelines responsible for manipulated content. However, this remains a challenging task due to the evolving nature of generation techniques and the subtlety of associated artifacts. The research addressed in the proposed workshop highlights the need for integrated approaches that combine reliable detection with detailed attribution, while addressing explainability and resilience to adversarial attacks. With the continued proliferation of synthetic media, such capabilities are essential for preserving the reliability of digital content and ensuring accountability in complex information ecosystems.

Acknowledgments

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.

The work of Claudio Vittorio Ragaglia has been supported by the Spoke 1 "Future HPC BigData" of the Italian Research Center on High-Performance Computing, Big Data and Quantum Computing (ICSC) funded by MUR Missione 4 Componente 2 Investimento 1.4: Potenziamento strutture di ricerca e creazione di "campioni nazionali di R&S (M4C2-19)" - Next Generation EU (NGEU).

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".

The work of Mirko Casu was supported by research PIAno di inCEntivi per la Ricerca di Ateneo 2024/2026 — Linea di Intervento i "Progetti di ricerca collaborativa" - SIAM Project - University of Catania, Italy.

References

[1] Wasim Ahmad, Yan-Tsung Peng, and Yuan-Hao Chang. 2025. FAME: A Lightweight Spatio-Temporal Network for Model Attribution of Face-Swap Deepfakes. *Expert Systems with Applications* (2025), 128571.

[2] Irene Amerini, Mauro Barni, Sebastiano Battiato, Paolo Bestagini, Giulia Boato, Tania Sari Bonaventura, Vittoria Bruni, Roberto Caldelli, Francesco De Natale, Rocco De Nicola, Luca Guarnera, Sara Mandelli, Gian Luca Marcialis, Marco Micheletto, Andrea Montibeller, Giulia Orrù, Alessandro Ortis, Pericle Perazzo, Giovanni Puglisi, Davide Salvi, Stefano Tubaro, Claudia Melis Tonti, Massimo Villari, and Domenico Vitulano. 2025. Deepfake Media Forensics: State of the Art and Challenges Ahead. In *Advances in Social Networks Analysis and Mining*, I-Hsien Ting, Reda Alhajj, Panagiotis Karampelas, and Min-Yuh Day (Eds.). Springer Nature Switzerland, Cham, 33–48.

[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] 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.

[5] Sara Concas, Gianpaolo Perelli, Gian Luca Marcialis, and Giovanni Puglisi. 2022. Tensor-Based Deepfake Detection In Scaled And Compressed Images. In *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 3121–3125.

[6] Andrea Di Pierno, Luca Guarnera, Dario Allegra, and Sebastiano Battiato. 2025. End-to-end Audio Deepfake Detection from RAW Waveforms: a RawNet-Based Approach with Cross-Dataset Evaluation. *arXiv preprint arXiv:2504.20923* (2025).

[7] Europol. 2022. Facing reality? Law enforcement and the challenge of deepfakes. In *an observatory report from the Europol Innovation Lab, Publications Office of the European Union, Luxembourg*.

[8] 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*. 2672–2680.

[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* (2024). https://doi.org/10.1145/3652027

[10] Luca Guarnera, Oliver Giudice, Matthias Nießner, and Sebastiano Battiato. 2022. On the Exploitation of Deepfake Model Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 61–70.

[11] 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.

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

[13] Nicholas Klein, Tianxiang Chen, Hemlata Tak, Ricardo Casal, and Elie Khoury. 2024. Source Tracing of Audio Deepfake Systems. In *Interspeech 2024*. ISCA, 1100–1104. https://doi.org/10.21437/interspeech.2024-1283

[14] 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.

[15] Nicolas Müller, Franziska Diekmann, and Jennifer Williams. 2022. Attacker Attribution of Audio Deepfakes. In *Interspeech 2022*. 2788–2792. https://doi.org/10.21437/Interspeech.2022-129

[16] Orazio Pontorno, Luca Guarnera, and Sebastiano Battiato. 2024. On the Exploitation of DCT-Traces in the Generative-AI Domain. In *2024 IEEE International Conference on Image Processing (ICIP)*. 3806–3812. https://doi.org/10.1109/ICIP51287.2024.10648013

[17] Orazio Pontorno, Luca Guarnera, and Sebastiano Battiato. 2025. DeepFeatureX Net: Deep Features EXtractors Based Network for Discriminating Synthetic from Real Images. In *International Conference on Pattern Recognition*. Springer, 177–193.

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

[19] 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*. 10684–10695.

[20] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep Unsupervised Learning Using Nonequilibrium Thermodynamics. In *International Conference on Machine Learning*. PMLR, 2256–2265.

[21] Adriana Stan, David Combei, Dan Oneata, and Horia Cucu. 2025. TADA: Training-free Attribution and Out-of-Domain Detection of Audio Deepfakes. *arXiv:2506.05802* (2025). arXiv:2506.05802 [eess.AS]

[22] Zhimin Sun, Shen Chen, Taiping Yao, Bangjie Yin, Ran Yi, Shouhong Ding, and Lizhuang Ma. 2023. Contrastive pseudo learning for open-world deepfake attribution. In *Proceedings of the IEEE/CVF international conference on computer vision*. 20882–20892.

[23] 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.

[24] Tianyun Yang, Ziyao Huang, Juan Cao, Lei Li, and Xirong Li. 2022. Deepfake network architecture attribution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 4662–4670.